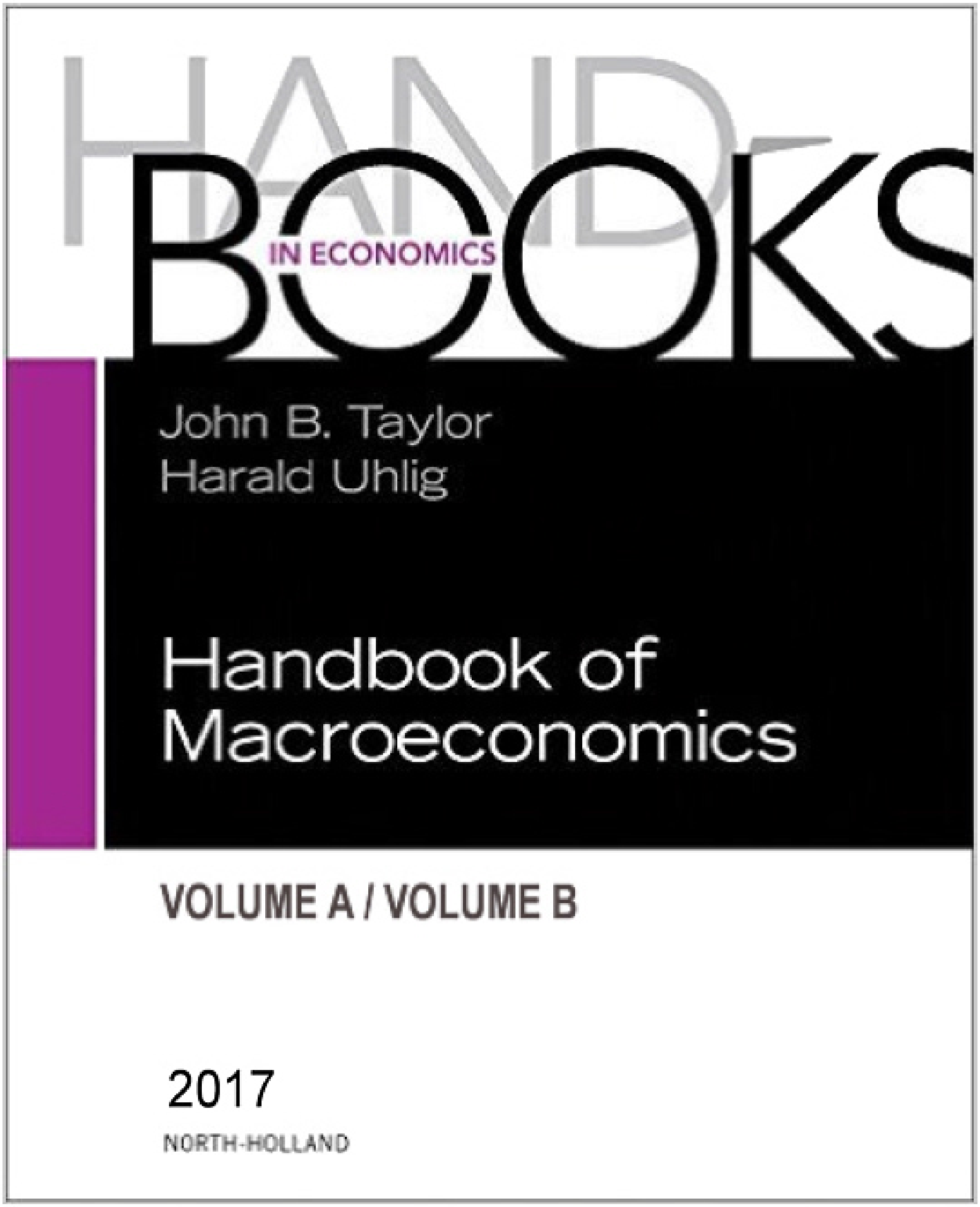


HANDBOOKS IN ECONOMICS

John B. Taylor
Harald Uhlig

Handbook of Macroeconomics

VOLUME A / VOLUME B

2017

NORTH-HOLLAND

Handbook of Macroeconomics, 2

FIRST EDITION

John B. Taylor

Stanford University, Stanford, CA, United States

Harald Uhlig

University of Chicago, Chicago, IL, United States

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo
North-Holland is an imprint of Elsevier

EDITOR'S BIOGRAPHY

John B. Taylor is the Mary and Robert Raymond Professor of Economics at Stanford University and the George P. Shultz Senior Fellow in Economics at Stanford's Hoover Institution. He is also the director of Stanford's Introductory Economics Center. His research focuses on macroeconomics, monetary economics, and international economics. He coedited Volume 1 of the Handbook of Macroeconomics and recently wrote *Getting Off Track*, one of the first books on the financial crisis, and *First Principles: Five Keys to Restoring America's Prosperity*. He served as senior economist and member of the President's Council of Economic Advisers. From 2001 to 2005, he served as undersecretary of the US Treasury for international affairs. Taylor was awarded the Hoagland Prize and the Rhodes Prize by Stanford University for excellence in undergraduate teaching. He received the Alexander Hamilton Award and the Treasury Distinguished Service Award for his policy contributions at the US Treasury. Taylor received a BA in economics summa cum laude from Princeton and a PhD in economics from Stanford.

Harald Uhlig, born 1961, is a professor at the Department of Economics of the University of Chicago since 2007 and was chairman of that department from 2009 to 2012. Previously, he held positions at Princeton, Tilburg University and the Humboldt-Universität zu Berlin. His research interests are in quantitative macroeconomics, financial markets, and Bayesian econometrics. He served as coeditor of *Econometrica* from 2006 to 2010 and as editor of the *Journal of Political Economy* since 2012 (head editor since 2013). He is a consultant of the Bundesbank, the European Central Bank, and the Federal Reserve Bank of Chicago. He is a fellow of the Econometric Society and a recipient of the Gossen Preis of the Verein für Socialpolitik, awarded annually to an economist in the German language area whose work has gained an international reputation.

CONTRIBUTORS

E. Afanasyeva
IMFS, Goethe University Frankfurt, Frankfurt, Germany

M. Aguiar
Princeton University, Princeton, NJ, United States

A. Alesina
Harvard University, Cambridge, MA, United States; IGIER, Bocconi University, Milan, Italy

G.-M. Angeletos
MIT; NBER, Cambridge, MA, United States

S. Basu
Boston College, Chestnut Hill; NBER, Cambridge, MA, United States

M.D. Bordo
Rutgers University, New Brunswick, NJ; NBER, Cambridge, MA, United States

J. Borovička
New York University, New York, NY; NBER, Cambridge, MA, United States

P. Brinca
Nova School of Business and Economics, Lisboa; Centre for Economics and Finance, University of Porto, Porto, Portugal

M.K. Brunnermeier
Princeton University, Princeton, NJ, United States

V.V. Chari
University of Minnesota; Federal Reserve Bank of Minneapolis, Minneapolis, MN, United States

S. Chatterjee
Federal Reserve Bank of Philadelphia, Philadelphia, PA, United States

H. Cole
University of Pennsylvania, Philadelphia, PA, United States

P. D'Erasmo
Federal Reserve Bank of Philadelphia, Philadelphia, PA, United States

D.W. Diamond
University of Chicago Booth School of Business, Chicago, IL; National Bureau of Economic Research, Cambridge, MA, United States

M. Doepke
Northwestern University, Evanston, IL, United States

E. Farhi
Harvard University, Cambridge, MA, United States

J. Fernández-Villaverde
University of Pennsylvania, Philadelphia, PA, United States

N. Fuchs-Schündeln
Goethe University Frankfurt, Frankfurt, Germany; CEPR, London, United Kingdom

M. Gertler
NYU, New York, NY; Princeton University, Princeton, NJ; Federal Reserve Board of Governors, Washington, DC, United States

M. Golosov
Princeton University, Princeton, NJ, United States

V. Guerrieri
University of Chicago, Chicago, IL; NBER, Cambridge, MA, United States

R.E. Hall
Hoover Institution, Stanford University, CA; National Bureau of Economic Research, Cambridge, MA, United States

J.D. Hamilton
University of California, San Diego, La Jolla, CA, United States

G.D. Hansen
UCLA, Los Angeles, CA; NBER, Cambridge, MA, United States

L.P. Hansen
University of Chicago, Chicago, IL; NBER, Cambridge, MA, United States

T.A. Hassan
CEPR, London, United Kingdom; University of Chicago, Chicago, IL; NBER, Cambridge, MA, United States

J. Hassler
Institute for International Economic Studies (IIES), Stockholm University, Stockholm; University of Gothenburg, Gothenburg, Sweden; CEPR, London, United Kingdom

C.L. House
NBER, Cambridge, MA; University of Michigan, Ann Arbor, MI, United States

E. Hurst
The University of Chicago Booth School of Business, Chicago, IL, United States

C.I. Jones
Stanford GSB, Stanford, CA; NBER, Cambridge, MA, United States

A.K. Kashyap
University of Chicago Booth School of Business, Chicago, IL; National Bureau of Economic Research, Cambridge, MA, United States

P.J. Kehoe
University of Minnesota; Federal Reserve Bank of Minneapolis, Minneapolis, MN, United States; University College London, London, United Kingdom

N. Kiyotaki
NYU, New York, NY; Princeton University, Princeton, NJ; Federal Reserve Board of Governors, Washington, DC, United States

D. Krueger
University of Pennsylvania, Philadelphia, PA, United States; CEPR, London, United Kingdom; CFS, Goethe University Frankfurt, Frankfurt, Germany; NBER, Cambridge, MA, United States; Netspar, Tilburg, The Netherlands

P. Krusell
Institute for International Economic Studies (IIES), Stockholm University, Stockholm; University of Gothenburg, Gothenburg, Sweden; CEPR, London, United Kingdom; NBER, Cambridge, MA, United States

M. Kuete
IMFS, Goethe University Frankfurt, Frankfurt, Germany

E.M. Leeper
Indiana University, IN; NBER, Cambridge, MA, United States

C. Leith
University of Glasgow, Glasgow, United Kingdom

C. Lian
MIT, Cambridge, MA, United States

J. Lindé
Sveriges Riksbank; Stockholm School of Economics, Stockholm, Sweden; CEPR, London, United Kingdom

E. McGrattan
University of Minnesota; Federal Reserve Bank of Minneapolis, Minneapolis, MN, United States

C.M. Meissner
NBER, Cambridge, MA; University of California, Davis, CA, United States

E.G. Mendoza
PIER, University of Pennsylvania, Philadelphia, PA; NBER, Cambridge, MA, United States

A. Mian
Princeton University, Princeton, NJ; NBER, Cambridge, MA, United States

K. Mitman
CEPR, London, United Kingdom; IIES, Stockholm University, Stockholm, Sweden

L.E. Ohanian
UCLA, Los Angeles; NBER, Cambridge, MA; Hoover Institution, Stanford University, Stanford, CA, United States

A. Passalacqua
Harvard University, Cambridge, MA, United States

F. Perri
CEPR, London, United Kingdom; Federal Reserve Bank of Minneapolis, Minneapolis, MN, United States

M. Piazzesi
Stanford University, Stanford, CA; NBER, Cambridge, MA, United States

E.C. Prescott
Arizona State University, Tempe, AZ; Federal Reserve Bank of Minneapolis, Minneapolis, MN, United States

A. Prestipino
NYU, New York, NY; Princeton University, Princeton, NJ; Federal Reserve Board of Governors, Washington, DC, United States

V.A. Ramey
University of California, San Diego, CA; NBER, Cambridge, MA, United States

J.F. Rubio-Ramírez
Emory University; Federal Reserve Bank of Atlanta, Atlanta, GA, United States; BBVA Research, Madrid, Madrid, Spain; Fulcrum Asset Management, London, England, United Kingdom

Y. Sannikov
Princeton University, Princeton, NJ, United States

M. Schneider
Stanford University, Stanford, CA; NBER, Cambridge, MA, United States

F. Schorfheide
University of Pennsylvania, Philadelphia, PA, United States

F. Smets
ECB, Frankfurt, Germany; KU Leuven, Leuven, Belgium; CEPR, London, United Kingdom

A.A. Smith, Jr.
NBER, Cambridge, MA; Yale University, New Haven, CT, United States

Z. Stangebye
University of Notre Dame, Notre Dame, IN, United States

J.H. Stock
Harvard University; The National Bureau of Economic Research, Cambridge, MA, United States

A. Sufi
University of Chicago Booth School of Business, Chicago, IL; NBER, Cambridge, MA, United States

J.B. Taylor
Stanford University, Stanford, CA, United States

M. Tertilt
University of Mannheim, Mannheim, Germany

A. Tsyvinski
Yale University, New Haven, CT, United States

H. Uhlig
University of Chicago, Chicago, IL; NBER, Cambridge, MA, United States; CEPR, London, United Kingdom

M.W. Watson
The Woodrow Wilson School, Princeton University, Princeton, NJ; The National Bureau of Economic Research, Cambridge, MA, United States

I. Werning
MIT, Cambridge, MA, United States

N. Werquin
Toulouse School of Economics, Toulouse, France

V. Wieland
IMFS, Goethe University Frankfurt, Frankfurt, Germany

R. Wouters
National Bank of Belgium, Brussels, Belgium; CEPR, London, United Kingdom

J. Yoo
IMFS, Goethe University Frankfurt, Frankfurt, Germany; Bank of Korea, Seoul, South Korea

J. Zhang
Federal Reserve Bank of Chicago, Chicago, IL, United States

Participants-Chicago

Left to right

Last row: Jesper Linde, John Heaton, Basu (slightly stepping forward), Zighuo He, Thorsten Drautzburg, Jonas Fischer, Andrea Prestipino, Jarda Borovicka, Mathias Trabandt

Second-from-last: Manuel Amador, Marty Eichenbaum (slightly stepping forward), Guido Lorenzoni, Luigi Boccola, Satyajit Chatterjee, Campbell Leith, Lawrence Christiano, Christopher House

Forth row: Marios Angeletos, Hal Cole, Markus Brunnermeier, Lars Hansen, Jeff Campbell

Third Row: Volker Wieland, Douglas Diamond, Anil Kashyap, Rüdiger Bachmann, Harald Uhlig, Nobu Kiyotaki

Second row: Frank Smets, Michele Tertilt, Eric Leeper, Jing Zhang, Enrique Mendoza

First row: Mathias Doepke, Veronica Guerrieri, Amir Sufi, Michael Weber, John Taylor

Not Pictured:

Authors: Ivan Werning, Christopher House, Erik Hurst, Raf Wouters, Yuliy Sannikov, Alberto Alesina, Andrea Passalacqua

Discussants: Stavros Panageas, Eric Sims, Thibaut Lamadon, Alessandra Voena, Kinda Cheryl Hachem, Casey Mulligan, Alp Simsek

Participants-Stanford

Left to right

Last row: Fabrizio Perri, Pablo Kurlat, Mikhail Golosov, Patrick Kehoe, Gary Hansen, Tarek Hassan, Pete Klenow, Christopher Meissner, Frank Schorfheide, Lee Ohanian (somewhat behind), Robert Hodrick

Middle row: Kurt Mitman, Dirk Krueger (slightly behind), Tony Smith, Harald Uhlig (slightly behind), Nicolas Werquin, Mark Watson (somewhat in front), Nicola Fuchs-Schündeln, Ellen McGrattan, Valerie Ramey (somewhat in front), Sebastian DiTella (to the right/behind Valerie Ramey), Pedro Brinca (somewhat in front), Charles Kolstad, Bob Hall, John Hassler, Carl Walsh

Front row: Chad Jones, Per Krusell, Jim Hamilton, Jim Stock, John Taylor, Steve Davis, Ed Prescott, Michael Bordo, Chari, Oscar Jorda, Serguei Maliar, Amir Kermani

Not Pictured:

Authors: Monika Piazzesi, Martin Schneider, Jesus Fernandez-Villaverde, Juan Rubio-Ramirez, Aleh Tsyvinski

Discussants: Bart Hobijn, Pierre Siklos, Arvin Krishnamurthy, Christopher Tonetti, Yuriy Gorodnichenko, John Cochrane, Michael Bauer, Cosmin Ilut

PREFACE

This *Handbook* aims to survey the state of knowledge and major advances during the past two decades in the field of macroeconomics. It covers empirical, theoretical, methodological, and policy issues, including fiscal, monetary, and regulatory policies to deal with unemployment, economic growth, and crises, taking account of research developments before, during, and after the global financial crisis of 2007–2009. It can serve as a textbook and as an introduction to frontier research.

THE STATE OF MACRO, THE FINANCIAL CRISIS, AND NEW CURRENTS

The *Handbook* displays an amazing range of new and different ideas. There are neoclassical chapters on real business cycles and there are new Keynesian chapters on monetary business cycles. There are also chapters extending well beyond traditional macro, including the macroeconomics of the family, natural experiments, environmental issues, time allocation, and the fast moving areas of the connection between financial and real factors, incomplete markets, incomplete contracts, heterogeneous agents, and recursive contracts. There are also treatments of macroprudential policies, the impact of fiscal policy at the zero lower bound on interest rates, the fiscal theory of the price level, and the political economy of bailouts and debt. And there are chapters essential for research on the latest estimation and solution techniques (in continuous and discrete time), as well as encyclopedic reviews of the key facts of economic growth and economic fluctuations both at the aggregate and individual level.

A widely debated question for macroeconomics is whether the 2007–2009 financial crisis demonstrated a failure of the field or whether there was a failure of policy to follow the advice implied by the field. The chapters in the *Handbook* written by active and experienced researchers in macroeconomics can help answer that question in ways that informal policy debates cannot, and we hope that this is an important contribution of the *Handbook*.

There is no question that the field of macroeconomics has continued to progress enormously since the advent of rational expectations, microeconomic foundations, dynamic optimization, and general equilibrium models. Using this paradigm macroeconomists—before and after the financial crisis—have been able to introduce real-world rigidities in price setting, learning, incomplete markets, and financial frictions.

Since the global financial crisis and the Great Recession, some view a lack of financial frictions in macroeconomic models as an indication of failure, and of course there is much in this new *Handbook* on financial frictions and the financial sector more generally in macro models. But the 1999 *Handbook* already included work on financial frictions as evidenced by the chapter written by Ben Bernanke, Mark Gertler, and Simon Gilchrist. And an important finding reported in the chapter in this *Handbook* by Jesper Linde, Frank Smets, and Raf Wouters is that when more financial factors are added to macro models used at central banks, they do not help that much in explaining the financial crisis.

SUMMARY

The 33 chapters of the *Handbook* are divided into five sections. Each chapter starts with a short summary written by its authors, and reading these is the best way to understand what is in the *Handbook*. This short summary of the whole book shows how the chapters are organized and fit together.

Section 1, *The Facts of Economic Growth and Economic Fluctuation*, starts off with examination of the fundamental facts upon which macroeconomic theories are built and with which they must be consistent. It covers both the long run—going back 100 years—and the short run—tracing how shocks impact and propagate over time and how changes in policy regimes or rules affect economic fluctuations. Emphasizing microeconomic underpinnings, the chapters in this section look at the time allocation by people and families, the impact of longer decisions take on debt or purchases houses, the way wage decisions affect the allocation of labor, and the historical impact of financial and fiscal crises.

Section 2 focuses *The Methodology of Macroeconomics*. It covers factor models, structural VARs, solution methods, estimation of DSGE models, recursive contracts, endogenously incomplete markets, heterogeneous agents, natural experiments, the use of "wedges" as accounting framework for business cycle models, incomplete information, coordination frictions, and comprehensive methods of comparing models and achieving robustness.

Section 3, *Financial-Real Connections*, covers bank runs, the real effects of financial crises, credit markets, booms and busts, the central role of the housing market, and quantitative models of sovereign debt crises. It also shows different ways to connect the real and the financial sector including through continuous-time methods and models of the term structure of uncertainty.

Section 4, *Models of Economic Growth and Fluctuations*, covers several approaches to modeling the economy, including neoclassical or real business cycle models and staggered wage and price models or other rigidities that can explain slow recoveries and long slums. It takes a macroeconomic perspective on environmental issues as well as family decisions.

Section 5, *Macroeconomic Policy*, contains a thorough review of models used by central banks for conducting monetary policy, the analysis of regulatory policy including liquidity requirements, the fiscal theory of the price level, fiscal multipliers, liquidity traps, currency unions, and the technical sustainability vs the political economy of government debt.

John B. Taylor
Harald Uhlig

ACKNOWLEDGMENTS

First, we would like to acknowledge the initial encouragement from Kenneth Arrow and Michael Intriligator who, as Co-Editors of the *Handbook of Economics* Series, suggested in May 2007—nearly a decade ago and before the Global Financial Crisis—the need for a new *Handbook of Macroeconomics*. We are also grateful to Scott Bentley, Elsevier Acquisition Editor, who worked closely with us to get helpful comments and reviews of our initial outline, to Michael Woodford, who succeeded Mike Intriligator as Co-Editor of the Series for his helpful suggestions on organization and coverage, and to Joslyn Chaiprasert-Paguio, Editorial Project Manager, who brought the project to completion.

We—along with all the economists and students who will benefit from this volume—owe a great deal of thanks to the 74 accomplished economists who agreed to contribute to this endeavor for the time and creativity they invested in their chapters. We thank them for attending the conferences, presenting draft chapters, and then revising the chapters under strict deadlines. We particularly thank the discussants of the draft chapters whose careful comments helped improve the chapters and the overall book itself, including Manuel Amador, Ruediger Bachmann, Bob Barsky, Michael Bauer, Luigi Bocola, Jeff Campbell, John Cochrane, Sebastian Di Tella, Thorsten Drautzberg, Jonas Fisher, Yuriy Gorodnichenko, Kinda Cheryl Hachem, Zhiguo He, John Heaton, Bart Hobijn, Robert Hodrick, Cosmin Ilut, Oscar Jorda, Amir Kermani, Peter Klenow, Charles Kolstad, Arvind Krishnamurthy, Pablo Kurlat, Thibaut Lamadon, Guido Lorenzoni, Serguei Maliar, Casey Mulligan, Stavros Panageas, Pierre Siklos, Eric Sims, Alp Simsek, Christopher Tonetti, Alessandra Voena, Carl Walsh, and Michael Weber.

The two conferences at Stanford and Chicago where authors presented draft chapters and received this critical commentary were essential to the completion of the *Handbook of Macroeconomics*. We thank the Hoover Institution at Stanford University and the Becker-Friedman Institute at the University of Chicago for the financial and logistical support that made these conferences possible. We also wish to thank Marie-Christine Slakey for managing the overall manuscript and communications with authors from the very beginning of the project.

John B. Taylor
Harald Uhlig

Volume 2A

SECTION 1

The Facts of Economic Growth and Economic Fluctuation

CHAPTER 1

The Facts of Economic Growth

C.I. Jones
Stanford GSB, Stanford, CA, United States
NBER, Cambridge, MA, United States

Contents

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.002

Abstract

Why are people in the richest countries of the world so much richer today than 100 years ago? And why are some countries so much richer than others? Questions such as these define the field of economic growth. This paper documents the facts that underlie these questions. How much richer are we today than 100 years ago, and how large are the income gaps between countries? The purpose of the paper is to provide an encyclopedia of the fundamental facts of economic growth upon which our theories are built, gathering them together in one place and updating them with the latest available data.

Keywords

Economic growth, Development, Long-run growth, Productivity

JEL Classification Codes

E01, O10, 04

> *"[T]he errors which arise from the absence of facts are far more numerous and more durable than those which result from unsound reasoning respecting true data."—Charles Babbage, quoted in (Rosenberg, 1994, p. 27).*
>
> *"[I]t is quite wrong to try founding a theory on observable magnitudes alone… It is the theory which decides what we can observe."—Albert Einstein, quoted in (Heisenberg, 1971, p. 63).*

Why are people in the United States, Germany, and Japan so much richer today than 100 or 1000 years ago? Why are people in France and the Netherlands today so much richer than people in Haiti and Kenya? Questions like these are at the heart of the study of economic growth.

Economics seeks to answer these questions by building quantitative models—models that can be compared with empirical data. That is, we'd like our models to tell us not only that one country will be richer than another, but by how much. Or to explain not only that we should be richer today than a century ago, but that the growth rate should be 2% per year rather than 10%. Growth economics has only partially achieved these goals, but a critical input into our analysis is knowing where the goalposts lie—that is, knowing the facts of economic growth.

The purpose of this paper is to lay out as many of these facts as possible. Kaldor (1961) was content with documenting a few key stylized facts that basic growth theory should hope to explain. Jones and Romer (2010) updated his list to reflect what we've learned over the last 50 years. The approach here is different. Rather than highlighting a handful of stylized facts, we draw on the last 30 years of the renaissance of growth economics to lay out what is known empirically about the subject. These facts are updated with the latest data and gathered together in a single place—potentially useful to newcomers to the field as well as to experts. The result, I hope, is a fascinating tour of the growth literature from the perspective of the basic data.

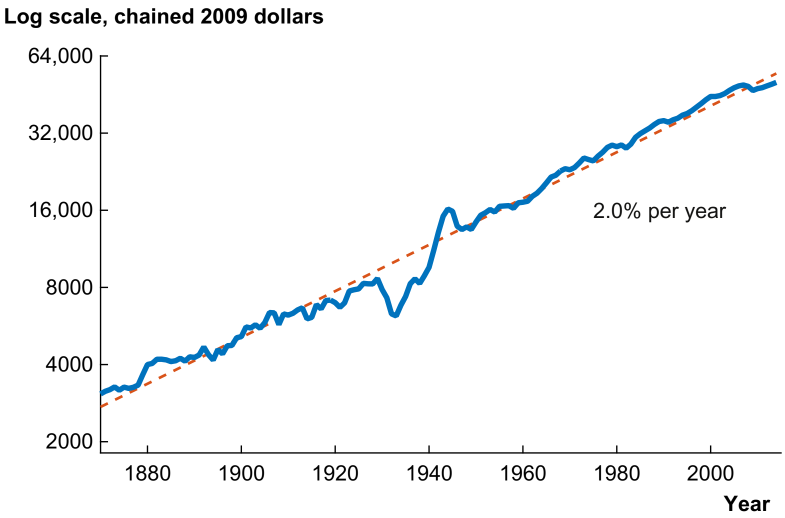


Fig. 1 GDP per person in the United States. Source: *Data for 1929–2014 are from the U.S. Bureau of Economic Analysis, NIPA table 7.1. Data before 1929 are spliced from Maddison, A. 2008. Statistics on world population, GDP and per capita GDP, 1-2006 AD. Downloaded on December 4, 2008 from http://www.ggdc.net/maddison/.*

The paper is divided broadly into two parts. First, I present the facts related to the growth of the "frontier" over time: what are the growth patterns exhibited by the richest countries in the world? Second, I focus on the spread of economic growth throughout the world. To what extent are countries behind the frontier catching up, falling behind, or staying in place? And what characteristics do countries in these various groups share?

1. GROWTH AT THE FRONTIER

We begin by discussing economic growth at the "frontier." By this I mean growth among the richest set of countries in any given time period. For much of the last century, the United States has served as a stand in for the frontier, and we will follow this tradition.

1.1 Modern Economic Growth

Fig. 1 shows one of the key stylized facts of frontier growth: For nearly 150 years, GDP per person in the US economy has grown at a remarkably steady average rate of around 2% per year. Starting at around $3,000 in 1870, per capita GDP rose to more than $50,000 by 2014, a nearly 17-fold increase.

Beyond the large, sustained growth in living standards, several other features of this graph stand out. One is the significant decline in income associated with the Great

Table 1 The stability of US Growth

Period	Growth Rate	Period	Growth Rate
1870–2007	2.03	1973–1995	1.82
1870–1929	1.76	1995–2007	2.13
1929–2007	2.23		
1900–1950	2.06	1995–2001	2.55
1950–2007	2.16	2001–2007	1.72
1950–1973	2.50		
1973–2007	1.93		

Note: Annualized growth rates for the data shown in Fig. 1.

Depression. However, to me this decline stands out most for how anomalous it is. Many of the other recessions barely make an impression on the eye: over long periods of time, economic growth swamps economic fluctuations. Moreover, despite the singular severity of the Great Depression—GDP per person fell by nearly 20% in just 4 years—it is equally remarkable that the Great Depression was *temporary*. By 1939, the economy is already passing its previous peak and the macroeconomic story a decade later is once again one of sustained, almost relentless, economic growth.

The stability of US growth also merits some discussion. With the aid of the trend line in Fig. 1, one can see that growth was slightly slower pre-1929 than post. Table 1 makes this point more precisely. Between 1870 and 1929, growth averaged 1.76%, vs 2.23% between 1929 and 2007 (using "peak to peak" dates to avoid business cycle problems). Alternatively, between 1900 and 1950, growth averaged 2.06% vs 2.16% since 1950. Before one is too quick to conclude that growth rates are increasing; however, notice that the period since 1950 shows a more mixed pattern, with rapid growth between 1950 and 1973, slower growth between 1973 and 1995, and then rapid growth during the late 1990s that gives way to slower growth more recently.

The interesting "trees" that one sees in Table 1 serves to support the main point one gets from looking at the "forest" in Fig. 1: steady, sustained exponential growth for the last 150 years is a key characteristic of the frontier. All modern theories of economic growth—for example, Solow (1956), Lucas (1988), Romer (1990), and Aghion and Howitt (1992)—are designed with this fact in mind.

The sustained growth in Fig. 1 also naturally raises the question of whether such growth can and will continue for the next century. On the one hand, this fact more than any other helps justify the focus of many growth models on the balanced growth path, a situation in which all economic variables grow at constant exponential rates forever. And the logic of the balanced growth path suggests that the growth can continue indefinitely. On the other hand, as we will see, there are reasons from other facts and theories to question this logic.

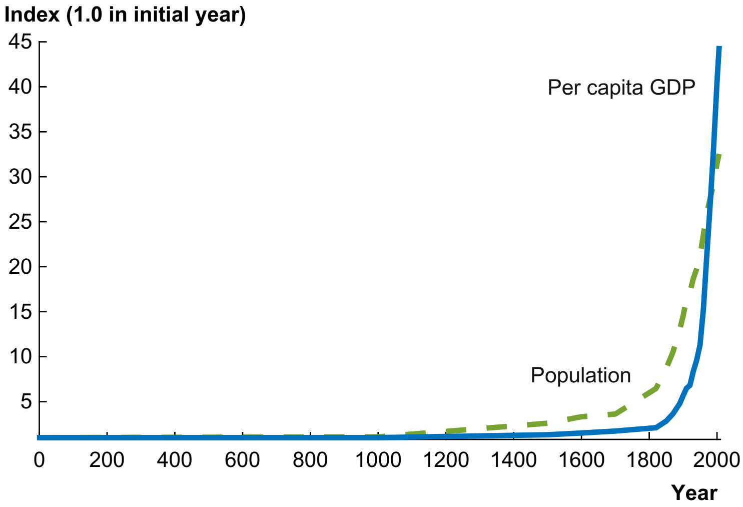


Fig. 2 Economic growth over the very long run. Source: *Data are from Maddison, A. 2008. Statistics on world population, GDP and per capita GDP, 1-2006 AD. Downloaded on December 4, 2008 from http://www.ggdc.net/maddison/ for the "West," ie, Western Europe plus the United States. A similar pattern holds using the "world" numbers from Maddison.*

1.2 Growth Over the Very Long Run

While the future of frontier growth is surely hard to know, the stability of frontier growth suggested by Fig. 1 is most certainly misleading as a guide to growth further back in history. Fig. 2 shows that sustained exponential growth in living standards is an incredibly recent phenomenon. For thousands and thousands of years, life was, in the evocative language of Thomas Hobbes, "nasty, brutish, and short." Only in the last two centuries has this changed, but in this relatively brief time, the change has been dramatic.[a]

Between the year 1 C.E. and the year 1820, living standards in the "West" (measured with data from Western Europe and the United States) essentially doubled, from around $600 per person to around $1200 per person, as shown in Table 2. Over the next 200 years; however, GDP per person rose by more than a factor of twenty, reaching $26,000.

The era of modern economic growth is in fact even more special than this. Evidence suggests that living standards were comparatively stagnant for thousands and thousands of years before. For example, for much of prehistory, humans lived as simple hunters and gatherers, not far above subsistence. From this perspective—say for the last 200,000 years

[a] Papers that played a key role in documenting and elaborating upon this fact include Maddison (1979), Kremer (1993), Maddison (1995), Diamond (1997), Pritchett (1997), and Clark (2001). This list neglects a long, important literature in economic history; see Clark (2014) for a more complete list of references.

Table 2 The Acceleration of world growth

Year	GDP per person	Growth rate	Population (millions)	Growth rate
1	590	–	19	–
1000	420	−0.03	21	0.01
1500	780	0.12	50	0.17
1820	1240	0.15	125	0.28
1900	3350	1.24	280	1.01
2006	26,200	1.94	627	0.76

Note: Growth rates are average annual growth rates in percent, and GDP per person is measured in real 1990 dollars.
Source: Data are from Maddison, A. 2008. Statistics on world population, GDP and per capita GDP, 1-2006 AD. Downloaded on December 4, 2008 from http://www.ggdc.net/maddison/ for the "West," ie, Western Europe plus the United States

or more—the era of modern growth is spectacularly brief. It is the economic equivalent of Carl Sagan's famous "pale blue dot" image of the earth viewed from the outer edge of the solar system.

Table 2 reveals several other interesting facts. First and foremost, over the very long run, economic growth at the frontier has accelerated—that is, the rates of economic growth are themselves increasing over time. Romer (1986) emphasized this fact for living standards as part of his early motivation for endogenous growth models. Kremer (1993) highlighted the acceleration in population growth rates, dating as far back as a million years ago, and his evidence serves as a very useful reminder. Between 1 million B.C.E. and 10,000 B.C.E., the average population growth rate in Kremer's data was 0.00035% per year. Yet despite this tiny growth rate, world population increased by a factor of 32, from around 125,000 people to 4 million. As an interesting comparison, that's similar to the proportionate increase in the population in Western Europe and the United States during the past 2000 years, shown in Table 2.

Various growth models have been developed to explain the transition from stagnant living standards for thousands of years to the modern era of economic growth. A key ingredient in nearly all of these models is Malthusian diminishing returns. In particular, there is assumed to be a fixed supply of land which is a necessary input in production.[b] Adding more people to the land reduces the marginal product of labor (holding technology constant) and therefore reduces living standards. Combined with some subsistence level of consumption below which people cannot survive, this ties the size of the population to the level of technology in the economy: a better technology can support a larger population.

[b] I have used this assumption in my models as well, but I have to admit that an alternative reading of history justifies the exact opposite assumption: up until very recently, land was completely elastic—whenever we needed more, we spread out and found greener pastures.

Various models then combine the Malthusian channel with different mechanisms for generating growth. Lee (1988), Kremer (1993), and Jones (2001) emphasize the positive feedback loop between "people produce ideas" as in the Romer model of growth with the Malthusian "ideas produce people" channel. Provided the increasing returns associated with ideas is sufficiently strong to counter the Malthusian diminishing returns, this mechanism can give rise to dynamics like those shown in Fig. 2. Lucas (2002) emphasizes the role of human capital accumulation, while Hansen and Prescott (2002) focus on a neoclassical model that features a structural transformation from agriculture to manufacturing. Oded Galor, with his coauthors, has been one of the most significant contributors, labeling this literature "unified growth theory." See Galor and Weil (2000) and Galor (2005).

2. SOURCES OF FRONTIER GROWTH

The next collection of facts related to economic growth are best presented in the context of the famous growth accounting decomposition developed by Solow (1957) and others. This exercise studies the sources of growth in the economy through the lens of a single aggregate production function. It is well known that the conditions for an aggregate production function to exist in an environment with a rich underlying microstructure are very stringent. The point is not that anyone believes those conditions hold. Instead, one often wishes to look at the data "through the lens of" some growth model that is much simpler than the world that generates the observed data. A long list of famous papers supports the claim that this is a productive approach to gaining knowledge, Solow (1957) itself being an obvious example.

While not necessary, it is convenient to explain this accounting using a Cobb–Douglas specification. More specifically, suppose final output Y_t is produced using stocks of physical capital K_t and human capital H_t:

$$Y_t = \underbrace{A_t M_t}_{\text{TFP}} K_t^{\alpha} H_t^{1-\alpha} \tag{1}$$

where α is between zero and one, A_t denotes the economy's stock of knowledge, and M_t is anything else that influences total factor productivity (the letter "M" is reminiscent of the "measure of our ignorance" label applied to the residual by Abramovitz (1956) and also is suggestive of "misallocation," as will be discussed in more detail later). The next subsection provides a general overview of growth accounting for the United States based on this equation, and then the remainder of this section looks more closely at each individual term in Eq. (1).

2.1 Growth Accounting

It is traditional to perform the growth accounting exercise with a production function like (1). However, that approach creates some confusion in that some of the accumulation of physical capital is caused by growth in total factor productivity (eg, as in a standard Solow model). If one wishes to credit such growth to total factor productivity, it is helpful to do the accounting in a slightly different way.[c] In particular, divide both sides of the production function by Y_t^{α} and solve for Y_t to get

$$Y_t = \left(\frac{K_t}{Y_t}\right)^{\frac{\alpha}{1-\alpha}} H_t Z_t \tag{2}$$

where $Z_t \equiv (A_t M_t)^{\frac{1}{1-\alpha}}$ is total factor productivity measured in labor-augmenting units. Finally, dividing both sides by the aggregate amount of time worked, L_t, gives

$$\frac{Y_t}{L_t} = \left(\frac{K_t}{Y_t}\right)^{\frac{\alpha}{1-\alpha}} \frac{H_t}{L_t} \cdot Z_t \tag{3}$$

In this form, growth in output per hour Y_t/L_t comes from growth in the capital-output ratio K_t/Y_t, growth in human capital per hour H_t/L_t, and growth in labor-augmenting TFP, Z_t. This can be seen explicitly by taking logs and differencing Eq. (3). Also, notice that in a neoclassical growth model, the capital-output ratio is proportional to the investment rate in the long-run and does not depend on total factor productivity. Hence the contributions from productivity and capital deepening are separated in this version, in a way that they were not in Eq. (1).

The only term we have yet to comment on is H_t/L_t, the aggregate amount of human capital divided by total hours worked. In a simple model with one type of labor, one can think of $H_t = h_t L_t$, where h_t is human capital per worker which increases because of education. In a richer setting with different types of labor that are perfect substitutes when measured in efficiency units, H_t/L_t also captures composition effects. The Bureau of Labor Statistics, from which I've obtained the accounting numbers discussed next, therefore refers to this term as "labor composition."

Table 3 contains the growth accounting decomposition for the United States since 1948, corresponding to Eq. (3). Several well-known facts emerge from this accounting. First, growth in output per hour at 2.5% is slightly faster than the growth in GDP per person that we saw earlier. One reason is that the BLS data measure growth for the private business sector, excluding the government sector (in which there is zero productivity growth more or less by assumption). Second, the capital-output ratio is relatively stable over this period, contributing almost nothing to growth. Third, labor composition (a rise in educational attainment, a shift from manufacturing to services, and the increased labor force participation of women) contributes 0.3 percentage points per year to growth. Finally, as documented by Abramovitz, Solow, and others, the "residual" of total factor

[c] Klenow and Rodriguez-Clare (1997), for example, takes this approach.

Table 3 Growth accounting for the United States

Period	Output per hour	Contributions from		
		K/Y	Labor composition	Labor-Aug. TFP
1948–2013	**2.5**	**0.1**	**0.3**	**2.0**
1948–1973	3.3	−0.2	0.3	3.2
1973–1990	1.6	0.5	0.3	0.8
1990–1995	1.6	0.2	0.7	0.7
1995–2000	3.0	0.3	0.3	2.3
2000–2007	2.7	0.2	0.3	2.2
2007–2013	1.7	0.1	0.5	1.1

Note: Average annual growth rates (in percent) for output per hour and its components for the private business sector, following Eq. (3).
Source: Authors calculations using Bureau of Labor Statistics, *Multifactor Productivity Trends*, August 21, 2014.

productivity accounts for the bulk of growth, coming in at 2.0 percentage points, or 80% of growth since 1948.

The remainder of Table 3 shows the evolution of growth and its decomposition over various periods since 1948. We see the rapid growth and rapid TFP growth of the 1948–1973 period, followed by the well-known "productivity slowdown" from 1973 to 1995. The causes of this slowdown are much debated but not convincingly pinned down, as suggested by the fact that the entirety of the slowdown comes from the TFP residual rather than from physical or human capital; Griliches (1988) contains a discussion of the slowdown.

Remarkably, the period 1995–2007 sees a substantial recovery of growth, not quite to the rates seen in the 1950s and 1960s, but impressive nonetheless, coinciding with the dot-com boom and the rise in the importance of information technology. Byrne et al. (2013) provide a recent analysis of the importance of information technology to growth over this period and going forward. Lackluster growth in output per hour since 2007 is surely in large part attributable to the Great Recession, but the slowdown in TFP growth (which some such as Fernald, 2014 date back to 2003) is troubling.[d]

2.2 Physical Capital

The fact that the contribution of the capital-output ratio was modest in the growth accounting decomposition suggests that the capital-output ratio is relatively constant over

[d] There are a number of important applications of growth accounting in recent decades. Young (1992) and Young (1995) document the surprisingly slow total factor productivity growth in the East Asian miracle countries. Krugman (1994) puts Young's accounting in context and relates it to the surprising finding of early growth accounting exercises that the Soviet Union exhibited slow TFP growth as well. Klenow and Rodriguez-Clare (1997) conduct a growth accounting exercise using large multicountry data sets and show the general importance of TFP growth in that setting.

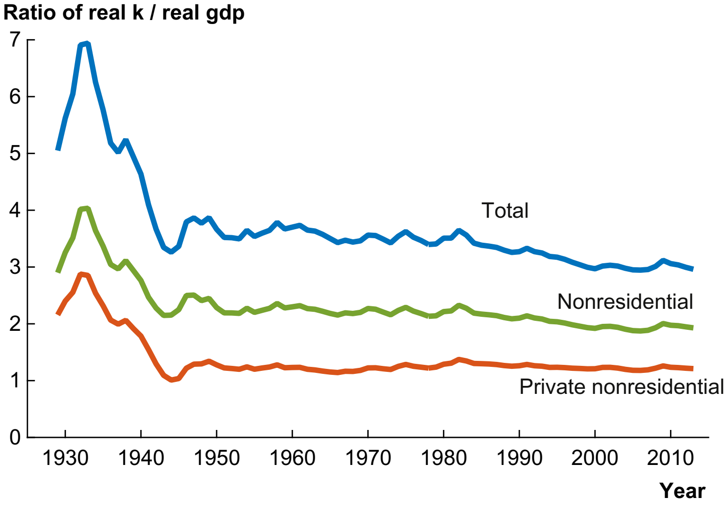


Fig. 3 The ratio of physical capital to GDP. Source: *Burea of Economic Analysis Fixed Assets tables 1.1 and 1.2. The numerator in each case is a different measure of the real stock of physical capital, while the denominator is real GDP.*

time. This suggestion is confirmed in Fig. 3. The broadest concept of physical capital (Total), including both public and private capital as well as both residential and nonresidential capital, has a ratio of 3 to real GDP. Focusing on nonresidential capital brings this ratio down to 2, and further restricting to private nonresidential capital leads a ratio of just over 1.

The capital stock is itself the cumulation of investment, adjusted for depreciation. Fig. 4 shows nominal spending on investment as a share of GDP back to 1929. The share is relatively stable for much of the period, with a notable decline during the last two decades.

In addition to cumulating investment, however, another step in going from the (nominal) investment rate series to the (real) capital-output ratio involves adjusting for relative prices. Fig. 5 shows the price of various categories of investment, relative to the GDP deflator. Two facts stand out: the relative price of equipment has fallen sharply since 1960 by more than a factor of 3 and the relative price of structures has risen since 1929 by a factor of 2 (for residential) or 3 (for nonresidential).

A fascinating observation comes from comparing the trends in the relative prices shown in Fig. 5 to the investment shares in Fig. 4: the nominal investment shares are relatively stable when compared to the huge trends in relative prices. For example, even though the relative price of equipment has fallen by more than a factor of 3 since 1960, the nominal share of GDP spent on equipment has remained steady.

The fall of equipment prices has featured prominently in parts of the growth literature; for example, see Greenwood et al. (1997) and Whelan (2003). These papers make the point that one way to reconcile the facts is with a two-sector model in which

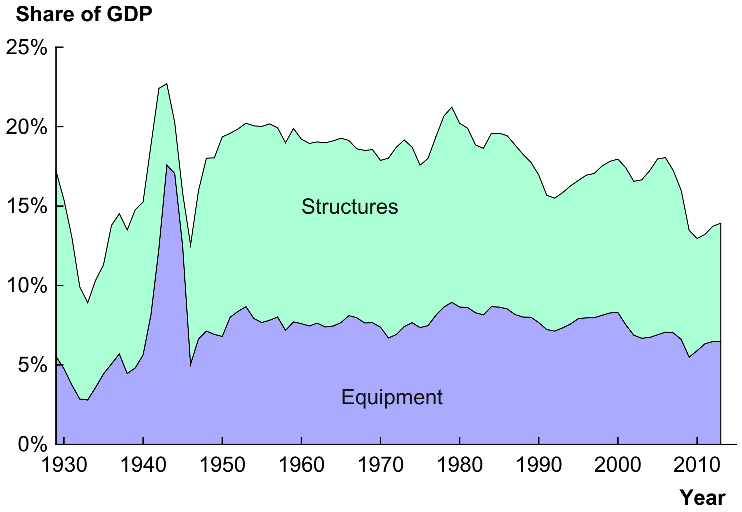


Fig. 4 Investment in physical capital (private and public), United States. Source: *National Income and Product Accounts, U.S. Bureau of Economic Analysis, table 5.2.5. Intellectual property products and inventories are excluded. Government and private investment are combined. Structures includes both residential and nonresidential investment. Ratios of nominal investment to GDP are shown.*

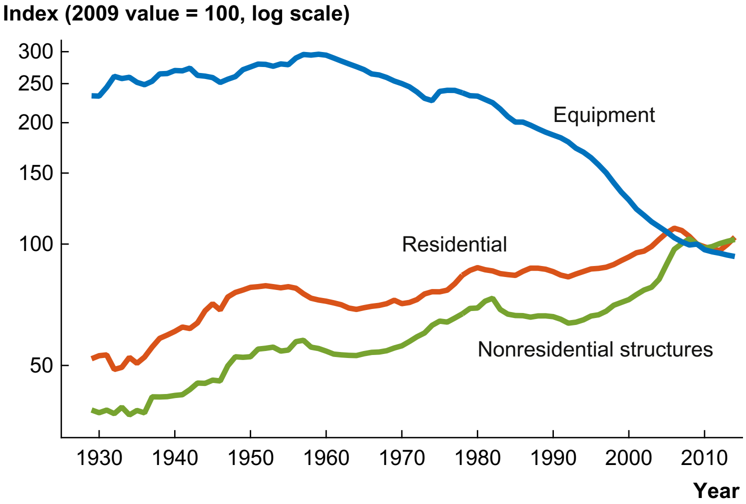


Fig. 5 Relative price of investment, United States. *Note*: The chained price index for various categories of private investment is divided by the chained price index for GDP. Source: *National Income and Product Accounts, U.S. Bureau of Economic Analysis table 1.1.4.*

technological progress in the equipment sector is substantially faster that technological progress in the rest of the economy—an assumption that rings true in light of Moore's Law and the tremendous decline in the price of a semiconductors. Combining this assumption with Cobb–Douglas production functions leads to a two-sector model that

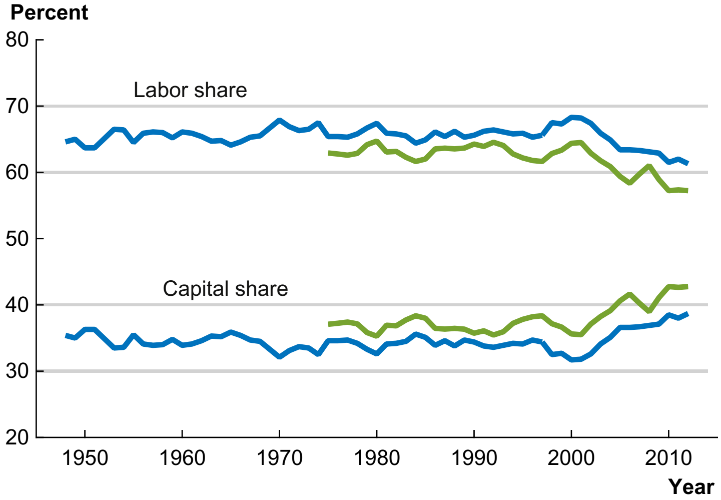


Fig. 6 Capital and labor shares of factor payments, United States. Source: *The series starting in 1975 are from Karabarbounis, L., Neiman, B. 2014. The global decline of the labor share. Q. J. Econ. 129 (1), 61–103. http://ideas.repec.org/a/oup/qjecon/v129y2014i1p61-103.html and measure the factor shares for the corporate sector, which the authors argue is helpful in eliminating issues related to self-employment. The series starting in 1948 is from the Bureau of Labor Statistics* Multifactor Productivity Trends, *August 21, 2014, for the private business sector. The factor shares add to 100%.*

is broadly consistent with the facts we've laid out. A key assumption in this approach is that better computers are equivalent to having more of the old computers, so that technological change is, at least partially, capital (equipment) augmenting. The Cobb–Douglas assumption ensures that this nonlabor augmenting technological change can coexist with a balanced growth path and delivers a stable nominal investment rate.[e]

2.3 Factor Shares

One of the original Kaldor (1961) stylized facts of growth was the stability of the shares of GDP paid to capital and labor. Fig. 6 shows these shares using two different data sets, but the patterns are quite similar. First, between 1948 and 2000, the factor shares were indeed quite stable. Second, since 2000 or so, there has been a marked decline in the labor share and a corresponding rise in the capital share. According to the data from the Bureau of Labor Statistics, the capital share rose from an average value of 34.2% between 1948 and 2000 to a value of 38.7% by 2012. Or in terms of the complement, the labor share declined from an average value of 65.8% to 61.3%.

[e] This discussion is related to the famous Uzawa theorem about the restrictions on technical change required to obtain balanced growth; see Schlicht (2006) and Jones and Scrimgeour (2008).

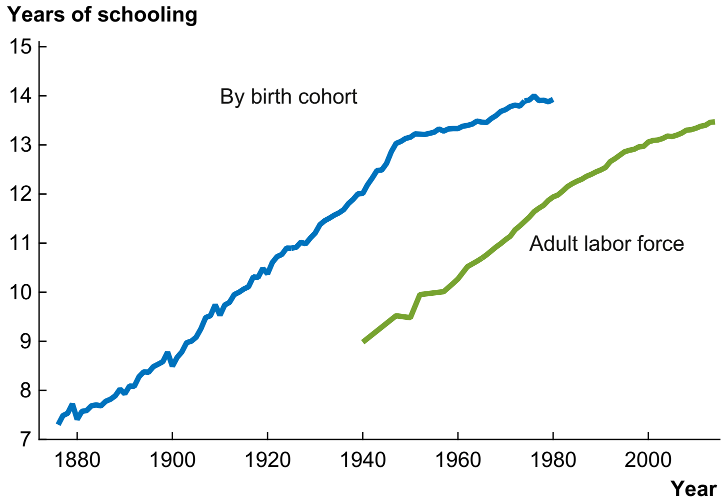


Fig. 7 Educational attainment, United States. Source: *The* blue (dark gray *in the print version) line shows educational attainment by birth cohort from Goldin, C., Katz, L.F. 2007. Long-run changes in the wage structure: narrowing, widening, polarizing. Brook. Pap. Econ. Act. 2, 135–165. The* green (gray *in the print version) line shows average educational attainment for the labor force aged 25 and over from the Current Population Survey.*

It is hard to know what to make of the recent movements in factor shares. Is this a temporary phenomenon, perhaps amplified by the Great Recession? Or are some more deeper structural factors at work? Karabarbounis and Neiman (2014) document that the fact extends to many countries around the world and perhaps on average starts even before 2000. Other papers seek to explain the recent trend by studying depreciation, housing, and/or intellectual property and include Elsby et al. (2013), Bridgman (2014), Koh et al. (2015), and Rognlie (2015).

A closely-related fact is the pattern of factor shares exhibited across industries within an economy and across countries. Jones (2003) noted the presence of large trends, both positive and negative, in the 35 industry (2-digit) breakdown of data in the United States from Dale Jorgenson. Gollin (2002) suggests that factor shares are uncorrelated with GDP per person across a large number of countries.

2.4 Human Capital

The other major neoclassical input in production is human capital. Fig. 7 shows a time series for one of the key forms of human capital in the economy, education. More specifically, the graph shows educational attainment by birth cohort, starting with the cohort born in 1875.

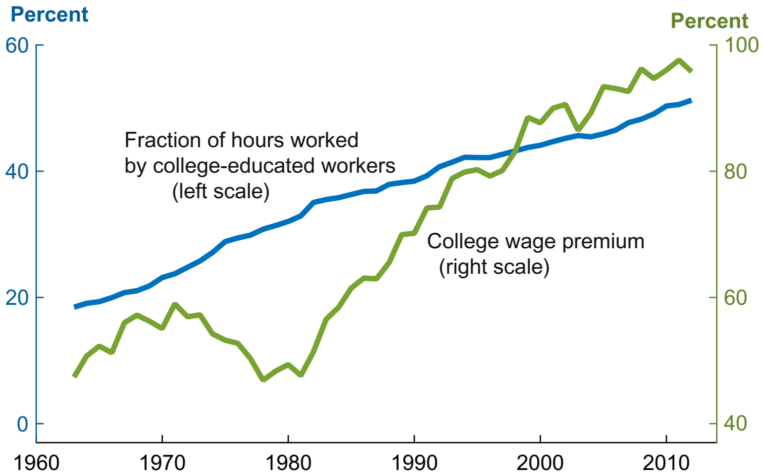


Fig. 8 The supply of college graduates and the college wage premium, 1963–2012. *Note:* The supply of US college graduates, measured by their share of total hours worked, has risen from below 20% to more than 50% by 2012. The US college wage premium is calculated as the average excess amount earned by college graduates relative to nongraduates, controlling for experience and gender composition within each educational group. Source: *Autor, D.H. 2014. Skills, education, and the rise of earnings inequality among the "other 99 percent". Science 344 (6186), 843–851, fig. 3.*

Two facts emerge. First, for 75 years, educational attainment rose steadily, at a rate of slightly less than 1 year per decade. For example, the cohort born in 1880 got just over 7 years of education, while the cohort born in 1950 received 13 years of education on average. As shown in the second (green) line in the figure, this translated into steadily rising educational attainment in the adult labor force. Between 1940 and 1980, for example, educational attainment rose from 9 years to 12 years, or about 3/4 of a year per decade. With a Mincerian return to education of 7%, this corresponds to a contribution of about 0.5 percentage points per year to growth in output per worker.

The other fact that stands out prominently, however, is the leveling-off of educational attainment. For cohorts born after 1950, educational attainment rose more slowly than before, and for the latest cohorts, educational attainment has essentially flattened out. Over time, one expects this to translate into a slowdown in the increase of educational attainment for the labor force as a whole, and some of this can perhaps be seen in the last decade of the graph.

Fig. 8 shows another collection of stylized facts related to human capital made famous by Katz and Murphy (1992). The blue line in the graph shows the fraction of hours worked in the US economy accounted for by college-educated workers. This fraction rose from less than 20% in 1963 to more than 50% by 2012. The figure also shows the college wage premium, that is the excess amount earned by college graduates over nongraduates after controlling for experience and gender. This wage premium averaged around 50% between 1963 and the early 1980s but then rose sharply through 2012 to peak at nearly 100%. Thus, even though the supply of college graduates was growing rapidly, the wage premium for college graduates was increasing sharply as well.

Katz and Murphy (1992) provide an elegant way to understand the dynamics of the college wage premium. Letting "coll" and "hs" denote two kinds of labor ("college graduates" and "high school graduates"), the human capital aggregate that enters production is given by a CES specification:

$$H = ((A_{coll}L_{coll})^{\rho} + (A_{hs}L_{hs})^{\rho})^{1/\rho} \tag{4}$$

An increase in the supply of college graduates lowers their marginal product, while an increase in the technology parameter A_{coll} raises their marginal product. Katz and Murphy (1992) show that with an elasticity of substitution of around 1.4, a constant growth rate of A_{coll}/A_{hs}, which Katz and Murphy call "skill-biased technical change," together with the observed movements in L_{coll}/L_{hs} can explain the time series for the college wage premium.

Human capital includes more than just education, of course. Workers continue to accumulate skills on the job. This human capital shows up as higher wages for workers, but separating this into a quantity of human capital and a price per unit of human capital requires work. One simple approach is to assume each year of work experience leads to a constant increase in human capital, and this approach is commonly pursued in growth accounting. Examples of richer efforts to measure human capital in a growth setting include Lucas (2009), Erosa et al. (2010), Lucas and Moll (2014), and Manuelli and Seshadri (2014).

2.5 Ideas

Our next set of facts relate to the economy's stock of knowledge or ideas, the A in the production function that we began with back in Eq. (1). It has long been recognized that the "idea production function" is hard to measure. Where do ideas come from? Part of the difficulty is that the answer is surely multidimensional. Ideas are themselves very heterogeneous, some clearly arise through intentional research, but others seem to arrive by chance out of seemingly nowhere. Confronted with these difficulties, Solow (1956) modeled technological change as purely exogenous, but this surely goes too far. The more people there are searching for new ideas, the more likely it is that discoveries will be made. This is true if the searching is intentional, as in research, but even if it is a byproduct of the production process itself as in models of learning by doing. The production of new ideas plays a fundamental role in the modern understanding of growth; see Romer (1990), Grossman and Helpman (1991), and Aghion and Howitt (1992).[f]

[f] Various perspectives on the idea production function are presented by Mokyr (1990), Griliches (1994), Weitzman (1998), and Fernald and Jones (2014).

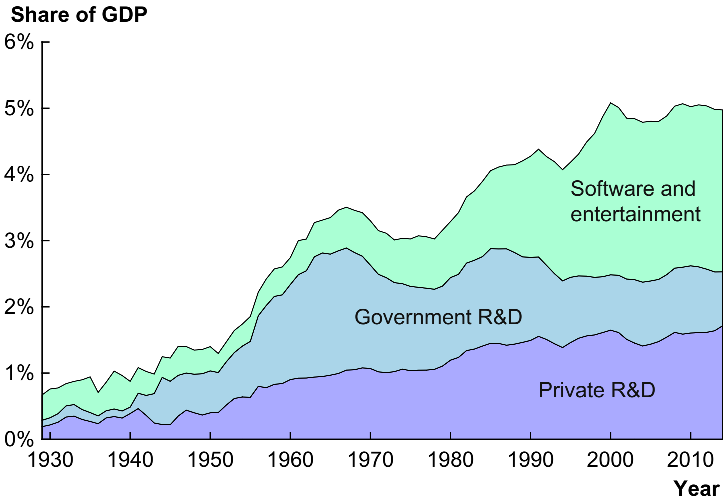


Fig. 9 Research and development spending, United States. Source: *National Income and Product Accounts, U.S. Bureau of Economic Analysis via FRED database. "Software and entertainment" combines both private and public spending. "Entertainment" includes movies, TV shows, books, and music.*

With this in mind, Fig. 9 shows spending on research and development, as a share of GDP, for the United States. These data can now be obtained directly from the National Income and Product Accounts, thanks to the latest revisions by the Bureau of Economic Analysis. The broadest measure of investment in ideas recorded by the NIPA is investment in "intellectual property products." This category includes traditional research and development, spending on computer software, and finally spending on "entertainment," which itself includes movies, TV shows, books, and music.

Several facts stand out in Fig. 9. First, total spending on investment in intellectual property products has risen from less than 1% of GDP in 1929 to nearly 5% of GDP in recent years. This overall increase reflects a large rise in private research and development and a large rise in software and entertainment investment, especially during the last 25 years. Finally, government spending on research and development has been shrinking as a share of GDP since peaking in the 1960s with the space program.

Fig. 10 provides an alternative perspective on R&D in two dimensions. First, it focuses on employment rather than dollars spent, and second it brings in an international perspective. The figure shows the number of researchers in the economy as a share of the population.[g]

[g] According to the OECD's Frascati Manual 2002, p. 93, researchers are defined as "professionals engaged in the conception or creation of new knowledge, products, processes, methods and systems and also in the management of the projects concerned."

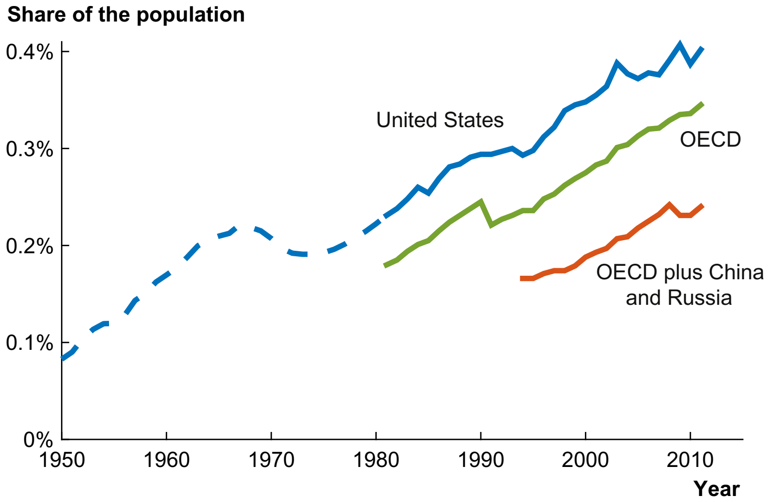


Fig. 10 Research employment share. Source: *Data for 1981–2001 are from OECD Main Science and Technology Indicators, http://stats.oecd.org/Index.aspx?DataSetCode=MSTI_PUB. Data prior to 1981 for the United States are spliced from Jones, C.I. 2002. Sources of U.S. economic growth in a world of ideas. Am. Econ. Rev. 92 (1), 220–239, which uses the NSF's definition of "scientists and engineers engaged in R&D."*

Each of the three measures in the figure tells the same story: the fraction of the population engaged in R&D has been rising in recent decades. This is true within the United States, within the OECD, and even if we incorporate China and Russia as well.

It is important to appreciate a significant limitation of the R&D data shown so far. In particular, these data only capture a small part of what an economist would call research. For example, around 70% of measured R&D occurs in the manufacturing industry. In 2012, only 18 million workers (out of US employment that exceeds 130 million) were employed by firms that conducted any official R&D.[h] According to their corporate filings, Walmart and Goldman-Sachs report doing zero R&D.

So far, we have considered the input side of the idea production function. We now turn to the output side. Unfortunately, the output of ideas is even harder to measure than the inputs. One of the more commonly-used measures is patents, and this measure is shown in Fig. 11.

On first glance, it appears that patents, like many other variables reviewed in this essay, have grown exponentially. Indeed, at least since 1980 one sees a very dramatic rise in the number of patents granted in the United States, both in total and to US inventors. The difference between these two lines—foreign patenting in the United States—is also interesting, and one testament to the global nature of ideas is that 56% of patents granted by the US patent office in 2013 were to foreigners.

[h] These numbers are from Wolfe (2014).

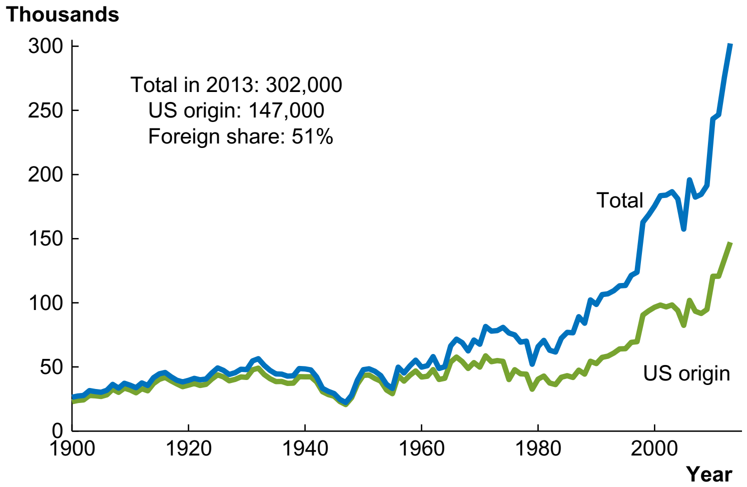


Fig. 11 Patents granted by the US Patent and Trademark Office. Source: *http://www.uspto.gov/web/offices/ac/ido/oeip/taf/h_counts.htm*.

A closer look at Fig. 11, though, reveals something equally interesting: the number of patents granted to US inventors in 1915, 1950, and 1985 was approximately the same. Put another way, during the first 85 years of the 20th century, the number of patents granted to US residents appears to be stationary, in sharp contrast to the dramatic increase since 1985 or so. Part of the increase since the 1980s is due to changes in patent policy, including extending patent protection to software and business models and changes in the judicial appeals process for patent cases (Jaffe and Lerner, 2006).

Griliches (1994) combined these two basic facts related to ideas (rapid growth in the inputs, stable production of patents) to generate a key implication: the productivity of research at producing patents fell sharply for most of the 20th century. Kortum (1997) developed a growth model designed to match these facts in which he emphasized that patents can be thought of as *proportional* improvements in productivity. If each patent raises GDP by a constant percent, then a constant flow of new patents can generate a constant rate of economic growth. The problem with this approach (or perhaps the problem with the patent data) is that it breaks down after 1980 or so. Since 1980, the number of patents has risen by more than a factor of four, while growth rates are more or less stable. The bottom line is that the idea production function remains something of a black box perhaps precisely because we do not have great measures of ideas or the inputs used to produce them.[i]

[i] Examples of progress include Caballero and Jaffe (1993), Acemoglu et al. (2013), L. (2013), and Akcigit et al. (2014b).

2.6 Misallocation

The organizing principle for this section of the paper is the production function given back in Eq. (1). In specifying that production function, I broke total factor productivity into two pieces: the stock of ideas, A, and everything else, which I labeled "M" either for the "measure of our ignorance" or for "misallocation." It is this latter interpretation that I wish to take up now.

One of the great insights of the growth literature in the last 15 years is that misallocation at the micro level can show up as a reduction in total factor productivity at a more aggregated level. This insight appears in various places, including Banerjee and Duflo (2005), Chari et al. (2007), Restuccia and Rogerson (2008) and Hsieh and Klenow (2009).

The essence of the insight is quite straightforward: when resources are allocated optimally, the economy will operate on its production possibilities frontier. When resources are misallocated, the economy will operate inside this frontier. But that is just another way of saying that TFP will be lower: a given quantity of inputs will produce less output.

As we explain in detail in the second part of this paper (in Section 4.7), there is a large literature on misallocation and development—this is our best candidate answer to the question of why are some countries so much richer than others. There is much less discussion of the extent to which misallocation is related to frontier growth, the subject at hand.

While it is clear conceptually that even the country or countries at the frontier of growth can suffer from misallocation and that changes in misallocation can contribute to growth, there has been little work quantifying this channel. Indeed, my own working hypothesis for many years was that this effect was likely small in the United States during the last 50 years. I now believe this is wrong.

Hsieh et al. (2013) highlight a striking fact that illustrates this point: in 1960, 94% of doctors and lawyers were white men; by 2008, this fraction was just 62%. Given that innate talent for these and other highly-skilled professions is unlikely to differ across groups, the occupational distribution in 1960 suggests that a large number of innately talented African Americans and white women were not working in the occupations dictated by comparative advantage. The paper quantifies the macroeconomic consequences of the remarkable convergence in the occupational distribution between 1960 and 2008 and finds that 15–20% of growth in aggregate output per worker is explained by the improved allocation of talent. In other words, declines in misallocation may explain a significant part of US economic growth during the last 50 years.

Examples to drive home these statistics are also striking. Sandra Day O'Connor—future Supreme Court Justice—graduated third in her class from Stanford Law School in 1952. But the only private sector job she could get upon graduation was as a legal secretary (Biskupic, 2006). Closer to our own profession, David Blackwell, of contraction mapping fame, was the first African American inducted into the National Academy of Sciences and the first tenured at U.C. Berkeley. Yet despite getting his Ph.D. at age 22 and obtaining a postdoc at the Institute for Advanced Studies in 1941, he was not permitted to attend lectures at Princeton and was denied employment at U.C. Berkeley for racial reasons. He worked at Howard University until 1954, when he was finally hired as a full professor in the newly-created statistics department at Berkeley.[j]

Another potential source of misallocation is related to the economics of ideas. It has long been suggested that knowledge spillovers are quite significant, both within and across countries. To the extent that these spillovers are increasingly internalized or addressed by policy—or to the extent that the opposite is true—the changing misallocation of knowledge resources may be impacting economic growth.[k]

As one final example, Hsieh and Moretti (2014) suggest that spatial misallocation within the United States may be significant. Why is it that Sand Hill Road in Palo Alto has Manhattan rents without the skyscrapers? Hsieh and Moretti argue that land use policies prevent the efficient spatial matching of people to land and to each other. They estimate that places like Silicon Valley and New York City would be four to eight times more populous in the efficient allocation.

Quantifying these and other types of misallocation affecting frontier growth is a fruitful direction for future research.

2.7 Explaining the Facts of Frontier Growth

While this essay is primarily about the facts of economic growth, it is helpful to step back and comment briefly on how multiple facts have been incorporated into our models of growth.

The basic neoclassical growth framework of Solow (1956) and Ramsey (1928) / Cass (1965) / Koopmans (1965) has long served as a benchmark organizing framework for understanding the facts of growth. The nonrivalry of ideas, emphasized by Romer (1990), helps us understand how sustained exponential growth occurs endogenously. I review this contribution and some of the extensive research it sparked in Jones (2005).[l]

The decline in the relative price of equipment and the rise in the college wage premium are looked at together in Krusell et al. (2000). That paper considers a setting in which equipment capital is complementary to skilled labor, so that the (technologically driven) decline in the price of equipment is the force of skill-biased technological change. That paper uses a general CES structure. One of the potential issues in that paper was that it could lead to movements in the labor share. But perhaps we are starting to see those in the data.

[j] See http://en.wikipedia.org/wiki/David_Blackwell. I'm grateful to Ed Prescott for this example.

[k] For evidence on knowledge spillovers, see Griliches (1992), Coe and Helpman (1995), Jones and Williams (1998), Klenow and Rodriguez-Clare (2005), and Bloom et al. (2013).

[l] Romer's insights are expanded upon in various directions. Aghion and Howitt (1992) and Grossman and Helpman (1991) emphasize the important role of creative destruction. Jones (1995), Kortum (1997), and Segerstrom (1998) clarify the way in which nonrivalry interacts with population growth to explain sustained growth in living standards.

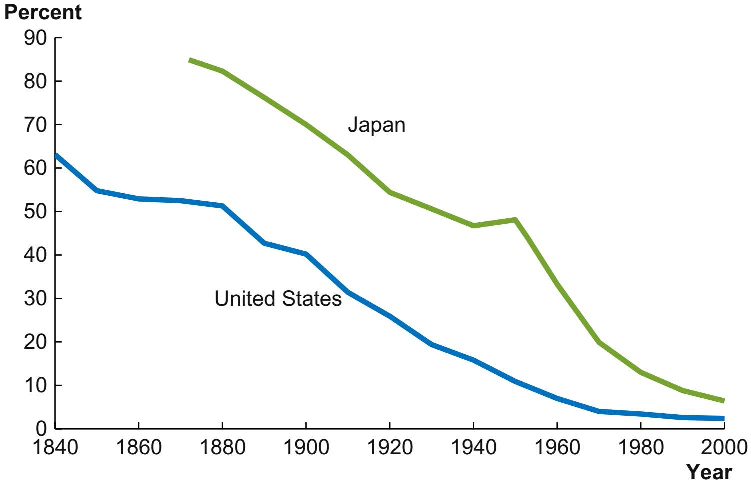


Fig. 12 Employment in agriculture as a share of total employment. Source: *Herrendorf, B., Rogerson, R., Valentinyi, A. 2014, Growth and structural transformation, In: Handbook of Economic Growth, vol. 2, Elsevier, pp. 855–941, http://ideas.repec.org/h/eee/grochp/2-855.html.*

The presence of trends in educational attainment and research investment opens up interesting opportunities for future research. Why are educational attainment and the share of labor devoted to research rising over time? What are the implications of these trends for future growth? Restuccia and Vandenbroucke (2013) suggest that skill-biased technological change is itself responsible for driving the rise in educational attainment. Acemoglu (1998) examines the further interactions when the direction of technological change is itself endogenous. Jones (2002) considers the implication of the trends in education and research intensity for future growth, suggesting that these trends have substantially raised growth during the last 50 years above the economy's long-run growth rate.

3. FRONTIER GROWTH: BEYOND GDP

The next collection of facts related to frontier growth look beyond the aggregate of GDP. These facts are related to structural change (the decline of agriculture and the rise of services, especially health), changes in leisure and fertility, rising inequality, and falling commodity prices.

3.1 Structural Change

Fig. 12 shows one of the most dramatic structural changes affecting frontier economies over the last 200 years and beyond: the decline of agriculture. In 1840, about two out of every three workers in the US economy worked in agriculture. By 2000, this share had fallen to just 2.4%. Similar changes can be seen in value-added in agriculture as a share of GDP as well as in other countries. For example, the chart also shows agriculture's share of employment in Japan, declining from 85% around 1870.[m]

The structural transformation has several other dimensions and connections in the growth literature. For example, the decline in agriculture is first associated with a rise in manufacturing, which is ultimately replaced by a rise in services, including health and education; more on this below.

Another form of structural transformation that has seen renewed interest is the possibility that machines (capital) may substitute for labor. Autor et al. (2003) look at detailed occupational classifications to study the impact of computerization on labor demand. They emphasize a polarization, with computerization being particularly substitutable for routine cognitive tasks that can be broken into specific rules but complementary to nonroutine, cognitive tasks. That is, computers substitute for bank tellers and low-level secretaries, while increasing the demand for computer programmers and leaving untouched manual jobs like janitorial work. Brynjolfsson and McAfee (2012) highlight broader ramifications of artificial intelligence, whereby computers might start driving cars, reading medical tests, and combing through troves of legal documents. That is, even many tasks thought to be cognitive and not easily routinized may be subject to computerization. What impact will such changes have on the labor market?

The answer to this question is obviously complicated and the subject of ongoing research.[n] One useful reference point is the enormous transformation that occurred as the agricultural share of the US labor force went from 2/3 to only 2%, largely because of mechanization and technological change. There is no doubt that this had a transformative affect on the labor market, but by and large this transformation was overwhelmingly beneficial. That's not to say that it must be that way in the future, but the example is surely worth bearing in mind.

3.2 The Rise of Health

A different structural transformation has been predominant during the last 50 years: the rise of health spending as a share of GDP. Fig. 13 shows this fact for the United States and for several other OECD countries. In the United States, the health share more than tripled since 1960, rising from 5% in 1960 to 17% in recent years. Large trends are

[m] The literature on structural transformation and economic growth is surveyed by Herrendorf et al. (2014). More recent contributions include Boppart (2014) and Comin et al. (2015), who emphasize demand systems with heterogeneous income effects.

[n] For some examples, see Acemoglu (1998), Zeira (1998), Caselli (1999), Hemous and Olsen (2014).

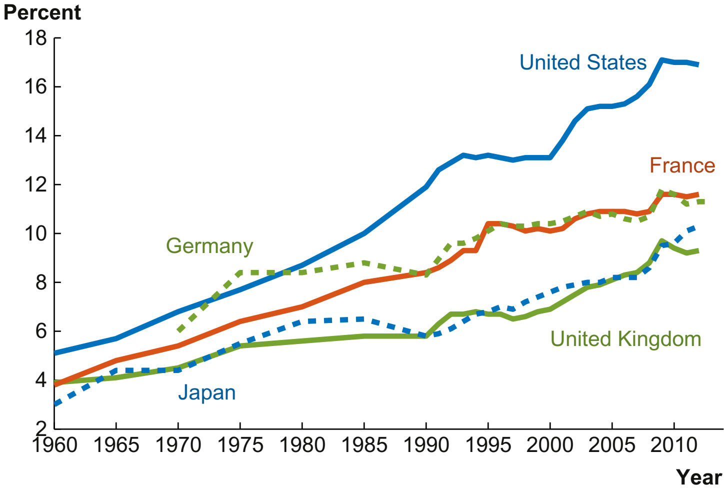


Fig. 13 Health spending as a share of GDP. Source: *OECD Health Statistics, 2014.*

present in other countries as well, with the share in France, for example, rising from under 4% to nearly 12%.

Hall and Jones (2007) propose that the widespread rise in the prominence of health care is a byproduct of economic growth. With standard preferences, the marginal utility of consumption declines rapidly. This is most easily seen for CRRA preferences in which the intertemporal elasticity of substitution is below one, in which case flow utility is bounded. As we get richer and richer, the marginal utility of consumption on any given day declines rapidly; what people really need are more days of life to enjoy their high level of consumption. Hence there is an income effect tilting spending toward life-saving categories.

One of the few time series related to economic growth that does not grow exponentially is life expectancy, where the increases tend to be arithmetic rather than exponential. Fig. 14 shows life expectancy at birth and at age 65 in the United States. Life expectancy at birth increased rapidly in the first half of the 20th century, thanks to improvements in public health and large declines in infant mortality. Since 1950, the rate of improvement has been more modest, around 1.8 years per decade. The figure also shows that the rise in life expectancy occurs at old ages. Life expectancy conditional on reaching age 65 has risen by just under 1 year per decade since 1950.[o] Interestingly,

[o] Nordhaus (2003) and Murphy and Topel (2006) discuss the rise in life expectancy and the economic returns to reducing mortality in more detail. Oeppen and Vaupel (2002) suggest that "record life expectancy" (ie, the maximum life expectancy across countries) has grown linearly at 2.5 years per decade for more than 150 years.

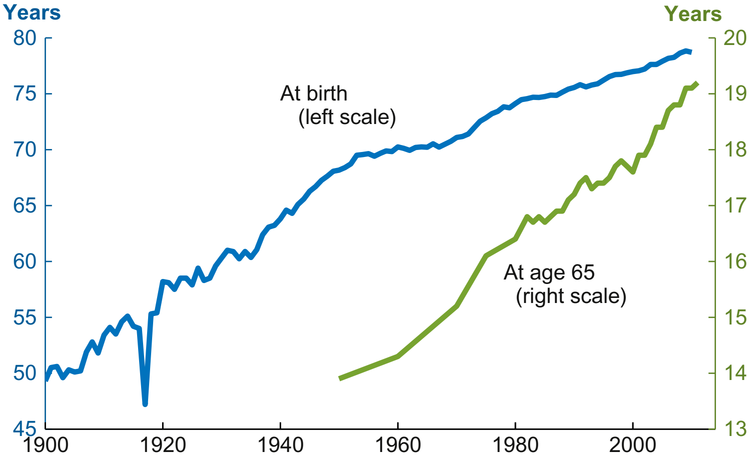


Fig. 14 Life expectancy at birth and at age 65, United States. Source: *Health, United States 2013 and https://www.clio-infra.eu*.

the mortality rate itself seems to grow exponentially with age, a phenomenon known as the Gompertz–Makeham Law; see Dalgaard and Strulik (2014).

3.3 Hours Worked and Leisure

A standard stylized fact in macroeconomics is that the fraction of the time spent working shows no trend despite the large upward trend in wages. The next two figures show that this stylized fact is not really true over the longer term, although the evidence is somewhat nuanced.

Fig. 15 shows average annual hours worked per person engaged in work from the Penn World Tables, which takes its data in turn from the Total Economy Database of the Conference Board. Among advanced countries, annual hours worked has fallen significantly since 1950. Average hours worked in the United States, for example, fell from 1909 in 1950 to 1704 in 2011. In France, the decline is even more dramatic, from 2159 to 1476. The decline starts slightly later in Japan after their recovery from World War II, with hours falling from 2222 in 1960 to 1706 in 2011.

Fig. 16 breaks the US evidence down into more detail, courtesy of Ramey and Francis (2009). First, the figure shows the split between men and women. Average weekly hours of market work by men fell sharply between 1900 and 1980, before leveling off. In contrast, market work by women has been on an upward trend. Ramey and Francis (2009) also use time diaries to estimate home production, and this is where the story gets more complicated. As men are substituting away from market work, they are also substituting into home production. Home production by men rose from just 4 h per week in 1900 to more than 16 h per week in 2005. The increase in leisure, then, was much smaller than the decline in market hours suggests.

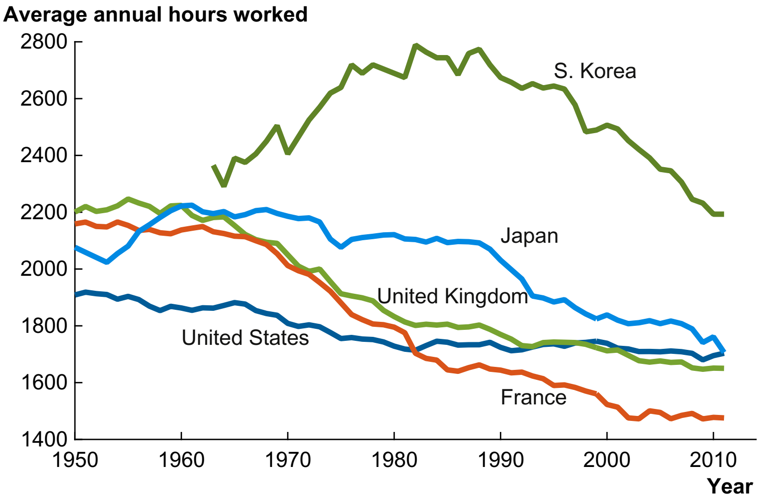


Fig. 15 Average annual hours worked, select countries. Source: *Average annual hours worked per person employed, from the Penn World Tables 8.0. See Feenstra, R.C., Inklaar, R., Timmer, M.P. 2015. The next generation of the Penn World Table. Am. Econ. Rev. 105 (10), 3150–3182. doi:10.1257/aer.20130954 and their excellent data appendix for details on the data.*

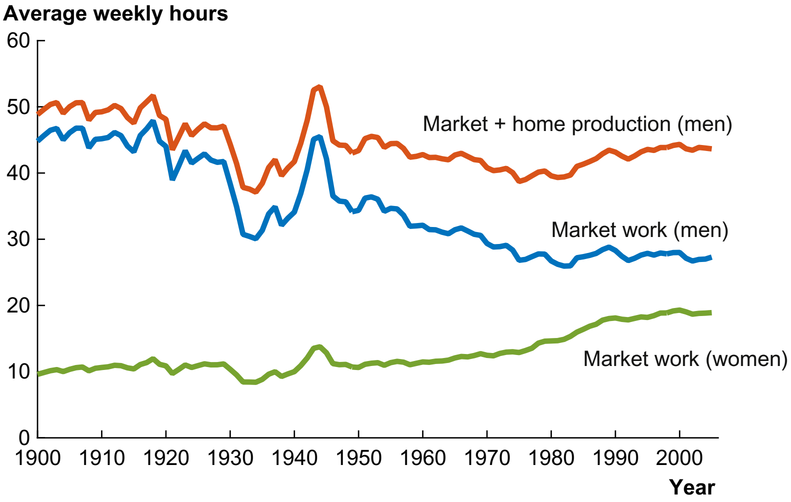


Fig. 16 Average weekly hours worked, United States. Source: *Average weekly hours per worker, from Ramey, V.A., Francis, N. 2009. A century of work and leisure. Am. Econ. J. Macroecon. 1 (2), 189–224. http://ideas.repec.org/a/aea/aejmac/v1y2009i2p189-224.html.*

3.4 Fertility

The facts we have presented thus far in this section—the decline in agriculture and the rise in services like health, the rise in life expectancy, the decline in hours worked—are all consistent with a particular kind of income effect. As people get richer, the marginal

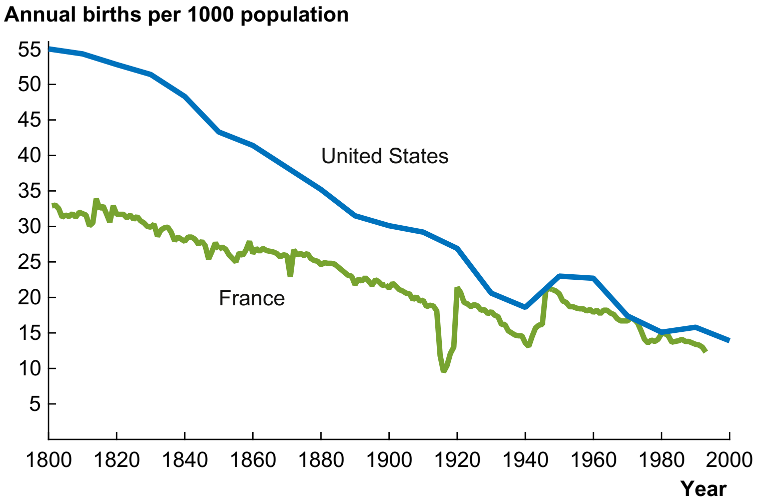


Fig. 17 Fertility in the United States and France. Source: *Data for the United States are from Haines, M. 2008, Fertility and mortality in the United States. In: Whaples, R., (Ed.), EH.Net Encyclopedia, http://eh.net/encyclopedia/fertility-andmortality-in-the-united-states/ and data for France are from Greenwood, J., Vandenbroucke, G. 2004. The baby boom and baby bust: O.E.C.D. fertility data. http://guillaumevdb.net/BabyBoom-data.pdf*.

utility of consumption falls and people substitute away from consumption and toward actions that conserve on the precious time endowment. Time is the one thing that technological progress cannot create!

The next fact on fertility raises interesting questions about this hypothesis. In particular, Fig. 17 shows the large decline in fertility dating back at least to 1800, known as the demographic transition. Since 1800, the birth rate has fallen from 5.5% in the United States and 3.3% in France down to less than 1.5% in recent years.

In dynastic models like Barro and Becker (1989), in which having more children is essentially a way of increasing one's own effective lifetime or time endowment, there is a force that tends to raise fertility, at least if income effects dominate substitution effects. But instead, we see strong declines in fertility in the data. A large literature seeks to understand the declines in fertility and the hump-shape in population growth that are together known as the demographic transition. A key part of the standard explanation is that children are themselves time intensive, in which case conserving on children also conserves on time as people get richer.[p]

[p] For example, see Galor and Weil (1996), Doepke (2005), Greenwood et al. (2005), Jones et al. (2010), Cordoba and Ripoll (2014), and Jones and Tertilt (forthcoming).

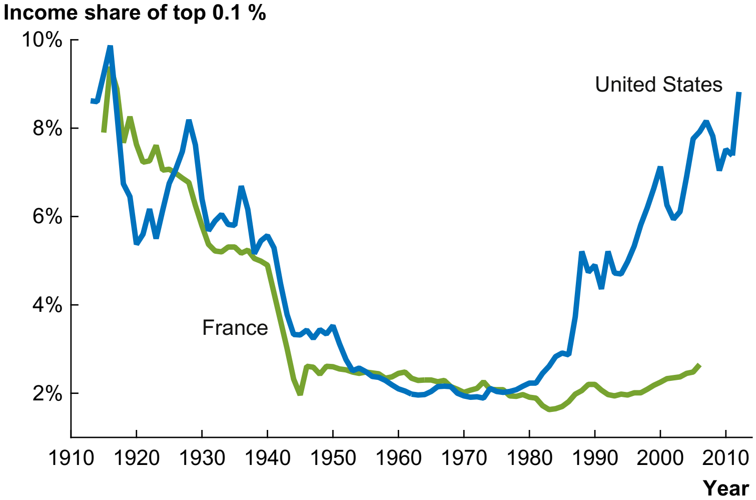


Fig. 18 Top income inequality in the United States and France. Source: *Alvaredo, F., Atkinson, A.B., Piketty, T., Saez, E. 2013. The World Top Incomes Database. Accessed on October 15, 2013, http://topincomes.g-mond.parisschoolofeconomics.eu/*.

3.5 Top Inequality

One of the more famous facts documented during the last decade is shown in Fig. 18. This is the top income inequality graph of Piketty and Saez (2003). In both the United States and France, the share of income earned by the top 0.1% of households was around 9% in 1920, and in both countries the share declined sharply until the 1950s to around 2%. It stayed at this low level until around 1980. But then a very large difference emerged, with top income shares rising in the United States to essentially the same level as in 1920, while the share in France remains relatively low. Much of the decline in the first part of the century is associated with capital income, and much of the rise in US inequality since 1980 is associated with labor (and business) income.[q]

It is also worth stepping back to appreciate the macroeconomic consequences of this inequality. Fig. 19 merges the Piketty–Saez top inequality data with the long-run data on GDP per person for the United States shown at the start of this paper in Fig. 1. In particular, the figure applies the Piketty–Saez inequality shares to average GDP per person to produce an estimate of GDP per person for the top 0.1% and the bottom 99.9%.[r]

[q] Possible explanations for this pattern are discussed by Castaneda et al. (2003), Cagetti and Nardi (2006), Atkinson et al. (2011), Benhabib et al. (2011), Aoki and Nirei (2013), Jones and Kim (2014), Piketty (2014), Piketty et al. (2014), and Saez and Zucman (2014).

[r] It is important to note that this estimate is surely imperfect. GDP likely does not follow precisely the same distribution as the Adjusted Gross Income data that forms the basis of the Piketty–Saez calculations: health benefits are more equally distributed, for example.

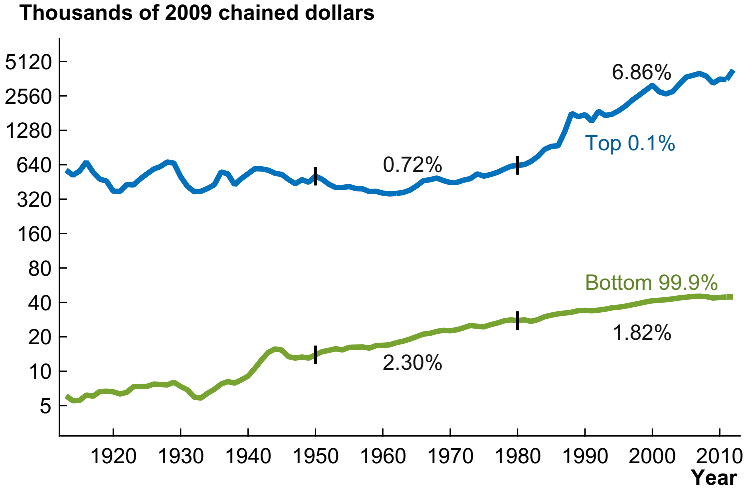


Fig. 19 GDP per person, top 0.1% and bottom 99.9%. *Note:* This figure displays an estimate of average GDP per person for the top 0.1% and the bottom 99.9%. Average annual growth rates for the periods 1950–1980 and 1980–2007 are also reported. Source: *Aggregate GDP per person data are from Fig. 1. The top income share used to divide the GDP is from the October 2013 version of the world top incomes database, from http://g-mond.parisschoolofeconomics.eu/topincomes/.*

Two key results stand out. First, until recently, there is surprisingly little growth in average GDP per person at the top: the value in 1977 is actually *lower* than the value in 1913. Instead, all the growth until around 1960 occurs in the bottom 99.9%. The second point is that this pattern changed in recent decades. For example, average growth in GDP per person for the bottom 99.9% declined by around half a percentage point, from 2.3% between 1950 and 1980 to only 1.8% between 1980 and 2007. In contrast, after being virtually absent for 50 years, growth at the top accelerated sharply: GDP per person for the top 0.1% exhibited growth more akin to China's economy, averaging 6.86% since 1980. Changes like this clearly have the potential to matter for economic welfare and merit the attention they've received.

3.6 The Price of Natural Resources

This next fact is very different from what we've been discussing, but it is one of the more surprising facts related to frontier growth. Fig. 20 shows the real price of industrial commodities, consisting of an equally-weighted basket of aluminum, coal, copper, lead, iron ore, and zinc, deflated by the consumer price index. During the 20th century, world demand for these industrial commodities exploded with the rise of the automobile, electrification, urbanization, and the general industrialization that occurred in the United States and around the world. The surprise shown in the figure is that the real price of these commodities *declined* over the 20th century. Moreover, the magnitude of the

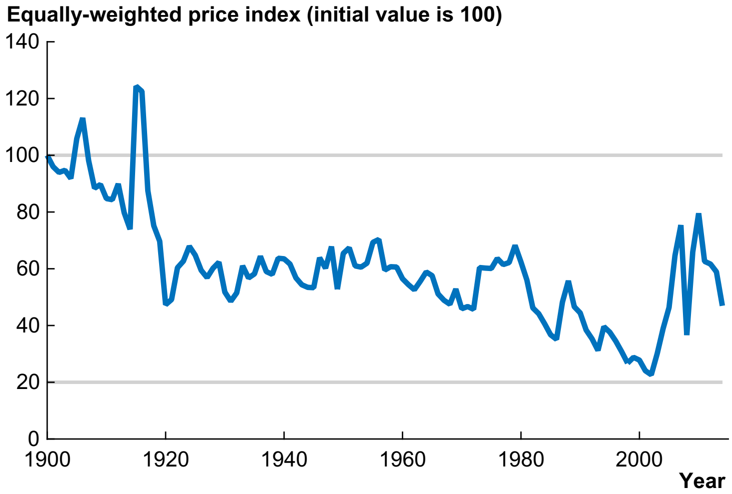


Fig. 20 The real price of industrial commodities. Source: *The price of an equally-weighted basket of aluminum, coal, copper, lead, iron ore, and zinc, deflated by the consumer price index. Commodity prices are from www.globalfinancialdata.com and the CPI is from www.measuringworth.com.*

decline was large—a factor of 5 between the year 1900 and 2000. Evidently, some combination of increased discoveries and technological changes led the effective supply to grow even faster than the enormous rise in demand.[s]

Also striking, though, is the large increase in the real price of these commodities since 2000. Part of the explanation could be the rapid growth of China and India over this period and the large increase in the demand for commodities that their growth entailed. Interestingly, we will see later that many developing countries performed quite well in the 2000s. Some of that growth contributed to the rise in demand for commodities, but some of that success may also reflect commodity-driven growth resulting from the rise in demand from China and India.

4. THE SPREAD OF ECONOMIC GROWTH

Up until now, we've been primarily concerned with the growth of the frontier: what are the facts about how the frontier is moving over time? Now, we turn to how growth is spreading across countries: how are different countries moving relative to the frontier?

4.1 The Long Run

One of the key facts about the spread of growth over the very long run is that it occurred at different points in time, resulting in what is commonly referred to as

[s] This fact has been noted before, for example by Simon (1981).

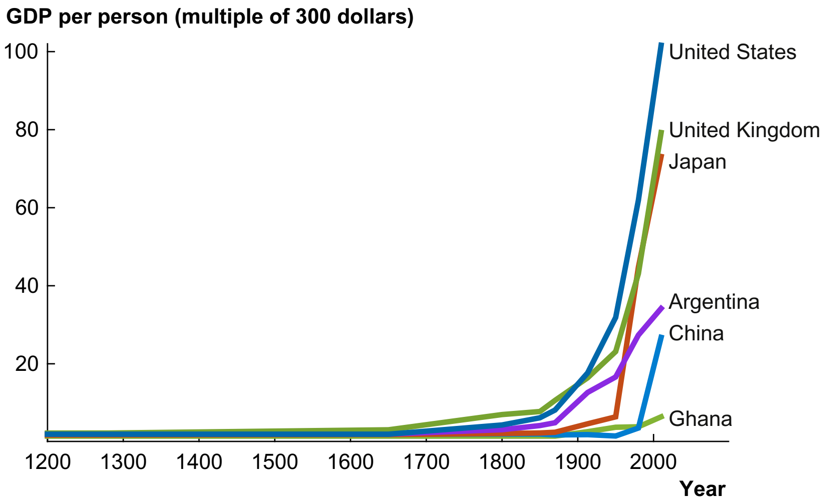


Fig. 21 The great divergence. *Note*: The graph shows GDP per person for various countries. The units are in multiples of 300 dollars and therefore correspond roughly to the ratio between a country's per capita income and the income in the poorest country in the world. Source: *Bolt, J., van Zanden, J.L. 2014. The Maddison Project: collaborative research on historical national accounts. Econ. Hist. Rev. 67 (3), 627–651.*

"The Great Divergence."[t] Fig. 21 illustrates this point. GDP per person differs modestly prior to the year 1600 according to The Maddison Project data. For example, GDP per person in the year 1300 ranges from a high of $1620 in the Netherlands (in 1990 dollars) to a low of $610 in Egypt. But Egypt was surely not the poorest country in the world at the time. Following an insight by Pritchett (1997), notice that the poorest countries in the world in 1950 had an income around $300, and this level—less than one dollar per day—seems very close to the minimum average income likely to prevail in any economy at any point in time. Therefore in 1300, the ratio of the richest country to the poorest was on the order of $\$1620/\$300 \approx 5$. Even smaller ratios are observed in Maddison's data prior to the year 1300.

Fig. 21 shows how this ratio evolved over time for a small sample of countries, and one sees the "Great Divergence" in incomes that occurs after the year 1600. The ratio of richest to poorest rises to more than 10 by 1870 (for the United Kingdom) and then to more than 100 by 2010 (for the United States). Across the range of countries, rapid growth takes hold at different points in time. Argentina is relatively rich by 1870 and growth takes off in Japan after World War II. In 1950, China was substantially poorer than Ghana—by more than a factor of two according to Maddison. Rapid growth since

[t] See Maddison (1995), Pritchett (1997), Lucas (2000), and Pomeranz (2009).

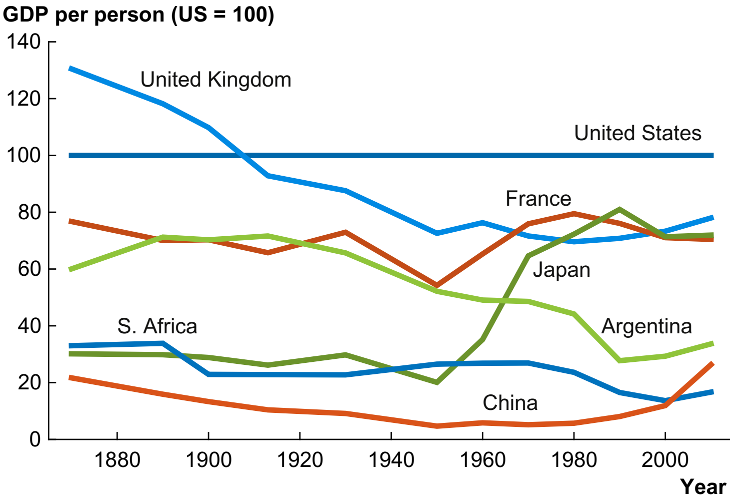


Fig. 22 The spread of economic growth since 1870. Source: *Bolt, J., van Zanden, J.L. 2014. The Maddison Project: collaborative research on historical national accounts. Econ. Hist. Rev. 67 (3), 627–651.*

1978 raises China's living standards to more than a factor of 25 over the benchmark level of $300 per year.

Fig. 22 shows the spread of growth since 1870 in an alternative way, by plotting incomes relative to the US level. A key fact that stands out when the data are viewed this way is the heterogeneity of experiences. Some countries like the United Kingdom, Argentina, and South Africa experience significant declines in their incomes relative to the United States, revealing the fact that their growth rates over long periods of time fell short of the 2% growth rate of the frontier. Other countries like Japan and China see large increases in relative incomes.

4.2 The Spread of Growth in Recent Decades

Fig. 23 focuses in on the last 30 years using the Penn World Tables 8.0 data, again showing GDP per person relative to the US Several facts then stand out. First, incomes in the countries of Western Europe have been roughly stable, around 75% of the US level. It is perhaps surprising that countries like France, Germany, and the United Kingdom are this far behind the United States. Prescott (2004) observes that a large part of the difference is in hours worked: GDP per hour is much more similar in these countries, and it is the fact that work hours per adult are substantially lower in Western Europe that explains their lower GDP per person. Jones and Klenow (2015) note that in addition to the higher leisure, Western Europeans tend to have higher life expectancy and lower consumption inequality. Taking all of these factors into account in constructing a consumption-equivalent welfare measure, the Western European countries look

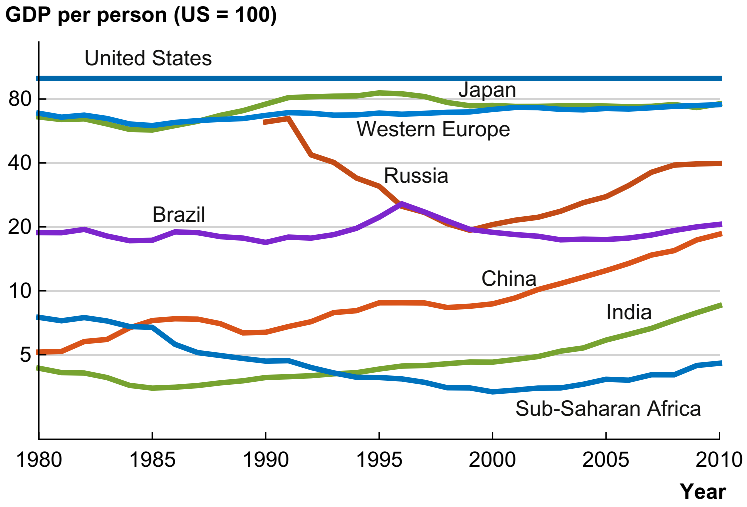


Fig. 23 The spread of economic growth since 1980. Source: *The Penn World Tables 8.0.*

much closer to US levels than the simple GDP per person numbers imply; this point is discussed further below in Section 3.

Fig. 23 also illustrates the "lost decades" that Japan has experienced. After rapid growth in the 1980s (and before), Japan peaked at an income relative to the United States of 85% in 1995. Since 1995, though, Japan has fallen back to around 75% of the US level. The rapid growth of China since 1980 and India since around 1990 are also evident in this figure. The contrast with sub-Saharan Africa is particularly striking, as that region as a whole falls from 7.5% of US income in 1980 to just 3.3% by 2000. Since 2000, many of the countries and regions shown in Fig. 23 exhibit catch-up to the United States.

Fig. 24 shows GDP per person relative to the United States in 1960 and 2011 for 100 countries. Countries scatter widely around the 45-degree line, and the first impression is that there is no systematic pattern to this scattering. Some countries are moving up relative to the United States and some countries are falling further behind, and the movements can be large, as represented by the deviations from the 45-degree line.

Looking more closely at the graph, there is some suggestion that there are more middle-income countries above the 45-degree line than below. At least between 1960 and 2011, countries in the middle of the distribution seemed more likely to move closer to the United States than to fall further behind. In contrast, for low income countries, the opposite pattern appears in the data: poor countries are on average more systematically below the 45-degree line rather than above.

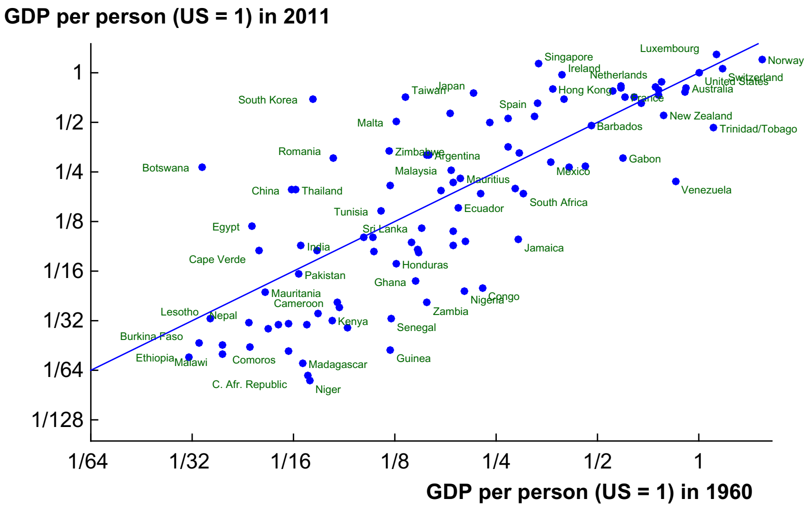


Fig. 24 GDP per person, 1960 and 2011. Source: *The Penn World Tables 8.0.*

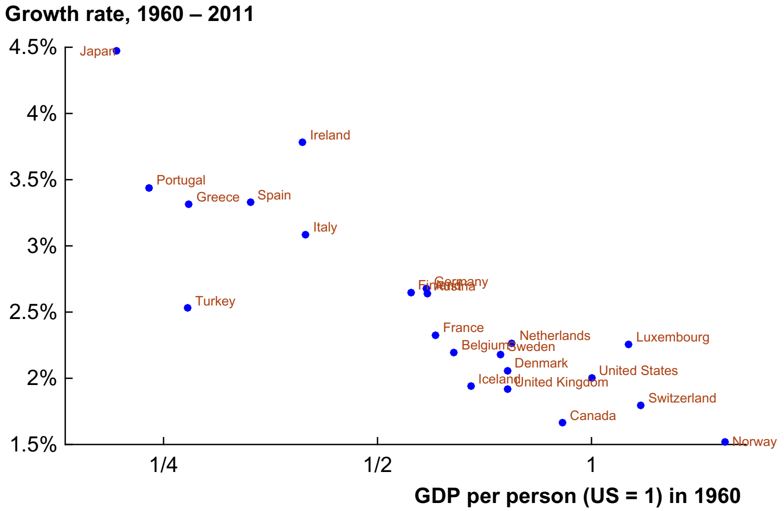


Fig. 25 Convergence in the OECD. Source: *The Penn World Tables 8.0. Countries in the OECD as of 1970 are shown.*

Fig. 25 shows one of the more famous graphs from the empirical growth literature, illustrating the "catch-up" behavior of OECD countries since 1960. Among OECD countries, those that were relatively poor in 1960—like Japan, Portugal, and Greece—grew rapidly, while those that were relatively rich in 1960—like Switzerland, Norway, and

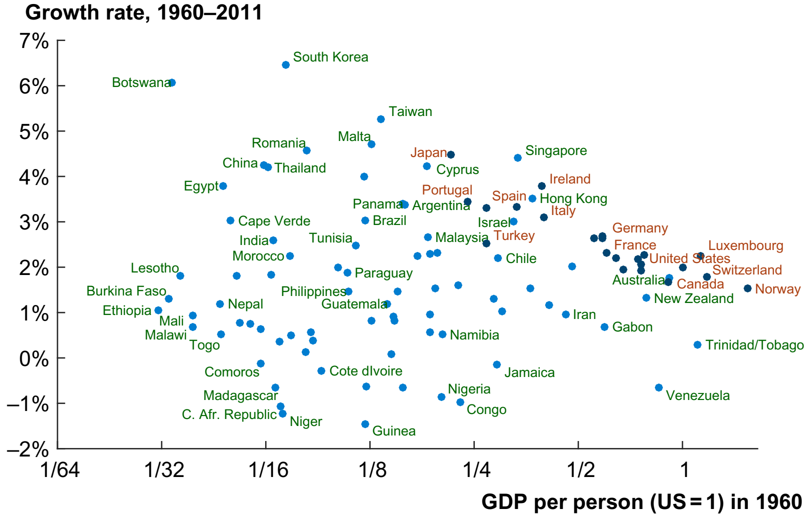


Fig. 26 The lack of convergence worldwide. Source: *The Penn World Tables 8.0.*

the United States—grew more slowly. The pattern is quite strong in the data; a simple regression line leads to an R-squared of 75%.[u]

Fig. 26 shows that a simplistic view of convergence does not hold for the world as a whole. There is no tendence for poor countries around the world to grow either faster or slower than rich countries. For every Botswana and South Korea, there is a Madagascar and Niger. Remarkably, 14 out of the 100 shown in the figure exhibited a negative growth rate of GDP per person between 1960 and 2011.

There is some question as to whether or not these persistent negative growth rates are entirely accurate. Young (2012) notes that the data on which these growth rates are based is often of very poor quality. For example, the United Nations National Accounts database publishes current and constant-price GDP numbers for 47 sub-Saharan African countries between 1991 and 2004, but as of mid-2006, the UN Statistical Office had actually received data for only one half of the observations, and had received no constant-price data at all for this period for 15 of these countries. Young uses measures of consumer durables (eg, radios, television sets, and bicycles) and other information from the Demographic and Health Surveys for developing countries to provide an alternative estimate of growth rates. He finds that living standards in sub-Saharan African countries were growing at around 3.5% per year during the last two decades, comparable to growth rates in other developing countries.

Barro (1991), Barro and Sala-i-Martin (1992), and Mankiw et al. (1992) provide a key insight into why the convergence pattern appears in Fig. 25 but not in Fig. 26. In

[u] See also Baumol (1986) and DeLong (1988).

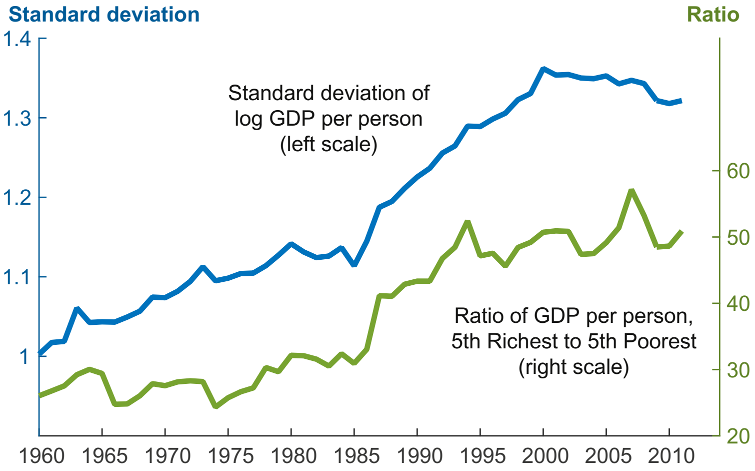


Fig. 27 Divergence since 1960. Source: *The Penn World Tables 8.0, calculated across a stable sample of 100 countries.*

particular, they show that the basic predictions of neoclassical growth theory hold for the world as a whole. Countries around the world are converging—but to their own steady-states, rather than to the frontier. If one conditions on determinants of a country's steady state (such as the investment rates in physical and human capital), then one sees that countries below their steady states grow rapidly and those above their steady states grow slowly (or even decline). The rate at which countries converge to their own steady state—often called the "speed of convergence"—seems to be around 2% per year, a fact sometimes known as "Barro's iron law of convergence." The interpretation of the OECD countries in Fig. 25, then, is that these countries have relatively similar steady state positions, so that even if we do not condition on these determinants formally, the convergence phenomenon appears. Confirming this logic, the implied speed of convergence for the OECD countries estimated from the slope of the best-fit line for Fig. 25 is 1.8% per year.[v]

These general patterns are examined in more detail in the following graphs and tables. Fig. 27 shows a time series of the cross-sectional standard deviation of log GDP per person for this stable 100-country sample. As an alternative measure of dispersion, it also shows the ratio of GDP per person between the 5th richest and 5th poorest countries in the sample. Both measures reveal the same thing: between 1960 and the late-1990s, there was a widening of the world income distribution, at least when each country is a unit of observation. In the last decade or so, this pattern seems to have stabilized. In fact, some of this pattern was already evident back in Fig. 24. The poorest countries in 1960 such as Ethiopia were only about 32 times poorer than the

[v] See Barro (2012) for a recent discussion of convergence.

Table 4 The very long-run distribution

	Distribution			
"Bin"	**1980**	**2010**	**Long run**	**Years to "shuffle"**
Less than 5%	18	21	15	1190
Between 5% and 10%	19	16	8	1100
Between 10% and 20%	22	16	11	920
Between 20% and 40%	13	18	14	270
Between 40% and 80%	19	18	32	950
More than 80%	9	12	20	1000

Entries under "Distribution" reflect the percentage of countries with relative (to the United States) GDP per person in each bin. "Years to shuffle" indicates the number of years after which the best guess as to a country's location is given by the long-run distribution (ie, within a percentage point, bin by bin), provided that the country begins in a particular bin.
Source: Computed following Jones, C.I. 1997. On the evolution of the world income distribution. J. Econ. Perspect. 11, 19–36 using the Penn World Tables 8.0 for 134 countries.

United States. By 2011, there are many countries with relative incomes below this level, and both Niger and the Central African Republic were more than 64 times poorer than the United States.

Table 4 examines the dynamics of the distribution of incomes across countries in a more systematic fashion, following Quah (1993). First, we sort the 134 countries for which we have data in both 1980 and 2010 into bins based on their income relative to the world frontier, represented by the United States. Then, using decadal growth rates between 1980 and 2010, we calculate the sample probabilities that countries move from one bin to another. Finally, we compute the stationary distribution of countries across the bins that is implied by assuming these sample probabilities are constant forever.

To begin, consider the 1980 and 2010 distributions shown in Table 4. The fraction of countries in the highest two bins increases slightly between 1980 and 2010. There is also a decrease in the fraction of countries between 5% and 20% of the US level.

Iterating over the dynamics implied by the sample transition probabilities leads to the stationary distribution shown in the third main column of the table.[w] Many countries are projected to move out of the lower and middle portions of the distribution and into the top.

Overall, the picture that emerges from this kind of analysis is that there is a basic dynamic in the data for the last 30 years that says that once countries get on the "growth escalator," good things tend to happen and they grow rapidly to move closer to the frontier. Where they end up depends, as we will discuss, on the extent to which their institutions improve. And this process is itself captured in the Markov transition dynamics estimated in Table 4. But whereas less than 30% of the countries were in the top two bins in 1980—with incomes

[w] Mathematically, the computation is easily illustrated. We estimate the Markov transition probabilities of countries across the bins. Multiplying this matrix by the initial distribution yields an estimate of the distribution of income for the next decade. Doing this many times yields an estimate of the long-run distribution.

greater than 40% of the US level—the Markov approach suggests that more than half of all countries will achieve this level of relative income in the long run.

One thing that may be misleading about this kind of exercise is that it implies that the stationary distribution is ergodic. That is, countries move around this distribution over time, so that, given enough time, even the United Kingdom can end up in the poorest bin. (The last column of the table suggests that these dynamics are very slow.) Perhaps this is correct—think about Argentina during the past 150 years. Alternatively, Lucas (2000), in his "Macroeconomics for the 21st Century," suggests that from the standpoint of the year 2100, the most striking fact of macroeconomics may end up being how many countries have moved close to the frontier. In other words, the Great Divergence of the last 200 years may turn into a Great Convergence over the next century. Perhaps the diffusion of good rules and good institutions leads to a more or less permanent improvement in the distribution, which is only partially captured by the kind of calculation done here.[x]

4.3 The Distribution of Income by Person, Not by Country

Up until now, we've taken the country as the unit of observation. This is appropriate for some purposes, but there is also a good case to be made for taking the person as the unit of observation: why should the 500,000 people in Luxembourg get the same weight as the 1.4 billion people in China?

Fig. 28 approaches the data from the standpoint of the individual. We assume each person in a country gets that country's GDP per person and then compute the world income distribution by person. More detailed calculations along these lines incorporate the distribution of income *within* each country as well and provide further support for the basic point in Fig. 28; see Bourguignon and Morrisson (2002) and especially Sala-i-Martin (2006).

In 1960, 51% of the world's population lived on less than 3 dollars per day (measured in 2005 US dollars). By 2011, less than 5% of the world's population lived below this level. The big difference, of course, is China and India, which between them contain more than one third of the world's population. In 1960, China and India were very poor, with incomes below $2.75 per day, while by 2011 average incomes were $10 per day in India and $22 per day in China. From the standpoint of the individual, the most outstanding fact of economic growth over the last 50 years is how many people have been elevated out of poverty.

4.4 Beyond GDP

It has long been recognized that GDP is an imperfect measure of living standards. Pollution, leisure, life expectancy, inequality, crime, and even freedom are some of

[x] Buera et al. (2011b) study the diffusion of market-oriented institutions in a setting where countries learn from the growth experiences of their neighbors. Kane (2015) explores a Markov approach like that discussed here and suggests that the escalator effects have been getting more prevalent over time.

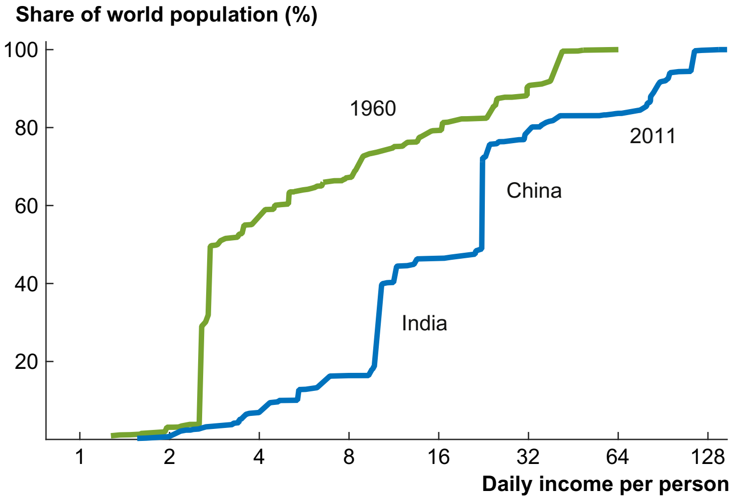


Fig. 28 The distribution of world income by population. Source: *The Penn World Tables 8.0, calculated across a stable sample of 100 countries.*

the factors that are incorporated imperfectly, if at all, in GDP. Various attempts have been made over the years to repair some of the omissions, or at least to judge how important they might be. Examples include Nordhaus and Tobin (1972), Becker et al. (2005), Fleurbaey and Gaulier (2009), and Stiglitz et al. (2009).

Jones and Klenow (2015) extend this literature by using a standard utility function and a "behind the veil of ignorance" approach to construct a consumption-equivalent welfare measure that values consumption, leisure, mortality, and inequality for a range of countries. Table 5 shows their baseline findings for a high-quality sample of countries for which household survey data can be used to compute welfare. The key finding comes in two parts. First, Western European countries like the United Kingdom and France have much higher living standards than their GDPs indicate. For example, compared to the United States, France has higher life expectancy, more leisure per person, and lower inequality of both consumption and leisure, and these differences make a substantial difference: whereas GDP per person in France is only about 2/3 of that in the United States, consumption-equivalent welfare is around 92% of the US level. Second, while many rich countries are richer than we might have thought, the opposite is true for poor countries. Life expectancy and leisure tend to be lower and inequality tends to be higher, all of which reduce welfare relative to GDP. As just one example, South Africa's GDP per person is about 16% of the US level, but consumption-equivalent welfare is only 7.4% of that in the United States.

In terms of growth rates, declining mortality has the largest impact: in most countries of the world—the notable exception being in sub-Saharan Africa—declining mortality has raised consumption-equivalent welfare growth substantially. In the United States and

Table 5 Beyond GDP: Welfare across countries

	Consumption-equivalent welfare	Income	Log ratio	Life exp.	Decomposition C/Y	Leisure	Cons. ineq.	Leis. ineq.
United States	100.0	100.0	0.000	0.000	0.000	0.000	0.000	0.000
United Kingdom	96.6	75.2	0.250	0.086	−0.143	0.073	0.136	0.097
France	91.8	67.2	0.312	0.155	−0.152	0.083	0.102	0.124
Italy	80.2	66.1	0.193	0.182	−0.228	0.078	0.086	0.075
Spain	73.3	61.1	0.182	0.133	−0.111	0.070	0.017	0.073
Mexico	21.9	28.6	−0.268	−0.156	−0.021	−0.010	−0.076	−0.005
Russia	20.7	37.0	−0.583	−0.501	−0.248	0.035	0.098	0.032
Brazil	11.1	17.2	−0.436	−0.242	0.004	0.005	−0.209	0.006
S. Africa	7.4	16.0	−0.771	−0.555	0.018	0.054	−0.283	−0.006
China	6.3	10.1	−0.468	−0.174	−0.311	−0.016	0.048	−0.014
Indonesia	5.0	7.8	−0.445	−0.340	−0.178	−0.001	0.114	−0.041
India	3.2	5.6	−0.559	−0.440	−0.158	−0.019	0.085	−0.028
Malawi	0.9	1.3	−0.310	−0.389	0.012	−0.020	0.058	0.028

Notes: The consumption-equivalent welfare numbers in the first column use a conventional utility function to "add up" the contributions from consumption, leisure, mortality, and inequality and express them in a consumption-equivalent manner. The income column reports GDP per person. The "decomposition" columns report an additive decomposition of the log difference between welfare and income.

Source: These numbers are taken from table 2 of Jones, C.I., Klenow, P.J. 2015. Beyond GDP: Welfare across countries and time. Stanford University, unpublished manuscript, and are based on data from household surveys in each country, from the World Bank (for mortality), and from the Penn World Tables 8.0 for a year close to 2005.

Western Europe, for example, growth rates since 1980 are arguably understated by around a full percentage point because of this factor.

4.5 Development Accounting

Development accounting applies the logic of Solow's growth accounting to explain differences in the levels of GDP per worker across countries. Countries can be rich because they have large quantities of inputs or because they use these inputs efficiently. Quantitatively, how important are each of these components?

An early version of development accounting is Denison (1967), who compared 8 European economies to the United States in 1960. Christensen et al. (1981) work with a similar set of countries and extend the analysis to include human capital. King and Levine (1994) focus on the role of physical capital vs TFP in a broad set of countries, and provide the first use of the phrase "development accounting" that I have found.[y] Klenow and Rodriguez-Clare (1997) and Hall and Jones (1999) incorporate human capital differences and consider a broad range of countries. Caselli (2005) provides a detailed overview and analysis of this literature.[z]

The basics of development accounting follow closely upon the analysis of growth accounting that we conducted back in Section 2 To see this link, recall the production function we considered there:

$$Y_t = \underbrace{A_t M_t}_{\text{TFP}} K_t^{\alpha} H_t^{1-\alpha}. \tag{5}$$

Some versions of development accounting work directly with this production function. The advantage is that it is the most straightforward approach. The disadvantage is familiar from growth accounting and the standard neoclassical growth model: differences in TFP induce capital accumulation that leads to differences in K across countries. Hence some of what is attributed to K in this approach might more naturally be attributed to TFP.

An alternative approach—pursued by Klenow and Rodriguez-Clare (1997) and Hall and Jones (1999)—is to rewrite the production function in a way that assigns any induced capital accumulation to TFP. This is accomplished by dividing both sides of the production function by Y_t^{α} and solving for Y_t to get

[y] Bob Hall and I (Hall and Jones, 1996) proposed the phrase "levels accounting" which doesn't have nearly the same ring!

[z] The papers cited to this point assume a known production function—typically Cobb-Douglas with an exponent on capital around 1/3. A related set of papers including Mankiw et al. (1992), Islam (1995), and Caselli et al. (1996) conduct a similar exercise in a regression framework. The limitation of the regression framework in its simplest form is that it imposes an orthogonality between inputs and total factor productivity which seems inappropriate. Estimating production functions is notoriously difficult.

$$\frac{Y_t}{L_t} = \left(\frac{K_t}{Y_t}\right)^{\frac{\alpha}{1-\alpha}} \frac{H_t}{L_t} \cdot Z_t \tag{6}$$

where $Z_t \equiv (A_t M_t)^{\frac{1}{1-\alpha}}$ is total factor productivity measured in labor-augmenting units. To understand why this equation assigns the induced capital accumulation to TFP, notice that in the steady state of a neoclassical growth model, the capital–output ratio K/Y is proportional to the investment rate and independent of TFP. Hence the contributions from productivity and capital deepening are separated in this version, in a way that they were not in Eq. (5). This was the equation on which we based our growth accounting, and we will use the same equation for development accounting.

The Penn World Tables, starting with Version 8.0, contains all the information needed to conduct the simplest form of development accounting as in Eq. (6). That data set contains measures of the economy's stock of physical capital and measures of human capital that are based on educational attainment data from Barro and Lee (2013) and Mincerian returns to education of 13.4% for the first 4 years, 10.1% for the second 4 years, and 6.8% for all additional years, as in Hall and Jones (1999). We conduct our growth accounting exercise using this data and Eq. (6), assuming $\alpha = 1/3$.[aa]

Table 6 shows the basic development accounting exercise using the Penn World Tables data for a sample of countries.[ab] To see how the accounting works, consider the row for Mexico. According to the Penn World Tables, Mexico has a GDP per worker that is about 1/3 that in the United States in 2010. This 1/3 number is the product of the next three terms in the row, following the formula in Eq. (6).

Remarkably little of the difference is due to physical capital: the capital–output ratio in Mexico is about 87% of that in the United States. Because of diminishing returns, though, it is the square root ($\alpha/(1-\alpha) = 1/2$ when $\alpha = 1/3$) of this difference that matters for income, and $\sqrt{0.87} \approx 0.93$, so differences in physical capital only lead to about a 7% difference in GDP per worker between the United States and Mexico.

In the next column, we see a larger contribution from human capital. In 2010, people aged 15 and over in Mexico had on average around 8.8 years of education according to Barro and Lee (2013). In contrast, educational attainment in the United States was 13.2. The difference is 4.6 years of schooling. With a return to each year of education of around 7%, this implies about a 32% difference due to education. The entry from human capital for Mexico is 0.76, and $1/0.76 \approx 1.32$, consistent with this reasoning.

The implied difference in TFP between the United States and Mexico is then $0.338/(0.931 \times 0.760) \approx 0.477$. Put another way, GDP per worker was 3 times higher

[aa] The Penn World Tables now contains its own measure of TFP as well. This measure is based on a Tornqvist index of inputs that incorporates variations in factor shares. The data appendix of Feenstra et al. (2015) contains a helpful and extensive discussion of the data and methods.

[ab] Data on all countries can be obtained in the data files available on the author's web page; see the file "DevelopmentAccounting.log."

Table 6 Basic development accounting, 2010

	GDP per worker, y	Capital/GDP $(K/Y)^{\alpha/(1-\alpha)}$	Human capital, h	TFP	Share due to TFP
United States	1.000	1.000	1.000	1.000	–
Hong Kong	0.854	1.086	0.833	0.944	48.9%
Singapore	0.845	1.105	0.764	1.001	45.8%
France	0.790	1.184	0.840	0.795	55.6%
Germany	0.740	1.078	0.918	0.748	57.0%
United Kingdom	0.733	1.015	0.780	0.925	46.1%
Japan	0.683	1.218	0.903	0.620	63.9%
South Korea	0.598	1.146	0.925	0.564	65.3%
Argentina	0.376	1.109	0.779	0.435	66.5%
Mexico	0.338	0.931	0.760	0.477	59.7%
Botswana	0.236	1.034	0.786	0.291	73.7%
South Africa	0.225	0.877	0.731	0.351	64.6%
Brazil	0.183	1.084	0.676	0.250	74.5%
Thailand	0.154	1.125	0.667	0.206	78.5%
China	0.136	1.137	0.713	0.168	82.9%
Indonesia	0.096	1.014	0.575	0.165	77.9%
India	0.096	0.827	0.533	0.217	67.0%
Kenya	0.037	0.819	0.618	0.073	87.3%
Malawi	0.021	1.107	0.507	0.038	93.6%
Average	0.212	0.979	0.705	0.307	63.8%
1/Average	4.720	1.021	1.418	3.260	69.2%

The product of the three input columns equals GDP per worker. The penultimate row, "Average," shows the geometric average of each column across 128 countries. The "Share due to TFP" column is computed as described in the text. The 69.2% share in the last row is computed looking across the columns, ie, as approximately 3.5/(3.5 + 1.5).
Source: Computed using the Penn World Tables 8.0 for the year 2010 assuming a common value of $\alpha = 1/3$.

in the United States than in Mexico. A factor of $1.07 \times 1.32 \approx 1.4$ of this difference is due to inputs, meaning a factor of 2.1 was due to TFP, since $1.4 \times 2.1 \approx 3$. From these numbers, one can also see easily how the "Share due to TFP" column is calculated. Notice that both 1.4 and 2.1 are simple multiples of 7: for each 2 parts due to inputs, 3 parts are due to TFP, hence the share due to TFP is around 60% (that is, 3/(2+3)).

More generally, several key findings stand out from Table 6. First, the capital-output ratio is remarkably stable across countries. Its average value is very close to one, and even the poorest country in the table, Malawi, is reported by the Penn World Tables to have a capital-output ratio very close to the US value. So differences in physical capital contribute almost nothing to differences in GDP per worker across countries. Caselli and Feyrer (2007) document a closely-related fact in great detail: the marginal product of capital (which here is proportional to the inverse of the capital-output ratio) is very similar in rich and poor countries.[ac]

[ac] The general lack of correlation between the capital-output ratio and GDP per person is discussed by Feenstra et al. (2015).

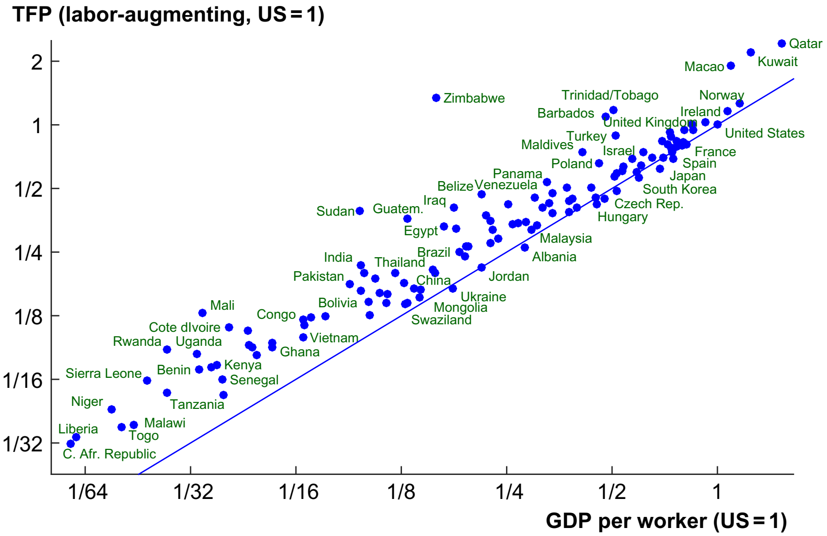


Fig. 29 Total factor productivity, 2010. Source: *Computed using the Penn World Tables 8.0 assuming a common value of* $\alpha = 1/3$.

Second, the contribution from educational attainment is larger, but still modest. For example, countries like India and Malawi only see their incomes reduced by a factor of 2 due to educational attainment. Loosely speaking, the poorest countries of the world have 4 or 5 years of education, while the richest have 13. Eight years of education with a Mincerian return of around 10% leads to a 0.8 difference in logs, and $exp(0.8) \approx 2$.

Finally, the implication of these first two points is that differences in TFP are the largest contributor to income differences in an accounting sense. Fig. 29 shows the levels of TFP plotted against GDP per worker for 128 countries in 2010. The two series are highly correlated at 0.96. And the differences in TFP are very large: the Central African Republic is about 64 times poorer than the United States and its TFP is about 32 times lower than the US level.

The large contribution from TFP is verified by the last column of Table 6, where the share explained by TFP ranges from just under 50% for Singapore and Hong Kong to more than 90% for Malawi. To understand the "Share due to TFP" column, consider the last row of Table 6. According to that row, the average country in the 128-country sample is just over 5 times poorer than the United States. Essentially none of this difference (a factor of 1.021) is due to differences in K/Y, while a factor of 1.42 is due to differences in educational attainment. Taken together, this means a factor of $1.021 \times 1.42 \approx 1.5$ is due to inputs, leaving a factor of about 3.5 attributed to TFP. We then compute the "Share due to TFP" as $3.5/(1.5 + 3.5) \approx 70\%$, as shown in the last entry in Table 6.[ad] The rest of the shares are computed in an analogous way. For example, for Malawi, about a factor

[ad] Or more exactly as $3.26/(3.26 + 1.021 * 1.418) \approx 69.2\%$.

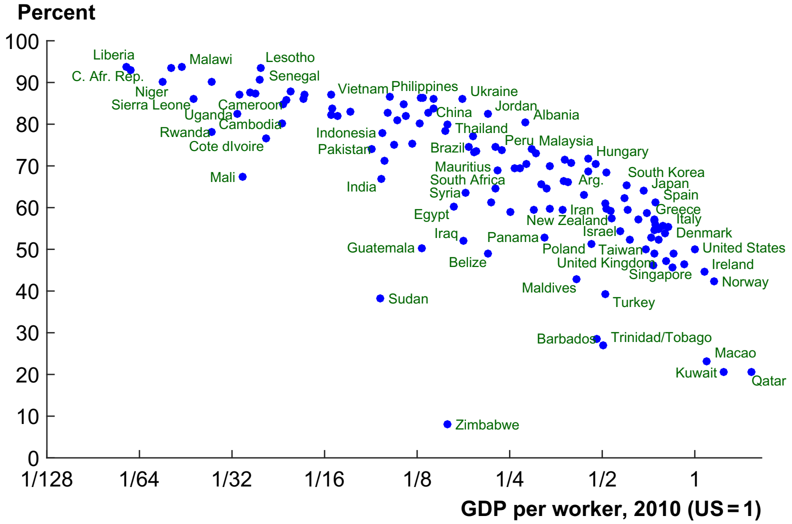


Fig. 30 The share of TFP in development accounting, 2010. Source: *Computed as described in the text and in Table 6 using the Penn World Tables 8.0 assuming a common value of* $\alpha = 1/3$.

of 2 is due to inputs and a factor of 26 is due to TFP, meaning that 26/28 ≈ 93% is due to TFP.

More generally, the share across all 128 countries is shown in Fig. 30. There, a systematic pattern is obvious. In the poorest countries of the world, well over 80% of the difference in GDP per worker relative to the United States is due to TFP differences. Moving across the graph to richer countries, one sees that less and less is due to TFP, and for the richest countries as a whole, TFP contributes around 50% of the differences.

4.6 Understanding TFP Differences

The basic finding that TFP differences account for the bulk of income differences across countries has led to a large body of research designed to explain what these differences are. This is exemplified by the title of a famous paper by Prescott (1998): "Needed: A Theory of TFP."

In the last 15 years, this challenge has been approached in two ways. First, several papers have improved our measures of inputs in various ways, chipping away at the contribution of the "measure of our ignorance." Second, the literature on misallocation has emerged to provide the kind of theory that Prescott was seeking. The remainder of this section will review the efforts to improve input measurement, while the next several sections will consider misallocation and its implications.

Caselli (2005) provides a detailed survey and analysis of the state of development accounting as of 2005. The interested reader should certainly look there to get up to speed. Caselli reviews progress on many dimensions: measuring the quality of education using test scores (Hanushek and Kimko, 2000); considering differences in the experience of the labor force across countries (Klenow and Rodriguez-Clare, 1997); sectoral differences in productivity, especially agriculture (Restuccia et al., 2008); differences in labor productivity due to health (Weil, 2007); differences in the quality of capital (Caselli and Wilson, 2004); and the potential role of nonneutral productivity (Caselli and Coleman, 2006).

Much additional progress has been made in the decade since Caselli's review was published, especially with respect to misallocation, as discussed in the next section. But there has also been much valuable work on measuring the inputs into development accounting. Lagakos et al. (2012) use household survey data from 35 countries to show that the returns to experience vary substantially across countries, with poorer countries typically having much flatter age-earnings profiles. Incorporating differential returns to experience in development accounting boosts the importance of the human capital term by about 50%. Hendricks and Schoellman (2014) use data on immigrants from 50 source countries into 11 different host countries to improve our measurement of labor quality differences, providing another boost to the human capital term—of about 30%. Hanushek and Woessmann (2008) survey the broad range of evidence highlighting the importance of educational quality and cognitive skills more generally.

Two recent papers study the role of human capital once we depart from the assumption that workers with different human capital (such as education) are perfect substitutes. Jones (2014) proposes a generalized aggregator over workers with heterogeneous education levels and argues that the traditional perfect-substitutes case delivers a lower bound for the role of human capital. If workers with different human capital are less than perfect substitutes, the share of income differences explained by human capital may rise dramatically. Caselli and Ciccone (2013), however, take a similar approach but find the opposite result: the perfect substitutes case gives an upper bound on the importance of human capital differences, which therefore must be small.

How can these two papers be reconciled? To see an answer, consider a factor-augmenting productivity term that multiplies the amount of labor of each type. Caselli and Ciccone effectively keep this parameter unchanged when they do their development accounting and find large neutral TFP differences. Jones implicitly assumes these productivities change when the quantities of the different types of human capital change across countries, so as to keep the relative wages across types (eg, the skilled wage premium) constant. In other words, (Ben) Jones requires large productivity residuals in his development accounting as well, it's just that he labels these residuals as being inside the "human capital" aggregator rather than being neutral TFP differences.[ae]

4.7 Misallocation: A Theory of TFP

One of the great insights of the growth literature in the last 15 years is that misallocation at the micro level can show up as a reduction in total factor productivity at a more aggregated level. This insight appears in various places, including Banerjee and Duflo (2005), Chari et al. (2007), Restuccia and Rogerson (2008) and Hsieh and Klenow (2009).

As we discussed briefly in the context of misallocation and frontier growth (in Section 2.6), the essence of this insight is quite straightforward: when resources are allocated optimally, the economy will operate on its production possibilities frontier. When resources are misallocated, the economy will operate inside this frontier. But that is just another way of saying that TFP will be lower: a given quantity of inputs will produce less output.

A simple example illustrates this point. Suppose output is produced using two tasks according to $Y = X_1^{\alpha} X_2^{1-\alpha}$. This could describe a firm, and the tasks could be management and the production line, or this could be the economy as a whole and the tasks could be manufacturing and services. Suppose that each task is accomplished very simply: one unit of labor can produce one unit of either task, and the economy is endowed with L units of labor. Finally, suppose that the allocation of labor is such that a fraction s works in the first task, and the fraction $1 - s$ works in the second task. We leave the source of this allocation unspecified: labor could be optimally allocated, or it could be misallocated because of taxes, poor management, information problems, unions, or many other reasons. But given this allocation, there is a reduced-form production function given by

$$Y = M(s)L \quad \text{where} \quad M(s) \equiv s^{\alpha}(1-s)^{1-\alpha} \tag{7}$$

[ae] This finding is related to an observation made by Caselli and Coleman (2006). They noted that the ratio of "skilled" to "unskilled" workers varies enormously across countries. For example, if we let high school completion be the dividing line, the ratio of skilled to unskilled workers is just 0.025 in Kenya vs 1.8 in the United States—a difference of a factor of 70. If college completion is the dividing line, the factor proportions differ by even more. When workers with different human capital levels are no longer perfect substitutes, this ratio becomes relevant. The difficulty is that it can then imply implausibly large differences in the return to schooling across countries if one is not careful. Caselli and Coleman introduce additional TFP terms so they can match the returns to education, but then the large differences in factor proportions shows up as enormous differences in these nonneutral TFP terms. A similar issue seems to arise in the approach taken by (Ben) Jones; in this sense, there is implicitly an omitted nonneutral (ie, skill biased) TFP term that differs across countries and that is not being taken into account.

In other words, total factor productivity in this economy is $M(s)$, which depends on the allocation of labor in the economy.[af] Moreover, it is easy to see that the output-maximizing allocation of labor in this example has $s^* = \alpha$, and any departure of the allocation from s^* will reduce total factor productivity. This is the essence of the literature on misallocation and TFP.

This insight is at the heart of our best candidate explanations for answering the question of why some countries are so much richer than others. Development accounting tells us that poor countries have low levels of inputs, but they are also remarkably inefficient in how they use those inputs. Misallocation provides the theoretical connection between the myriad distortions in poor economies and the TFP differences that we observe in development accounting.

The remainder of this section explores various facts related to misallocation.

4.8 Institutions and the Role of Government

Countries are a natural unit of analysis for growth economists for the simple reason that national borders are the places where different political and economic institutions begin and end. It has long been conjectured that differences in these institutions are fundamental determinants of long-run economic success. But what is the evidence for such a claim? How do we know that the income differences we see across countries are not primarily driven by differences in natural resources or other aspects of geography?

One of the best sources of evidence on this question was provided by Olson (1996). Olson observed that history itself provides us with “natural experiments” that allow us to see the large impact of institutions on economic success. For example, prior to World War II, North and South Korea were not separate countries. As a rough approximation, the north and south of Korea contained people that shared a cultural heritage, a government, institutions, and even a geography. In fact, if anything, North Korea was economically advantaged relative to the South, containing a disproportionate share of electricity production and heavy industry.[ag] After the Korean War ended in 1953, North and South Korea were divided and governed according to very different rules. The resulting economic growth of the next half century was dramatically different, as illustrated in

[af] One could easily assume both capital and labor are used to produce each X so that the result in Eq. (7) would be $Y = M(s)K^{\beta}L^{1-\beta}$, which makes the connection between $M(s)$ and TFP even more apparent.

[ag] “[Under Japanese rule before World War II], the colonial industries were unevenly distributed between South Korea and North Korea. Heavy and chemical industries were concentrated in the North, while many light industries, such as textile, food, printing and wood, were located in the South. In 1940, North Korea’s share of the total production in the metal industry was 96%, and 82% for the chemical industry. Also, 92% of the total electricity production originated from the North in 1945 (Lee, D-G, 2002: 39). Thus, in 1945 when Japan withdrew from Korea and when Korea was divided into two separate political regimes, the South Korean economy in general and industry in particular were severely crippled.” Yang (2004), p. 16.

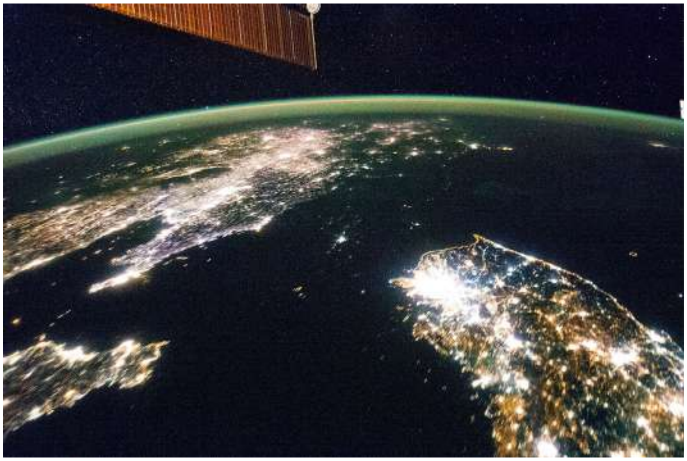


Fig. 31 Korea at night. *Note*: North Korea is the dark area in the center of the figure, between China to the north and South Korea to the south. Pyongyang is the isolated cluster in the center of the picture. Source: *http://commons.wikimedia.org/wiki/File:North_and_South_Korea_at_night.jpg*.

Fig. 31. The picture in this figure was taken by an astronaut on the International Space Station in early 2014 and shows North and South Korea at night. South Korea is brightly lit, while North Korea is almost completely dark, barely indistinguishable from the ocean. Whatever was different between North and South Korea after 1953 apparently had an enormous influence on their long-term economic success.

As a brief aside, it is worth noting that during the last several years, a number of papers have used satellite data on lights at night to study economic growth. Henderson et al. (2012) introduce this data and argue that it provides useful information on growth and standards of living. They also note that it can be used to study growth at the regional level, where income measures are not often available in developing countries. Michalopoulos and Papaioannou (2014) take up this latter point and compare the importance of national policies with subnational/cultural institutions using the light data.

"Natural" experiments similar to the North/South Korea example can be observed in East and West Germany after World War II, Hong Kong and southeastern China, and across the Rio Grande between Mexico and Texas. These examples make clear that something malleable matters for economic success even if they do not specifically identify what that something is.

Another fascinating piece of evidence comes from Acemoglu et al. (2002) and is illustrated in Fig. 32, the so-called reversal of fortune. Restricting our attention to former European colonies, economic success 500 years ago is *negatively* correlated with economic

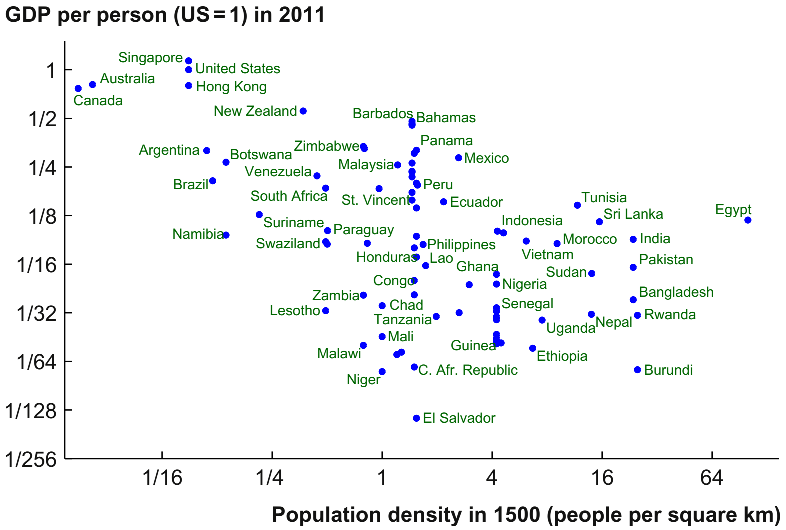


Fig. 32 The reversal of fortune. *Note:* Former European colonies that were proserous (at least in terms of population density) in 1500 are on average poorer today rather than richer. Source: *Population density is from Acemoglu, D., Johnson, S., Robinson, J.A. 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. Q. J. Econ. 117 (4), 1231–1294 and GDP per person is from the Penn World Tables 8.0.*

success today. That is, the places that were most successful 500 years ago, as measured by population density or urbanization, are on average comparatively poor today.

A classic example of this phenomenon, highlighted by Engerman and Sokoloff (1997), is the New World:

> *[Latin America] began with—by European standards of the time—vast supplies of land and natural resources per person and were among the most prosperous and coveted of the colonies in the seventeenth and eighteenth centuries. Indeed, so promising were these other regions that Europeans of the time generally regarded the thirteen British colonies on the North American mainland and Canada as of relatively marginal economic interest—an opinion evidently shared by Native Americans who had concentrated disproportionately in the areas the Spanish eventually developed. Yet, despite their similar, if not less favorable, factor endowments, the United States and Canada ultimately proved to be far more successful than the other colonies in realizing sustained economic growth over time (pp. 260–261).*

These examples suggest that economic success is not permanently given, for example by geographic endowments, but rather can be changed by the rules that are put in place. Engerman and Sokoloff (1997), Acemoglu et al. (2002), and others suggest that the institutions adopted by Europeans in response to these initial conditions influenced subsequent growth. In places that were already economically successful in 1500, Europeans tended to set up “extractive” institutions to transfer the economic gains back to Europe.

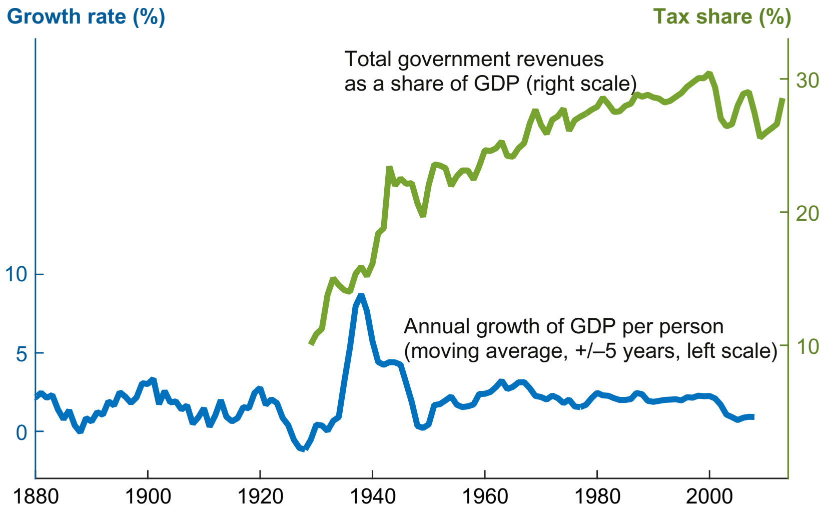


Fig. 33 Taxes and growth in the United States. Source: *This graph updates a similar figure in Stokey, N.L., Rebelo, S. 1995. Growth effects of flat-rate taxes. J. Polit. Econ. 103, 519–550. Total government current receipts are from NIPA table 3.1 via the FRED database and include federal, state, and local revenues. Real GDP per person is constructed as in Fig. 1. The growth rate is smoothed by taking a moving average across the 5 years before and after the relevant date.*

In contrast, Europeans themselves migrated to places that were sparsely populated in 1500, setting up "European" institutions that were conducive to long-run economic success.[ah]

Dell (2010) provides a detailed analysis of the long-reaching nature of one such institution in Peru. The "mita" was a forced labor system conscripting one seventh of the adult male population in a region of Peru to work in local silver and mercury mines between 1573 and 1812. A regression discontinuity analysis reveals that today—200 years later after this system ended—consumption is lower inside the former mita by 25%, educational attainment is lower, and the region is less well connected by roads and infrastructure.

4.9 Taxes and Economic Growth

One of the most obvious and readily quantified measures of government involvement in the economy is taxes. It is easy to write down models in which governments that tax heavily reduce the long-run success of their economies. The facts, however, are not so clear.

Fig. 33 shows the growth rate of real GDP per person in the United States since 1980 as well as the total government tax revenues (including state and local) as a share of GDP, updating a graph first highlighted by Stokey and Rebelo (1995). The remarkable fact that

[ah] For example, see also Hall and Jones (1999), Acemoglu et al. (2001), and Acemoglu and Robinson (2012).

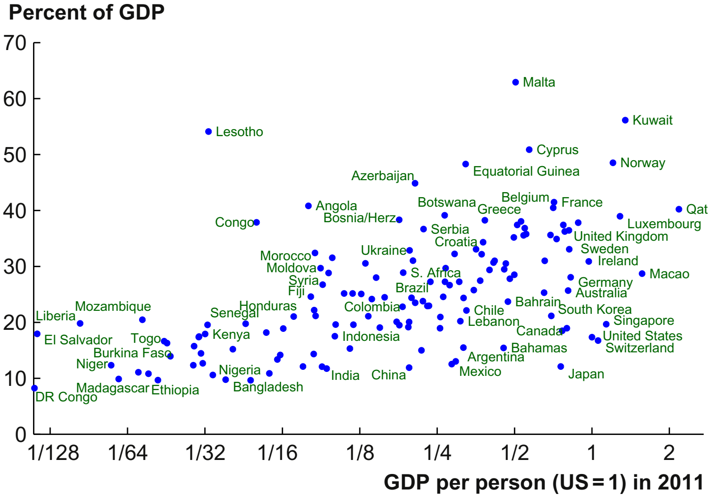


Fig. 34 Tax revenues as a share of GDP. *Note:* Tax revenue is averaged for the available years between 2000 and 2014, is for the central government only, and includes receipts for social insurance programs. Source: *This is an updated graph of a figure from Acemoglu, D. 2005. Politics and economics in weak and strong states. J. Monet. Econ. 52 (7), 1199–1226. http://ideas.repec.org/a/eee/moneco/v52y2005i7p1199-1226.html. The World Bank,* World Development Indicators. *GDP per person is from the Penn World Tables 8.0.*

emerges from this graph is that taxes have increased enormously, from around 10% of GDP in 1929 to more than 30% of GDP at their peak in 2000. But as we already noted earlier, growth rates over the 20th century were remarkably stable—if anything, they were higher after 1950 than before.

Fig. 34 shows a related fact by looking across the countries of the world: tax revenues as a share of GDP are *positively* correlated with economic success, not negatively correlated.

Of course, these are just simple correlations, and the nature of causality is likely to be very complicated. Governments do not simply throw the tax revenue that they collect into the ocean. Instead, this revenue—at least to some extent—is used to fund the good purposes that governments serve: providing a stable rule of law, a judicial system, education, public health, highways, basic research, and so on. Alternatively, perhaps only rich countries can afford large governments. Besley et al. (2013) and Pritchett and Aiyar (2015) consider issues along these lines.

4.10 TFPQ vs TFPR

An important realization related to the measurement of either labor productivity or TFP emerged during the last decade. Specifically, to measure true productivity, one needs detailed information on micro level prices. Foster et al. (2008) introduced the labels "TFPQ" and "TFPR," which will be explained in detail below. This distinction plays a crucial role in Hsieh and Klenow (2009). For more discussion, see the recent survey by Syverson (2011).

To see this point most easily, consider the following setup. The economy consists of a unit measure of heterogeneous varieties that enter the utility function via a CES aggregator:

$$C = \int_0^1 (\alpha_i Y_i)^\rho di \tag{8}$$

where α_i are taste parameters related to each variety and $0 < \rho < 1$ governs the elasticity of substitution between varieties.

Each variety is assumed to be produced by different monopolistically-competitive firms using labor:

$$Y_i = A_i L_i \tag{9}$$

where A_i is the (exogenous) productivity with which each variety is produced. Finally, assume labor is homogenous and can be hired by any firm at a wage rate w.

It is well known that in this kind of setup, monopolistically-competitive firms charge a price p_i for their variety that is a markup over marginal cost:

$$p_i = \frac{1}{\rho} \cdot \frac{w}{A_i}. \tag{10}$$

This implies that sales revenue for each firm is $p_i Y_i = wL_i/\rho$.

Now consider measuring firm-level productivity in this economy. Since we've left capital out of this example, we focus on labor productivity. But exactly the same issues apply to TFP as well.

In general, the data we have available on firms include sales revenues $p_i Y_i$ and employment L_i. This leads immediately to an important point: if one does not have data on the firm-level price p_i, then one cannot recover A_i. For example, deflating revenue by an industry-level price deflator is not the same as deflating by p_i because firms are heterogeneous and have different productivity levels. In the absence of firm-level prices, one typically recovers

$$\text{Revenue Productivity, TFPR}_\text{i}: \ \frac{p_i Y_i}{L_i} = \frac{w}{\rho}.$$

If firms have identical markups and do not face any distortions, revenue productivity should be equated across heterogeneous firms. Workers have to earn the same wage at each firm, and this equates the marginal revenue product of labor across firms, which is all that is being recovered here.

The same argument applies to total factor productivity, and gives rise to the label TFPR, where the "R" denotes "revenue" TFP. The marginal revenue product of capital should also be equated across firms in a simple model like this one, so weighted averages of the average revenue products of capital and labor—which is what TFPR is—should be equated across firms.

By this point, the reader should realize that in a world of heterogeneous goods, it is not even obvious how to compare "true productivity." How do we compare Ford's productivity in making pickup trucks to Tesla's productivity in making electric cars? Or how do we compare Dell's productivity at making PC's with Apple's productivity at making Macs? Even if we had detailed data on the price of Ford trucks and Tesla cars—even if we recovered the A_i's using these prices—they would not be comparable, because the products are different!

Both of these issues are addressed by having knowledge of the utility function, in this case the C aggregator in Eq. (8). In particular, knowledge of the utility function allows one to compute the marginal rate of substitution between different products—it tells us how to compare Fords and Teslas or Apples and Dells.

To see how, notice that the demand curve from utility maximization of (8) is

$$\lambda p_i = \rho \alpha_i^\rho Y_i^{\rho-1}. \tag{11}$$

where λ is the Lagrange multiplier from the budget constraint; we'll choose units so that $\lambda = 1$ in what follows. Sales revenue for variety i is then

$$p_i Y_i = \rho(\alpha_i Y_i)^\rho \tag{12}$$

and we can invert this equation to obtain

$$\alpha_i Y_i = \left(\frac{p_i Y_i}{\rho}\right)^{1/\rho}. \tag{13}$$

But this is what we require: $\alpha_i Y_i$ is the term that enters the utility aggregator C. The α_i tell us how to combine Fords and Teslas:[ai]

$$\text{True Productivity, TFPQ}_\text{i}: \ \frac{\alpha_i Y_i}{L_i} = \frac{(p_i Y_i)^{1/\rho}}{L_i} = \alpha_i A_i.$$

That is, TFPQ (the "Q" denotes "quantity" TFP) indicates how effective a firm is at taking a unit of labor (in this case, with no capital) and using it to produce Fords or Teslas in comparable units.

Notice that TFPQ is measured in this case using only the data we typically have—sales revenue and employment—together with knowledge of the elasticity of demand.

[ai] I'm dropping the $\rho^{1/\rho}$ term in the equation to keep things clear.

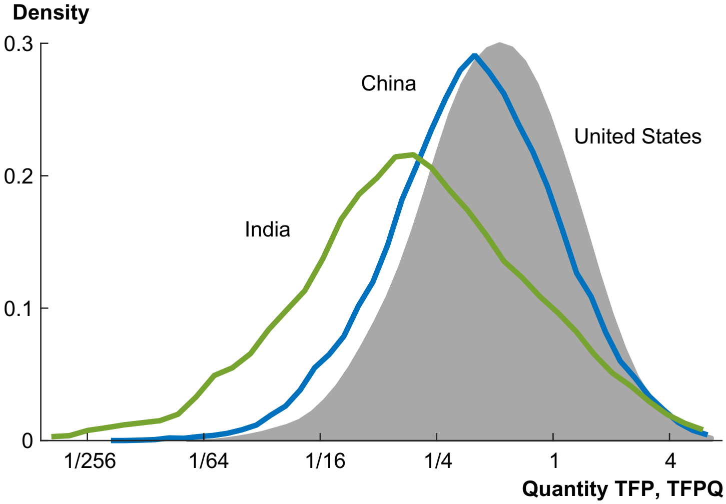


Fig. 35 The distribution of TFPQ in 4-digit manufacturing industries. *Note:* This is the average distribution of TFPQ within 4-digit manufacturing industries for the United States in 1997, China in 2005, and India in 1994, computed as described in the text. The means across countries are not meaningful. Source: *Hsieh, C.T., Klenow, P.J. 2009. Misallocation and manufacturing TFP in China and India. Q. J. Econ. 124 (4), 1403–1448; data provided by Chang Hsieh.*

TFPQ reflects both A_i and α_i—these are both part of fundamental productivity in this economic environment. In contrast, the somewhat conventional measure TFPR is actually independent of A_i and α_i.[aj]

4.11 The Hsieh–Klenow Facts

To quantify the effect of misallocation on aggregate TFP, Hsieh and Klenow (2009) use the insight that TFPR should be equated across plants if resources are allocated optimally. In particular, they use variation in TFPR across plants within 4-digit manufacturing industries to measure misallocation in the United States, China, and India.

The first part of their approach has already been explained in the previous section: they assume CES demand and use the constant elasticity to back out prices and real output from sales revenue. This allows them to measure "true" TFP, called "TFPQ," for each establishment in their data. The average distribution of TFPQ within 4 digit manufacturing that they recover is shown in Fig. 35.

As shown in this figure, the distributions within industry of TFPQ in the United States and China are relatively similar, while the distribution is significantly wider in

[aj] In richer settings, TFPR can depend on A_i, for example if there are fixed costs in production.

India.[ak] What is surprising, perhaps, is just how large the differences in TFPQ are within an industry. Hsieh and Klenow (2009) find that the 90–10 ratio of TFPQ across plants is 8.8 in the United States, 22.4 in India, and 11.5 in China. One way of thinking about these large differences is to note that employment differences across plants is very large. Why does a large textile manufacturer employ many more workers than a family shop? One answer is that TFPQ is much higher for the large plant.

Hsieh and Klenow's most valuable contribution, however, is to quantify misallocation. To see how they do this, consider a plant that produces with a Cobb–Douglas production function, using capital and labor, and that faces distortions τ_K and τ_L in choosing its inputs. These distortions are modeled as if they are taxes, but the literature interprets the distortions much more broadly to include credit market frictions, hiring and firing costs, quantity restrictions, and so on. The profit-maximizing firm will hire capital and labor until the marginal revenue product of these factors equals their gross-of-distortion rental price.[al] Written differently, payments to factors will equal the product of the factor exponents and the distortion terms:

$$\frac{rK_i}{p_i Y_i} = \alpha_K \cdot \frac{1}{1+\tau_K} \tag{14}$$

and

$$\frac{wL_i}{p_i Y_i} = \alpha_L \cdot \frac{1}{1+\tau_L}. \tag{15}$$

Roughly speaking, Hsieh and Klenow observe the left-hand side of these expressions in the data—they observe the share of revenues spent on labor and capital. They then use these observed spending shares to back out the distortions.

The key identification issue in recovering the τ_K and τ_L is this: when a manufacturing plant pays a large fraction of its revenue to labor, is that because it faces a low τ_L, or is that because its technology is labor intensive (a high α_L)? Hsieh and Klenow solve this identification problem by assuming that all firms within a 4-digit industry have common α_K and α_L Cobb–Douglas exponents. Then variation in factor shares across plants reflects distortions rather than technologies. This is one of the reasons why their approach works well within industries but would run into problems across industries.

TFPR is a summary measure of distortions, equal to a weighted average of the marginal revenue product of capital and the marginal revenue product of labor, relative to the

[ak] However, the authors note that small firms are underrepresented in the Chinese data, so this could reflect differences in the sample.

[al] For example, assuming no depreciation, the firm's profit maximization problem can be written as

$$\max_{K,L} \;\; p_i F(K_i, L_i) - w(1+\tau_L)L_i - r(1+\tau_K)K_i.$$

The first order conditions are then given in the text.

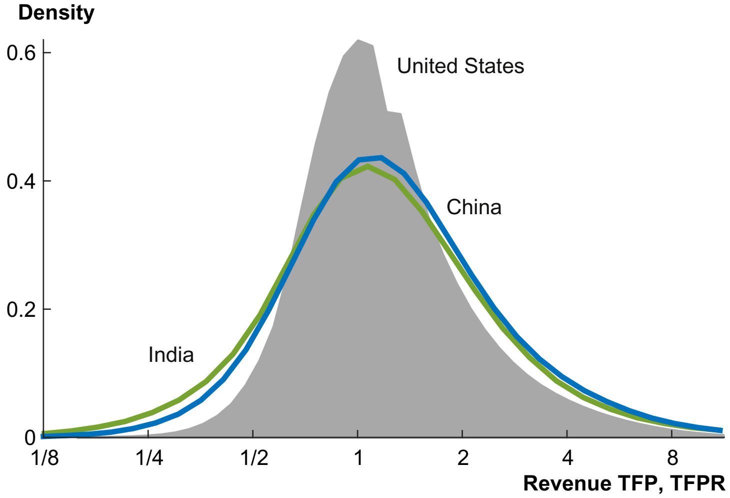


Fig. 36 The distribution of TFPR in 4-digit manufacturing industries. *Note:* This is the average distribution of TFPR within 4-digit manufacturing industries for the United States in 1997, China in 2005, and India in 1994, computed as described in the text. Source: *Hsieh, C.T., Klenow, P.J. 2009. Misallocation and manufacturing TFP in China and India. Q. J. Econ. 124 (4), 1403–1448; data provided by Chang Hsieh.*

average values in the industry. With no distortions, TFPR would take a value of one, as marginal revenue products get equated across firms in an efficient allocation. In the presence of distortions, TFPR equals a weighted average of the $1 + \tau_K$ and $1 + \tau_L$ distortions, where the weights are the Cobb–Douglas exponents in the production function.[am]

The average distribution of TFPR within 4-digit manufacturing industries is shown in Fig. 36. The first thing to note about this figure is that TFPR is not equal to one for every firm, not even in the United States. One interpretation of this fact is that resources are misallocated even in the United State and the deviations from unity can be used to measure US misallocation. An alternative interpretation is that there is measurement error in the US data, and some of the deviations reflect this measurement error. Both interpretations presumably have merit.

The second point to note in Fig. 36 is that the dispersion of TFPR in India and China is significantly larger than the dispersion in the United States. To the extent that this does not reflect larger measurement error in India and China, it suggests that the misallocation of capital and labor across establishments within 4-digit industries in China and India is a factor reducing GDP in those economies. Hsieh and Klenow quantify these effects and find that if China and India had the same dispersion of TFPR as the United States, their aggregate TFP would be higher by 30% to 50% in China and 40% to

[am] For aggregating the τ_K and τ_L into a single index of "TFPR" and to measure the effect of the distortions on TFP, Hsieh and Klenow therefore require values for the Cobb–Douglas exponents. They assume the US average shares are undistorted and assume China and India have the same Cobb–Douglas technology as the United States.

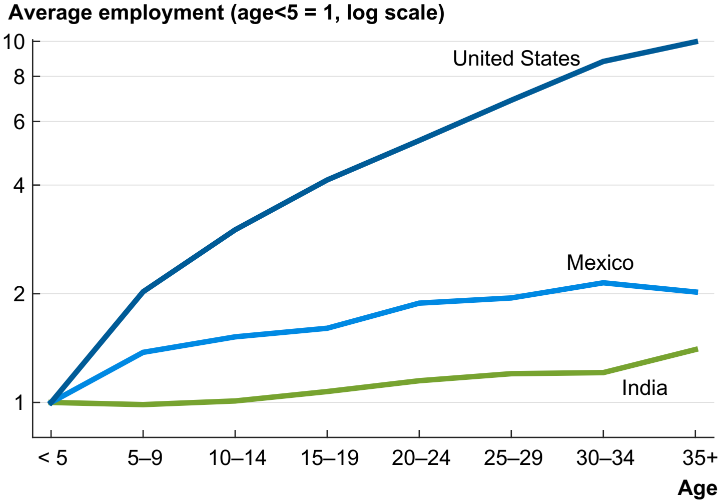


Fig. 37 Average employment over the life cycle. *Note:* The graph compares average employment per surviving plant in a later year to average employment per operating plant in an earlier year from the same cohort using census data for the manufacturing industry in the United States, Mexico, and India. Source: *Hsieh, C.T., Klenow, P.J. 2014. The life cycle of plants in India and Mexico. Q. J. Econ. 129 (3), 1035–1084; data provided by Chang Hsieh.*

60% in India. Long-run GDP would be higher by approximately twice this amount as capital accumulates in response to the higher TFP.

In a recent follow-up paper, Hsieh and Klenow attempt to understand what could be causing this misallocation. Hsieh and Klenow (2014) looks at how establishments in the United States, India, and Mexico grow as they age. Their remarkable finding is summarized in Fig. 37: plants in the United States get much larger as they age, while this is barely true at all in India.

To be more precise, plants that are more than 35 years old in the United States have more than 8 times the employment of plants that are less than 5 years old. In contrast, old plants in Mexico are only twice as large as young plants, while plants in India exhibit even less employment growth. The suggestion, explored in detail in this paper, is that distortions in Mexico and India prevent the most productive plants from growing in size, and this is one cause of the lower aggregate TFP in these economies. Hsieh and Klenow estimate that moving from the US life cycle to the Indian or Mexican life cycle of plant growth could reduce aggregate TFP by about 25%.

Motivated by facts like these, a growing number of recent papers explore various kinds of misallocation their effects on TFP. Asker et al. (2011) examine the role of volatility and adjustment costs in explaining variation in TFPR and TFPQ. Buera et al. (2011a), Midrigan and Xu (2014) and Moll (2014) study the extent to which credit market frictions can generate misallocation and TFP losses. Peters (2013) considers the role of heterogeneous markups in accounting for misallocation. Guner et al. (2008), Gourio and Roys (2014), and Garicano et al. (2014) consider the effect of regulations

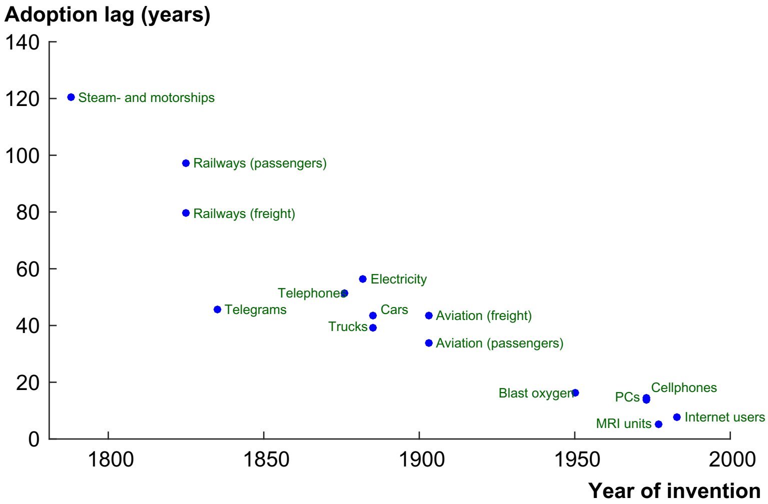


Fig. 38 Technology adoption is speeding up over time. *Note:* Adoption lags for each country measure the amount of time between when a technology is invented and when it was adopted in the country. The figure reports averages estimated across 166 countries spanning the period 1820–2003. Source: *Comin, D., Hobijn, B. 2010. An exploration of technology diffusion. Am. Econ. Rev. 100 (5), 2031–2059. doi:10.1257/aer.100.5.2031.*

tied to the size of firms. Akcigit et al. (2014a) suggest that incentive problems for managers limit the ability of potentially highly-productive small firms to expand, leading to lower TFP. Hopenhayn (2014) and Buera et al. (2015) provide excellent overviews of the recent literature.

4.12 The Diffusion of Ideas

Fig. 38 shows our next fact: lags in the adoption of new technologies have declined sharply over the last 200 years. This fact is taken from Comin and Hobijn (2010) and is based on the CHAT ("Cross-country Historical Adoption of Technology") database that these authors previous assembled. The database contains information on the diffusion of more than 100 technologies, in more than 150 countries, since 1800.

For 15 technologies, the graph plots the year of invention of each technology vs an average adoption lag across the sample of countries. More precisely, the adoption lag for each country/technology observation measures the number of years between the date a technology was invented and the date it was adopted in the country. The adoption lag shown in Fig. 38 is the average of this statistic across 166 countries.

A strong negative correlation is evident in the graph, suggesting the fact that technology adoption lags have been shrinking over time. Comin and Hobijn (2010) estimate that technologies invented 10 years later are on average adopted 4.3 years faster.

4.13 Urbanization

We remarked earlier on the inverse correlation between urbanization in the year 1500 and income per person today. Fig. 39 highlights the large trend in urbanization over time.

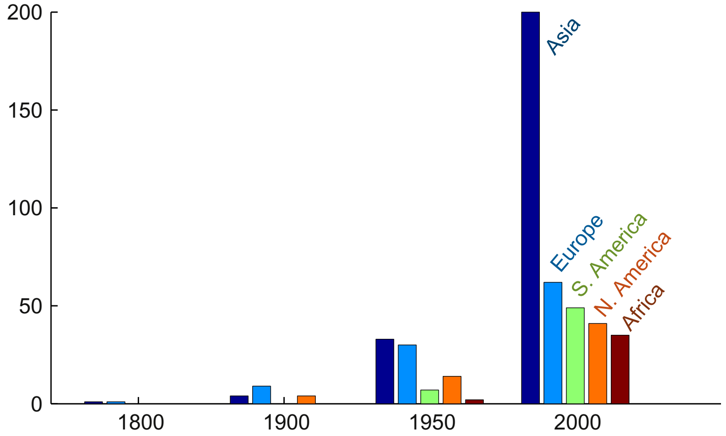


Fig. 39 The number of "million cities." *Note:* The histogram shows the number of cities on each continent with populations greater than 1 million. Oceania is included with Asia. Source: *Satterthwaite, D. 2005. The scale of urban change worldwide 1950–2000 and its underpinnings. International Institute for Environment and Development, unpublished paper, table 3.*

The figure shows the number of cities containing more than a million people, by continent, since 1800. This figure also emphasizes the extent to which urbanization is even stronger in Asia than in Europe and North America, at least as measured by the number of large cities.

5. CONCLUSION

While this paper has covered a large number of facts, there are still important facts that I have neglected. Partly this is because the paper is long enough, and partly this is because some of these facts are not as fully established as we might like. Therefore, this final discussion also can be read as a suggestion for places where further research might be particularly helpful.

One of the important facts omitted here is globalization and its effects on growth: the decline of trade barriers, the decline in transportation and communication costs, the rise of vertical supply chains. There is an old stylized fact from Sachs and Warner (1995) that "open" economies grew more rapidly over the period 1970 to 1989 than countries that were "closed." However, the definition they used of open and closed is so broad as to include other forces, and this dichotomy is surely correlated with institutional quality more generally.[an]

There are many other facts related to institutions that we wish we understood better. It is relatively clear from the kinds of evidence reviewed earlier that "institutions matter." But how important are different institutions, and how do institutions change?

[an] See also Ben-David (1993), Frankel and Romer (1999), and Feyrer (2009).

Is democracy conducive to growth, or does democracy typically result from growth, or both (Barro, 1999; Acemoglu et al., 2014)? Does human capital accumulation lead to good institutions, or vice versa, or both (Glaeser et al., 2004)? What is the relationship between culture, "fractionalization," institutions, and growth (Alesina et al., 2003)?

Another fact that we'd like to know more about is the extent of knowledge spillovers across countries. It is well appreciated that each country benefits from knowledge created elsewhere in the world, but quantifying these benefits is difficult. Eaton and Kortum (1999) suggest that only 60% of US growth in recent decades comes from knowledge created in the United States, and the numbers for local knowledge in Japan (35%) and the United Kingdom (13%) are even smaller. A related fact courtesy of Freeman (2010) is this: In the 1970s, China produced a negligible number of Ph.D.'s in science and engineering, but by 2010, China was producing 26% *more* than the United States. In a world of ideas, the economic development of China and India may have a profound effect on growth in the future. How many future Thomas Edisons, Albert Einsteins, Steve Jobs, and Sam Waltons are out there, waiting for their talents to be appropriately nurtured?

ACKNOWLEDGMENTS

I am grateful to Kevin Bryan, Antonio Ciccone, John Cochrane, John Haltiwanger, Jonathan Haskel, Bart Hobijn, Robert Inklaar, Jihee Kim, Pete Klenow, Marti Mestieri, John Seater, Chris Tonetti, seminar participants at the Stanford Handbook conference, and the editors John Taylor and Harald Uhlig for helpful comments. All the data used in this paper can be obtained in the file FactsofGrowth-Replication.zip, available on the author's web page.

REFERENCES

Abramovitz, M., 1956. Resource and output trends in the united states since 1870. Am. Econ. Assoc. Pap. Proc. 46 (2), 5–23.

Acemoglu, D., 1998. Why do new technologies complement skills? Directed technical change and wage inequality. Q. J. Econ. 113, 1055–1089.

Acemoglu, D., Robinson, J.A., 2012. Why Nations Fail: The Origins of Power, Prosperity and Poverty. Crown Business.

Acemoglu, D., Johnson, S., Robinson, J.A., 2001. The colonial origins of comparative development: an empirical investigation. Am. Econ. Rev. 91 (5), 1369–1401.

Acemoglu, D., Johnson, S., Robinson, J.A., 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. Q. J. Econ. 117 (4), 1231–1294.

Acemoglu, D., Akcigit, U., Bloom, N., Kerr, W.R., 2013. Innovation, reallocation and growth. National Bureau of Economic Research, Inc. http://ideas.repec.org/p/nbr/nberwo/18993.html.

Acemoglu, D., Naidu, S., Restrepo, P., Robinson, J.A., 2014. Democracy does cause growth. National Bureau of Economic Research, Inc. http://ideas.repec.org/p/nbr/nberwo/20004.html.

Aghion, P., Howitt, P., 1992. A model of growth through creative destruction. Econometrica 60 (2), 323–351.

Akcigit, U., Alp, H., Peters, M., 2014a. Lack of selection and limits to delegation: firms dynamics in developing countries.

Akcigit, U., Celik, M.A., Greenwood, J., 2014b. Buy, keep or sell: economic growth and the market for ideas. University of Pennsylvania manuscript.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., Wacziarg, R., 2003. Fractionalization. J. Econ. Growth 8 (2), 155–194.

Aoki, S., Nirei, M., 2013. Pareto distributions and the evolution of top incomes in the U.S. University Library of Munich, Germany. http://ideas.repec.org/p/pra/mprapa/47967.html.

Asker, J., Collard-Wexler, A., Loecker, J.D., 2011. Productivity volatility and the misallocation of resources in developing economies. National Bureau of Economic Research, Inc. http://ideas.repec.org/p/nbr/nberwo/17175.html.

Atkinson, A.B., Piketty, T., Saez, E., 2011. Top incomes in the long run of history. J. Econ. Lit. 49 (1), 3–71.

Autor, D.H., Levy, F., Murnane, R.J., 2003. The skill content of recent technological change: an empirical exploration. Q. J. Econ. 118 (4), 1279–1333. http://ideas.repec.org/a/tpr/qjecon/v118y2003i4p1279-1333.html.

Banerjee, A.V., Duflo, E., 2005. Growth theory through the lens of development economics. In: Aghion, P., Durlauf, S.A. (Eds.), Handbook of Economic Growth. North Holland, New York, NY, pp. 473–552.

Barro, R.J., 1991. Economic growth in a cross section of countries. Q. J. Econ. 106, 407–443.

Barro, R.J., 1999. Determinants of democracy. J. Polit. Econ. 107 (S6), S158–S183.

Barro, R.J., 2012. Convergence and modernization revisited. National Bureau of Economic Research, Inc. http://ideas.repec.org/p/nbr/nberwo/18295.html.

Barro, R.J., Becker, G.S., 1989. Fertility choice in a model of economic growth. Econometrica 57 (2), 481–501.

Barro, R.J., Sala-i-Martin, X., 1992. Convergence. J. Polit. Econ. 100 (2), 223–251.

Barro, R., Lee, J.W., 2013. A new data set of educational attainment in the world, 1950-2010. J. Dev. Econ. 104 (C), 184–198. http://EconPapers.repec.org/RePEc:eee:deveco:v:104:y:2013:i:c:p:184-198.

Baumol, W.J., 1986. Productivity growth, convergence and welfare: what the long-run data show. Am. Econ. Rev. 76, 1072–1085.

Becker, G.S., Philipson, T.J., Soares, R.R., 2005. The quantity and quality of life and the evolution of world inequality. Am. Econ. Rev. 95 (1), 277–291. http://ideas.repec.org/a/aea/aecrev/v95y2005i1p277-291.html.

Ben-David, D., 1993. Equalizing exchange: trade liberalization and income convergence. Q. J. Econ. 108 (3), 653–680.

Benhabib, J., Bisin, A., Zhu, S., 2011. The distribution of wealth and fiscal policy in economies with finitely lived agents. Econometrica 79 (1), 123–157. http://ideas.repec.org/a/ecm/emetrp/v79y2011i1p123-157.html.

Besley, T., Ilzetzki, E., Persson, T., 2013. Weak states and steady states: the dynamics of fiscal capacity. Am. Econ. J. Macroecon. 5 (4), 205–235. http://dx.doi.org/10.1257/mac.5.4.205.

Biskupic, J., 2006. Sandra Day O'Connor: How the First Woman on the Supreme Court Became Its Most Influential Justice. HarperCollins, New York, NY.

Bloom, N., Schankerman, M., Reenen, J.V., 2013. Identifying technology spillovers and product market rivalry. Econometrica 81 (4), 1347–1393. http://ideas.repec.org/a/ecm/emetrp/v81y2013i4p1347-1393.html.

Boppart, T., 2014. Structural change and the Kaldor facts in a growth model with relative price effects and non-Gorman preferences. Econometrica 82, 2167–2196. http://ideas.repec.org/a/wly/emetrp/v82y2014ip2167-2196.html.

Bourguignon, F., Morrisson, C., 2002. Inequality among world citizens: 1820-1992 inequality among world citizens: 1820-1992. Am. Econ. Rev. 92 (4), 727–744.

Bridgman, B., 2014. Is labor's loss capital's gain? Gross versus net labor shares. Bureau of Economic Analysis manuscript.

Brynjolfsson, E., McAfee, A., 2012. Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity and Irreversibly Transforming Employment and the Economy. Digital Frontier Press. http://books.google.com/books?id=IhArMwEACAAJ. ISBN 9780984725113.

Buera, F.J., Kaboski, J.P., Shin, Y., 2011a. Finance and development: a tale of two sectors. Am. Econ. Rev. 101 (5), 1964–2002. http://dx.doi.org/10.1257/aer.101.5.1964.

Buera, F.J., Monge-Naranjo, A., Primiceri, G.E., 2011b. Learning the wealth of nations. Econometrica 1468-0262. 79 (1), 1–45. http://dx.doi.org/10.3982/ECTA8299.

Buera, F.J., Kaboski, J.P., Shin, Y., 2015. Entrepreneurship and financial frictions: a macro-development perspective. National Bureau of Economic Research, Inc. https://ideas.repec.org/p/nbr/nberwo/21107.html.

Byrne, D.M., Oliner, S.D., Sichel, D.E., 2013. Is the information technology revolution over? Intern. Product. Monit. 25, 20–36. http://ideas.repec.org/a/sls/ipmsls/v25y20133.html.

Caballero, R.J., Jaffe, A.B., 1993. How high are the giants' shoulders? In: Blanchard, O., Fischer, S. (Eds.), NBER Macroeconomics Annual. MIT Press, Cambridge, MA, pp. 15–74.

Cagetti, M., Nardi, M.D., 2006. Entrepreneurship, frictions, and wealth. J. Polit. Econ. 114 (5), 835–870. http://ideas.repec.org/a/ucp/jpolec/v114y2006i5p835-870.html.

Caselli, F., 1999. Technological revolutions. Am. Econ. Rev. 89 (1), 78–102. http://ideas.repec.org/a/aea/aecrev/v89y1999i1p78-102.html.

Caselli, F., 2005. Accounting for cross country income differences. In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth. Elsevier.

Caselli, F., Ciccone, A., 2013. The contribution of schooling in development accounting: results from a nonparametric upper bound. J. Dev. Econ. 104 (C), 199–211. https://ideas.repec.org/a/eee/deveco/v104y2013icp199-211.html.

Caselli, F., Coleman, W.J., 2006. The world technology frontier. Am. Econ. Rev. 96 (3), 499–522.

Caselli, F., Feyrer, J., 2007. The marginal product of capital. Q. J. Econ. 122 (2), 535–568.

Caselli, F., Wilson, D.J., 2004. Importing technology. J. Monet. Econ. 51 (1), 1–32. http://ideas.repec.org/a/eee/moneco/v51y2004i1p1-32.html.

Caselli, F., Esquivel, G., Lefort, F., 1996. Reopening the convergence debate: a new look at cross-country growth empirics. J. Econ. Growth 1, 363–390.

Cass, D., 1965. Optimal growth in an aggregative model of capital accumulation. Rev. Econ. Stud. 32, 233–240.

Castaneda, A., Diaz-Gimenez, J., Rios-Rull, J.V., 2003. Accounting for the U.S. earnings and wealth inequality. J. Polit. Econ. 111 (4), 818–857. http://ideas.repec.org/a/ucp/jpolec/v111y2003i4p818-857.html.

Chari, V., Kehoe, P., McGrattan, E., 2007. Business cycle accounting. Econometrica 75 (3), 781–836.

Christensen, L.R., Cummings, D., Jorgenson, D.W., 1981. Relative productivity levels, 1947-1973. Eur. Econ. Rev. 16 (1), 61–94.

Clark, G., 2001. The secret history of the industrial revolution. U.C. Davis mimeo.

Clark, G., 2014. The industrial revolution. In: Handbook of Economic Growth, vol. 2. Elsevier, pp. 217–262. http://ideas.repec.org/h/eee/grochp/2-217.html.

Coe, D.T., Helpman, E., 1995. International R&D spillovers. Eur. Econ. Rev. 39 (5), 859–887.

Comin, D., Hobijn, B., 2010. An exploration of technology diffusion. Am. Econ. Rev. 100 (5), 2031–2059. http://dx.doi.org/10.1257/aer.100.5.2031.

Comin, D., Lashkari, D., Mestieri, M., 2015. Structural transformations with long-run income and price effects. Dartmouth College, unpublished manuscript.

Cordoba, J.C., Ripoll, M., 2014. The elasticity of intergenerational substitution, parental altruism, and fertility choice. Iowa State University, Department of Economics. http://ideas.repec.org/p/isu/genres/37766.html.

Dalgaard, C.J., Strulik, H., 2014. Optimal aging and death: understanding the Preston curve. J. Eur. Econ. Assoc. 12 (3), 672–701. http://ideas.repec.org/a/bla/jeurec/v12y2014i3p672-701.html.

Dell, M., 2010. The persistent effects of Peru's Mining Mita. Econometrica 78 (6), 1863–1903. http://ideas.repec.org/a/ecm/emetrp/v78y2010i6p1863-1903.html.

DeLong, J.B., 1988. Productivity growth, convergence, and welfare: comment. Am. Econ. Rev. 78, 1138–1154.

Denison, E.F., 1967. Why Growth Rates Differ. The Brookings Institution, Washington, DC.

Diamond, J., 1997. Guns, Germs, and Steel. W.W. Norton and Co., New York, NY.

Doepke, M., 2005. Child mortality and fertility decline: does the Barro-Becker model fit the facts? J. Popul. Econ. 18 (2), 337–366. http://ideas.repec.org/a/spr/jopoec/v18y2005i2p337-366.html.

Eaton, J., Kortum, S.S., 1999. International technology diffusion: theory and measurement. Int. Econ. Rev. 40, 537–570.

Elsby, M.W.L., Hobijn, B., Şahin, A., 2013. The decline of the U.S. labor share. Brook. Pap. Econ. Act. 2013 (2), 1–63.

Engerman, S.L., Sokoloff, K.L., 1997. Factor endowments, institutions, and differential paths of growth among new world economies. In: Haber, S. (Ed.), How Latin America Fell Behind. Stanford University Press, Stanford, CA.

Erosa, A., Koreshkova, T., Restuccia, D., 2010. How important is human capital? A quantitative theory assessment of world income inequality. Rev. Econ. Stud. 77 (4), 1421–1449. http://ideas.repec.org/a/bla/restud/v77y2010i4p1421-1449.html.

Feenstra, R.C., Inklaar, R., Timmer, M.P., 2015. The next generation of the Penn World Table. Am. Econ. Rev. 105 (10), 3150–3182. http://dx.doi.org/10.1257/aer.20130954.

Fernald, J., 2014. Productivity and potential output before, during, and after the great recession. In: NBER Macroeconomics Annual 2014, Volume 29, University of Chicago Press.

Fernald, J.G., Jones, C.I., 2014. The future of US economic growth. Am. Econ. Rev. Pap. Proc. 104 (5), 44–49. http://ideas.repec.org/a/aea/aecrev/v104y2014i5p44-49.html.

Feyrer, J., 2009. Trade and income–exploiting time series in geography. National Bureau of Economic Research, Inc. https://ideas.repec.org/p/nbr/nberwo/14910.html.

Fleurbaey, M., Gaulier, G., 2009. International comparisons of living standards by equivalent incomes. Scand. J. Econ. 111 (3), 597–624. http://ideas.repec.org/a/bla/scandj/v111y2009i3p597-624.html.

Foster, L., Haltiwanger, J., Syverson, C., 2008. Reallocation, firm turnover, and efficiency: selection on productivity or profitability? Am. Econ. Rev. 98 (1), 394–425. http://ideas.repec.org/a/aea/aecrev/v98y2008i1p394-425.html.

Frankel, J.A., Romer, D., 1999. Does trade cause growth? Am. Econ. Rev. 89 (3), 379–399.

Freeman, R.B., 2010. What does global expansion of higher education mean for the United States? In: American Universities in a Global Market, University of Chicago Press, pp. 373–404.

Galor, O., 2005. From stagnation to growth: unified growth theory. Handb. Econ. Growth 1, 171–293.

Galor, O., Weil, D., 2000. Population, technology, and growth: from the malthusian regime to the demographic transition. Am. Econ. Rev. 90, 806–828.

Galor, O., Weil, D.N., 1996. The gender gap, fertility, and growth. Am. Econ. Rev. 86 (3), 374–387.

Garicano, L., Lelarge, C., Van Reenen, J., 2014. Firm size distortions and the productivity distribution: evidence from France. .

Glaeser, E.L., La Porta, R., Lopez-de Silanes, F., Shleifer, A., 2004. Do institutions cause growth? J. Econ. Growth 9 (3), 271–303.

Gollin, D., 2002. Getting income shares right. J. Polit. Econ. 110 (2), 458–474.

Gourio, F., Roys, N., 2014. Size-dependent regulations, firm size distribution, and reallocation. Quant. Econ. 5, 377–416. https://ideas.repec.org/a/wly/quante/v5y2014ip377-416.html.

Greenwood, J., Hercowitz, Z., Krusell, P., 1997. Long-run implications of investment-specific technological change. Am. Econ. Rev. 87 (3), 342–362.

Greenwood, J., Seshadri, A., Vandenbroucke, G., 2005. The baby boom and baby bust. Am. Econ. Rev. 95 (1), 183–207. http://ideas.repec.org/a/aea/aecrev/v95y2005i1p183-207.html.

Griliches, Z., 1988. Productivity puzzles and R&D: another nonexplanation. J. Econ. Perspect. 2, 9–21.

Griliches, Z., 1992. The search for R&D spillovers. Scand. J. Econ. 94, 29–47.

Griliches, Z., 1994. Productivity, R&D and the data constraint. Am. Econ. Rev. 84 (1), 1–23.

Grossman, G.M., Helpman, E., 1991. Innovation and Growth in the Global Economy. MIT Press, Cambridge, MA.

Guner, N., Ventura, G., Yi, X., 2008. Macroeconomic implications of size-dependent policies. Rev. Econ. Dyn. 11 (4), 721–744. http://ideas.repec.org/a/red/issued/07-73.html.

Hall, R.E., Jones, C.I., 1996. The productivity of nations. NBER Working Paper No. 5812.

Hall, R.E., Jones, C.I., 1999. Why do some countries produce so much more output per worker than others? Q. J. Econ. 114 (1), 83–116.

Hall, R.E., Jones, C.I., 2007. The value of life and the rise in health spending. Q. J. Econ. 122 (1), 39–72.

Hansen, G.D., Prescott, E.C., 2002. Malthus to solow. Am. Econ. Rev. 92 (4), 1205–1217.

Hanushek, E.A., Kimko, D.D., 2000. Schooling, labor-force quality, and the growth of nations. Am. Econ. Rev. 90 (5), 1184–1208.

Hanushek, E.A., Woessmann, L., 2008. The role of cognitive skills in economic development. J. Econ. Lit. 46 (3), 607–668. http://dx.doi.org/10.1257/jel.46.3.607.

Heisenberg, W., 1971. Physics and Beyond, Encounters and Conversations. Harper & Row, New York, NY.

Hemous, D., Olsen, M., 2014. The rise of the machines: automation, horizontal innovation and income inequality. Society for Economic Dynamics. http://ideas.repec.org/p/red/sed014/162.html.

Henderson, J.V., Storeygard, A., Weil, D.N., 2012. Measuring economic growth from outer space. Am. Econ. Rev. 102 (2), 994–1028. http://ideas.repec.org/a/aea/aecrev/v102y2012i2p994-1028.html.

Hendricks, L., Schoellman, T., 2014. Human capital and development accounting: new evidence from immigrant earnings. Society for Economic Dynamics. http://ideas.repec.org/p/red/sed014/702.html.

Herrendorf, B., Rogerson, R., Valentinyi, A., 2014. Growth and structural transformation. In: Handbook of Economic Growth, vol. 2. Elsevier, pp. 855–941. http://ideas.repec.org/h/eee/grochp/2-855.html.

Hopenhayn, H.A., 2014. Firms, misallocation, and aggregate productivity: a review. Ann. Rev. Econ. 6 (1), 735–770. https://ideas.repec.org/a/anr/reveco/v6y2014p735-770.html.

Hsieh, C.T., Klenow, P.J., 2009. Misallocation and manufacturing TFP in China and India. Q. J. Econ. 124 (4), 1403–1448.

Hsieh, C.T., Klenow, P.J., 2014. The life cycle of plants in India and Mexico. Q. J. Econ. 129 (3), 1035–1084.

Hsieh, C.T., Moretti, E., 2014. Growth in cities and countries. U.C. Berkeley manuscript, unpublished.

Hsieh, C.T., Hurst, E., Jones, C.I., Klenow, P.J., 2013. The allocation of talent and U.S. economic growth. Stanford University, unpublished paper.

Islam, N., 1995. Growth empirics: a panel data approach. Q. J. Econ. 110, 1127–1170.

Jaffe, A.B., Lerner, J., 2006. Innovation and its discontents. Capital. Soc. 1(3).

Jones, C.I., 1995. R&D-based models of economic growth. J. Polit. Econ. 103 (4), 759–784.

Jones, C.I., 2001. Was an industrial revolution inevitable? economic growth over the very long run. Adv. Macroecon. 1 (2). http://www.bepress.com\-/bejm\-/advances/vol1\-/iss2/art1. Article 1.

Jones, C.I., 2002. Sources of U.S. economic growth in a world of ideas. Am. Econ. Rev. 92 (1), 220–239.

Jones, C.I., 2003. Growth, capital shares, and a new perspective on production functions. U.C. Berkeley mimeo.

Jones, C.I., 2005. Growth and ideas. In: Aghion, P., Durlauf, S.A. (Eds.), Handbook of Economic Growth. North Holland, New York, NY, pp. 1063–1111.

Jones, B.F., 2014. The human capital stock: a generalized approach. Am. Econ. Rev. 104 (11), 3752–3777. http://ideas.repec.org/a/aea/aecrev/v104y2014i11p3752-77.html.

Jones, C.I., Kim, J., 2014. A Schumpeterian model of top income inequality. Stanford University manuscript.

Jones, C.I., Klenow, P.J., 2015. Beyond GDP: welfare across countries and time. Stanford University, unpublished manuscript.

Jones, C.I., Romer, P.M., 2010. The new Kaldor Facts: ideas, institutions, population, and human capital. Am. Econ. J. Macroecon. 1945-7707. 2 (1), 224–245.

Jones, C.I., Scrimgeour, D., 2008. A new proof of Uzawa's steady-state growth theorem. Rev. Econ. Stat. 90 (1), 180–182. http://ideas.repec.org/a/tpr/restat/v90y2008i1p180-182.html.

Jones, C.I., Williams, J.C., 1998. Measuring the social return to R&D. Q. J. Econ. 113 (4), 1119–1135.

Jones, L., Tertilt, M., 2016. An economic history of fertility in the U.S.: 1826-1960. In: Rupert, P. (Ed.), Frontiers of Family Economics. Emerald Press, pp. 165–230.

Jones, L.E., Schoonbroodt, A., Tertilt, M., 2010. Fertility theories: can they explain the negative fertility-income relationship? In: Demography and the Economy, National Bureau of Economic Research, Inc, pp. 43–100. NBER Chapters, http://ideas.repec.org/h/nbr/nberch/8406.html.

Kaldor, N., 1961. Capital accumulation and economic growth. In: Lutz, F., Hague, D. (Eds.), The Theory of Capital. St. Martins Press, pp. 177–222.

Kane, T., 2015. Accelerating convergence in the world income distribution. Stanford University, unpublished paper.

Karabarbounis, L., Neiman, B., 2014. The global decline of the labor share. Q. J. Econ. 129 (1), 61–103. http://ideas.repec.org/a/oup/qjecon/v129y2014i1p61-103.html.

Katz, L., Murphy, K., 1992. Changes in relative wages, 1963-1987: supply and demand factors. Q. J. Econ. 107 (1), 35–78.

King, R.G., Levine, R., 1994. Capital fundamentalism, economic development, and economic growth. Carnegie-Rochester Conf. Ser. Publ. Pol. 0167-2231. 40, 259–292. http://dx.doi.org/10.1016/0167-2231(94)90011-6.http://www.sciencedirect.com/science/article/pii/0167223194900116.

Klenow, P.J., Rodriguez-Clare, A., 1997. The neoclassical revival in growth economics: Has it gone too far? In: Bernanke, B.S., Rotemberg, J.J. (Eds.), NBER Macroeconomics Annual 1997. MIT Press, Cambridge, MA.

Klenow, P.J., Rodriguez-Clare, A., 2005. Extenalities and growth. In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth. Elsevier, Amsterdam.

Koh, D., Santaeulalia-Llopis, R., Zheng, Y., 2015. Labor share decline and intellectual property products capital. Washington University St. Louis, unpublished paper.

Koopmans, T.C., 1965. On the concept of optimal economic growth. In: The Econometric Approach to Development Planning, North Holland, Amsterdam.

Kortum, S.S., 1997. Research, patenting, and technological change. Econometrica 65 (6), 1389–1419.

Kremer, M., 1993. Population growth and technological change: one million B.C. to 1990. Q. J. Econ. 108 (4), 681–716.

Krugman, P., 1994. The myth of Asia's miracle. Fore. Aff. 73 (6), 62–78.

Krusell, P., Ohanian, L., Rios-Rull, J.V., Violante, G., 2000. Capital-skill complementarity and inequality: a macroeconomic analysis. Econometrica 68 (5), 1029–1053.

Lagakos, D., Moll, B., Porzio, T., Qian, N., Schoellman, T., 2012. Experience matters: human capital and development accounting. National Bureau of Economic Research, Inc. http://ideas.repec.org/p/nbr/nberwo/18602.html.

Lee, R.D., 1988. Induced population growth and induced technological progress: their interaction in the accelerating stage. Math. Popul. Stud. 1 (3), 265–288.

Lucas, R.E., 1988. On the mechanics of economic development. J. Monet. Econ. 22 (1), 3–42.

Lucas, R.E., 2000. Some macroeconomics for the 21st century. J. Econ. Perspect. 14 (1), 159–168.

Lucas, R.E., 2002. The industrial revolution: past and future. Lect. Econ. Growth, 109–188.

Lucas, R.E., 2009. Ideas and growth. Economica 76 (301), 1–19. http://ideas.repec.org/a/bla/econom/v76y2009i301p1-19.html.

Lucas, R.E., Moll, B., 2014. Knowledge growth and the allocation of time. J. Polit. Econ. 122 (1), 1–51.

Maddison, A., 1979. Per capita output in the long run. Kyklos 32, 412–419.

Maddison, A., 1995. Monitoring the World Economy 1820-1992. Organization for Economic Cooperation and Development, Paris.

Mankiw, N., Romer, D., Weil, D., 1992. A contribution to the empirics of economic growth. Q. J. Econ. 107 (2), 407–438.

Manuelli, R.E., Seshadri, A., 2014. Human capital and the wealth of nations. Am. Econ. Rev. 104 (9), 2736–2762. http://dx.doi.org/10.1257/aer.104.9.2736.

Michalopoulos, S., Papaioannou, E., 2014. National institutions and subnational development in Africa. Q. J. Econ. 129 (1), 151–213. http://ideas.repec.org/a/oup/qjecon/v129y2014i1p151-213.html.

Midrigan, V., Xu, D.Y., 2014. Finance and misallocation: evidence from plant-level data. Am. Econ. Rev. 104 (2), 422–458. http://ideas.repec.org/a/aea/aecrev/v104y2014i2p422-58.html.

Mokyr, J., 1990. The Lever of Riches. Oxford University Press, New York, NY.

Moll, B., 2014. Productivity losses from financial frictions: can self-financing undo capital misallocation? Am. Econ. Rev. 104 (10), 3186–3221.

Murphy, K.M., Topel, R.H., 2006. The value of health and longevity. J. Polit. Econ. 114 (5), 871–904.

Nordhaus, W.D., 2003. The health of nations: the contribution of improved health to living standards. In: Murphy, K.M., Topel, R. (Eds.), Measuring the Gains from Medical Research: An Economic Approach. University of Chicago Press, Chicago, IL, pp. 9–40.

Nordhaus, W.D., Tobin, J., 1972. Is growth obsolete? In: Economic Research: Retrospect and Prospect, Vol 5: Economic Growth, National Bureau of Economic Research, Inc, pp. 1–80. http://ideas.repec.org/h/nbr/nberch/7620.html.

OECD, 2002. Frascati Manual 2002. OECD Publishing. http://dx.doi.org/10.1787/9789264199040-en.

Oeppen, J., Vaupel, J.W., 2002. Broken limits to life expectancy. Science 296 (5570), 1029–1031.

Olson, M., 1996. Big bills left on the sidewalk: why some nations are rich, and others poor. J. Econ. Perspect. 10 (2), 3–24.

Peters, M., 2013. Heterogeneous mark-ups, growth, and endogenous misallocation. Yale University manuscript.

Piketty, T., 2014. Capital in the Twenty-first Century. Harvard University Press.

Piketty, T., Saez, E., 2003. Income inequality in the united states, 1913-1998. Q. J. Econ. 118 (1), 1–39. http://ideas.repec.org/a/tpr/qjecon/v118y2003i1p1-39.html.

Piketty, T., Saez, E., Stantcheva, S., 2014. Optimal taxation of top labor incomes: a tale of three elasticities. Am. Econ. J. Econ. Pol. 6 (1), 230–271. http://ideas.repec.org/a/aea/aejpol/v6y2014i1p230-71.html.

Pomeranz, K., 2009. The Great Divergence: China, Europe, and the Making of the Modern World Economy. Princeton University Press.

Prescott, E.C., 1998. Needed: a theory of total factor productivity. Int. Econ. Rev. 39 (3), 525–551.

Prescott, E.C., 2004. Why do americans work so much more than Europeans? Q. Rev. 28, 2–13. http://ideas.repec.org/a/fip/fedmqr/y2004ijulp2-13nv.28no.1.html.

Pritchett, L., 1997. Divergence: big time. J. Econ. Perspect. 11 (3), 3–17.

Pritchett, L., Aiyar, Y., 2015. Taxes: price of civilization or tribute to leviathan? Center for Global Development Working Paper 412.

Quah, D., 1993. Empirical cross-section dynamics in economic growth. Eur. Econ. Rev. 37, 426–434.

Ramey, V.A., Francis, N., 2009. A century of work and leisure. Am. Econ. J. Macroecon. 1 (2), 189–224. http://ideas.repec.org/a/aea/aejmac/v1y2009i2p189-224.html.

Ramsey, F., 1928. A mathematical theory of saving. Econ. J. 38, 543–559.

Restuccia, D., Rogerson, R., 2008. Policy distortions and aggregate productivity with heterogeneous plants. Rev. Econ. Dyn. 11, 707–720.

Restuccia, D., Vandenbroucke, G., 2013. The evolution of education: a macroeconomic analysis. Int. Econ. Rev. 1468-2354. 54 (3), 915–936. http://dx.doi.org/10.1111/iere.12022.

Restuccia, D., Yang, D.T., Zhu, X., 2008. Agriculture and aggregate productivity: a quantitative cross-country analysis. J. Monet. Econ. 55 (2), 234–250.

Rognlie, M., 2015. Deciphering the fall and rise in the net capital share. In: Brookings Papers on Economic Activity, Conference Draft, March.

Romer, P.M., 1986. Increasing returns and long-run growth. J. Polit. Econ. 94, 1002–1037.

Romer, P.M., 1990. Endogenous technological change. J. Polit. Econ. 98 (5), S71–S102.

Rosenberg, N., 1994. Exploring the Black Box: Technology, Economics, and History. Cambridge University Press, New York, NY.

Sachs, J.D., Warner, A., 1995. Economic reform and the process of global integration. Brook. Pap. Econ. Act. 1, 1–95.

Saez, E., Zucman, G., 2014. Wealth inequality in the United States since 1913. U.C. Berkeley slides.

Sala-i-Martin, X., 2006. The world distribution of income: falling poverty and convergence, period. Q. J. Econ. 121 (2), 351–397.

Schlicht, E., 2006. A variant of Uzawa's theorem. Econ. Bull. 5 (6), 1–5. http://www.economicbulletin.com/2006/volume5/EB-06E10001A.pdf.

Segerstrom, P., 1998. Endogenous growth without scale effects. Am. Econ. Rev. 88 (5), 1290–1310.

Simon, J.L., 1981. The Ultimate Resource. Princeton University Press, Princeton, NJ.

Solow, R.M., 1956. A contribution to the theory of economic growth. Q. J. Econ. 70 (1), 65–94.

Solow, R.M., 1957. Technical change and the aggregate production function. Rev. Econ. Stat. 39 (3), 312–320.

Stiglitz, J.E., Sen, A., Fitoussi, J.P., 2009. Report by the commission on the measurement of economic performance and social progress. .

Stokey, N.L., Rebelo, S., 1995. Growth effects of flat-rate taxes. J. Polit. Econ. 103, 519–550.

Syverson, C., 2011. What determines productivity? J. Econ. Lit. 49 (2), 326–365. http://ideas.repec.org/a/aea/jeclit/v49y2011i2p326-65.html.

Weil, D.N., 2007. Accounting for the effect of health on economic growth. Q. J. Econ. 122 (3), 1265–1306.

Weitzman, M.L., 1998. Recombinant growth. Q. J. Econ. 113, 331–360.

Whelan, K., 2003. A two-sector approach to modeling U.S. NIPA data. J. Money, Credit, Bank. 35 (4), 627–656.

Wolfe, R.M., 2014. Business R&D performance in the United States tops $300 billion in 2012. NCSES Info Brief, NSF, 15-303.

Yang, J., 2004. Colonial legacy and modern economic growth in Korea: a critical examination of their relationships. Devel. Soc. 33 (1), 1–24.

Young, A., 1992. A tale of two cities: Factor accumulation and technical change in Hong Kong and Singapore. In: Blanchard, O., Fischer, S. (Eds.), NBER Macroeconomics Annual. MIT Press, Cambridge, MA, pp. 13–54.

Young, A., 1995. The tyranny of numbers: confronting the statistical realities of the East Asian growth experience. Q. J. Econ. 110 (3), 641–680.

Young, A., 2012. The African growth miracle. J. Polit. Econ. 120 (4), 696–739.

Williams, L.H., 2013. Intellectual property rights and innovation: evidence from the human genome. J. Polit. Econ. 121 (1), 1–27.

Zeira, J., 1998. Workers, machines, and economic growth. Q. J. Econ. 113 (4), 1091–1117. http://ideas.repec.org/a/tpr/qjecon/v113y1998i4p1091-1117.html.

CHAPTER 2

Macroeconomic Shocks and Their Propagation

V.A. Ramey
University of California, San Diego, CA, United States
NBER, Cambridge, MA, United States

Contents

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.003

Abstract

This chapter reviews and synthesizes our current understanding of the shocks that drive economic fluctuations. The chapter begins with an illustration of the problem of identifying macroeconomic shocks, followed by an overview of the many recent innovations for identifying shocks. It then reviews in detail three main types of shocks: monetary, fiscal, and technology. After surveying the literature, each section presents new estimates that compare and synthesize key parts of the literature. The penultimate section briefly summarizes a few additional shocks. The final section analyzes the extent to which the leading shock candidates can explain fluctuations in output and hours. It concludes that we are much closer to understanding the shocks that drive economic fluctuations than we were 20 years ago.

Keywords

Macroeconomic shocks, Monetary policy, Fiscal policy, Technology shocks, News, Identification, SVARs, DSGE estimation

JEL Classification Codes

E3, E5, E6

1. INTRODUCTION

At the beginning of the 20th century, economists began to recognize the importance of impulses and propagation mechanisms for explaining business cycle fluctuations. A key question was how to explain regular fluctuations in a model with dampened oscillations. In 1927, the Russian statistician Slutsky published a paper titled “The Summation of Random Causes as a Source of Cyclic Processes.” In this paper, Slutsky demonstrated the surprising result that moving sums of random variables could produce time series that looked very much like the movements of economic time series—"sequences of rising and falling movements, like waves…with marks of certain approximate uniformities and regularities."[a] This insight, developed independently by British mathematician Yule in 1926 and extended by Frisch (1933) in his paper "Propagation Problems and Impulse Problems in Dynamic Economics," revolutionized the study of business cycles. Their insights shifted the focus of research from developing mechanisms to support a metronomic view of business cycles, in which each boom created conditions leading to the next bust, to a search for the sources of the random shocks. Since then, economists have offered numerous candidates for these "random causes," such as crop failures, wars, technological innovation, animal spirits, government actions, and commodity shocks.

Research from the 1940s through the 1970s emphasized fiscal and monetary policy shocks, identified from large-scale econometric models or single equation analyses. The 1980s witnessed two important innovations that fundamentally changed the direction of the research. First, Sims' (1980a) paper "Macroeconomics and Reality" revolutionized the study of systems driven by random impulses by introducing vector autoregressions (VARs). Sims' VARs made the link between innovations to a linear system and macroeconomic shocks. Using his method, it became easier to talk about identification assumptions, to estimate impulse response functions, and to do innovation accounting using forecast error decompositions. The second important innovation was the expansion of the inquiry beyond policy shocks to consider important nonpolicy shocks, such as technology shocks (Kydland and Prescott, 1982).

These innovations led to a flurry of research on shocks and their effects. In his 1994 paper "Shocks," John Cochrane took stock of the state of knowledge at that time by using the by-then standard VAR techniques to conduct a fairly comprehensive search for the shocks that drove economic fluctuations. Surprisingly, he found that none of the popular candidates could account for the bulk of economic fluctuations. He proffered the rather pessimistic possibility that "we will forever remain ignorant of the fundamental causes of economic fluctuations" (Cochrane, 1994, abstract).

Are we destined to remain forever ignorant of the fundamental causes of economic fluctuations? Are Slutsky's "random causes" unknowable? In this chapter, I will summarize the new methodological innovations and what their application has revealed about the propagation of the leading candidates for macroeconomic shocks and their importance in explaining economic fluctuations since Cochrane's speculation.

The chapter progresses as follows. Section 2 begins by defining what a macroeconomic shock is. It then summarizes the many tools used for identifying macroeconomic shocks and computing impulse responses. It also highlights some of the complications and pitfalls, such as the effects of foresight and nonlinearities.

[a] Page 105 of the 1937 English version of the article published in *Econometrica*.

The topic of Section 3 is monetary shocks and their effects on the macroeconomy. The section summarizes the existing literature and the challenges to identification. It then explores the effects of several leading monetary shocks in a framework that incorporates some of the newer innovations.

Section 4 discusses fiscal shocks. It begins by summarizing results on government spending shocks and highlights the importance of anticipations. It estimates the effects of several leading identified shocks in a common framework. The second part of the section looks at tax shocks. It summarizes the literature on both unanticipated tax shocks and news about future tax changes and conducts some robustness checks.

Section 5 summarizes the literature on technology shocks, including total factor productivity (TFP) shocks, investment-specific technology (IST) shocks, and marginal efficiency of investment (MEI) shocks. It also discusses news about future technology. It compares a wide variety of identified shocks from the literature.

Section 6 briefly discusses four other candidate shocks: oil shocks, credit shocks, uncertainty shocks, and labor supply (or "wage markup") shocks.

Section 7 concludes by synthesizing what we have learned about shocks. It conducts a combined forecast error variance decomposition for output and hours to determine how much of the fluctuations can be accounted for by some of the leading shocks discussed in the earlier sections. It concludes that we have made substantial progress in understanding the shocks that drive the macroeconomy.

2. METHODS FOR IDENTIFYING SHOCKS AND ESTIMATING IMPULSE RESPONSES

2.1 Overview: What Is a Shock?

What, exactly, are the macroeconomic shocks that we seek to estimate empirically? There is some ambiguity in the literature about the definition because of some researchers' use of the term *shock* when they mean *innovation* (ie, the residuals from a reduced form VAR model) or *instrument*. Sims (1980a) equated innovations with macroeconomic shocks, despite claiming to be atheoretical. Others have used the word *shock* when they mean *instrument* (eg, Cochrane, 2004). In this chapter, I view shocks, VAR innovations, and instruments to be distinct concepts, although *identification assumptions* may equate them in many cases. Shocks are most closely related to the structural disturbances in a simultaneous equation system. I adopt the concept of shocks used by researchers such as Blanchard and Watson (1986), Bernanke (1986), and Stock and Watson (forthcoming). According to Bernanke (1986), the shocks should be *primitive* exogenous forces that are uncorrelated with each other and they should be *economically meaningful* (pp. 52–55).

I view the shocks we seek to estimate as the empirical counterparts to the shocks we discuss in our theories, such as shocks to technology, monetary policy, and fiscal policy. Therefore, the shocks should have the following characteristics: (1) they should be exogenous with respect to the other current and lagged endogenous variables in the model; (2) they should be uncorrelated with other exogenous shocks; otherwise, we cannot identify the unique causal effects of one exogenous shock relative to another; and (3) they should represent either unanticipated movements in exogenous variables or *news* about future movements in exogenous variables. With regard to condition (2), one might counter with situations in which both fiscal and monetary policies respond to some event and argue that therefore the fiscal and monetary shocks would be correlated. I would respond that these are not primitive shocks, but rather the *endogenous* responses of policies to a primitive shock. A primitive shock may directly enter several of the equations in the system. For example, a geopolitical event might lead to a war that causes both fiscal and monetary policy to respond endogenously. The geopolitical event would be the primitive shock from the standpoint of our economic models (though it might be considered an endogenous response from the standpoint of a political science model).[b]

To match these theoretical shocks, we can link the innovations in a structural vector autoregression (SVAR) to these theoretical (structural) shocks, estimate them in a structural dynamic stochastic general equilibrium (DSGE) model, or measure them directly using rich data sources.

2.2 Illustrative Framework

In this section, I lay out a simple framework in order to discuss the problem of identification and to illustrate some of the leading identification methods. I begin with the problem of identifying shocks to fiscal policy in a simple model with no dynamics. I then generalize the model to a dynamic trivariate model.

Consider first a simple model of the link between fiscal variables and GDP in a static setting. Suppose the structural relationships are given by the following equations:

$$\begin{aligned} \tau_t &= b_{\tau g} g_t + b_{\tau y} y_t + \varepsilon_{\tau t} \\ g_t &= b_{g\tau} \tau_t + b_{gy} y_t + \varepsilon_{gt} \\ y_t &= b_{y\tau} \tau_t + b_{yg} g_t + \varepsilon_{yt} \end{aligned} \tag{1}$$

where τ is taxes, g is government spending, and y is GDP. The εs are the macroeconomic shocks we seek to identify. We assume that they are uncorrelated and that, in this simple example, each one affects only one equation. $\varepsilon_{\tau t}$ is the tax shock; it might represent legislation resulting from a change in political power. ε_{gt} might capture the sudden outbreak of war, which raises desired military spending. ε_{yt} might capture technological progress. The *b*s capture the usual interactions. For example, we would expect that government spending would raise output, while taxes would lower it, so $b_{yg} > 0$ and $b_{y\tau} < 0$. Because

[b] Of course, the war might be caused by something like rainfall, in which case the primitive shock would be the rainfall. This shock would enter even more equations, such as the equations for government spending, GDP, and productivity.

of automatic stabilizers, however, the fiscal variables might also respond to GDP, ie, $b_{gy} < 0$ and $b_{\tau y} > 0$. This means that a simple regression of GDP on government spending and taxes will not uncover b_{yg} and $b_{y\tau}$ because g_t and τ_t are correlated with the shock to GDP, ε_{yt}. For example, we might observe no correlation between GDP and government spending, but this correlation is consistent both with no structural relationship between GDP and government spending (ie, $b_{yg} = b_{gy} = 0$) and with b_{yg} and b_{gy} being large, but with opposite signs. Without further assumptions or data, we cannot identify either the parameters or the shocks.

Now let us move to a simple trivariate model with three endogenous variables, Y_1, Y_2, and Y_3 in which dynamics are potentially important.[c] In the monetary context, these variables could be industrial production, a price index, and the federal funds rate; in the fiscal context, they could be GDP, government purchases, and tax revenue; and in the technology shock context, they could be labor productivity, hours, and investment. Let $Y_t = [Y_{1t}, Y_{2t}, Y_{3t}]$ be the vector of endogenous variables. Suppose that the dynamic behavior of Y_t is described by the following structural model:

$$Y_t = B(L)Y_t + \Omega\varepsilon_t \tag{2}$$

where $B(L) = B_0 + \sum_{k=1}^{p} B_k L^k$ and $E\left[\varepsilon_t \varepsilon_s'\right] = D$ if $t = s$, and 0 otherwise, where D is a diagonal matrix. The εs are the primitive structural shocks. Since a primitive shock can in principle affect more than one variable, I initially allow Ω to have nonzero off-diagonal elements.

The elements of B_0 are the same as the bs from Eq. (1), with $b_{jj} = 0$. Thus, the easiest way to address the dynamics is to recast the problem in terms of the *innovations* from a reduced form VAR:

$$A(L)Y_t = \eta_t \tag{3}$$

where $A(L)$ is a polynomial in the lag operator and $A(L) = I - \sum_{k=1}^{p} A_k L^k$. $\eta_t = [\eta_{1t}, \eta_{2t}, \eta_{3t}]$ are the reduced form VAR innovations. We assume that $E[\eta_t] = 0$, $E\left[\eta_t \eta_t'\right] = \Sigma_\eta$ and that $E\left[\eta_t \eta_s'\right] = 0$ for $s \neq t$. We then can link the innovations η in the reduced form VAR equation (3) to the unobserved structural shocks, ε, in the structural equation (2) as follows:

$$\eta_t = B_0 \eta_t + \Omega\varepsilon_t \quad \text{or} \tag{4a}$$

$$\eta_t = H\varepsilon_t, \quad \text{where } H = [I - B_0]^{-1}\Omega \tag{4b}$$

I will now write out the system in Eq. (4a) explicitly in a way that incorporates a commonly used identification assumption and a normalization. These restrictions are (i) Ω is the identity matrix (meaning each shock enters only one equation); and (ii) the structural

[c] See chapter Stock and Watson (forthcoming) in this handbook for a more precise analysis of identification using SVARs.

shocks have unit effect (ie, the diagonal elements of H are unity).[d] The system can then be written as

$$\begin{aligned}\eta_{1t} &= b_{12}\eta_{2t} + b_{13}\eta_{3t} + \varepsilon_{1t}\\ \eta_{2t} &= b_{21}\eta_{1t} + b_{23}\eta_{3t} + \varepsilon_{2t}\\ \eta_{3t} &= b_{31}\eta_{1t} + b_{32}\eta_{2t} + \varepsilon_{3t}\end{aligned} \tag{5}$$

This equation is the dynamic equivalent of Eq. (1). The only difference is that instead of writing the structural relationships in terms of the variables such as GDP, government spending, and taxes themselves, we now write them in terms of the reduced form VAR innovations. The interpretations of the bs, however, are the same if the structural relationships depend on contemporaneous interactions.

As discussed at the start of this section, we cannot identify the coefficients or the shocks without more restrictions. We require at least three more restrictions for identification of all three shocks, potentially fewer if we want to identify only one shock. Since a number of the common identification methods depend on contemporaneous restrictions, I will refer to the system of equations in Eq. (5) when discussing them.

2.3 Common Identification Methods

In this section, I briefly overview some of the most common methods for identification. This section is not meant to be comprehensive. See Stock and Watson (forthcoming) for more detailed treatments of the methods I summarize, as well as for a few other methods I do not summarize, such as set identification and identification through heteroscedasticity. I use the term "policy variable" for short, but it should be understood that it can represent any variable from which we want to extract a shock component.

2.3.1 Cholesky Decompositions

The most commonly used identification method in macroeconomics imposes alternative sets of recursive zero restrictions on the contemporaneous coefficients. This method was introduced by Sims (1980a) and is also known as "triangularization." The following are two widely used alternatives:

A. The policy variable does not respond within the period to the other endogenous variables. This could be motivated by decision lags on the part policymakers or other adjustment costs. Let Y_1 be the policy variable and η_1 be its reduced form innovation. Then this scheme involves constraining $b_{12} = b_{13} = 0$ in Eq. (5), which is equivalent to ordering the policy variable first in the Cholesky ordering. For example, Blanchard and Perotti (2002) impose this constraint to identify the shock to government

[d] An alternative normalization to (ii) is the assumption that the structural shocks have unit standard deviation (ie, the variances of the εs are unity).

spending; they assume that government spending does not respond to the contemporaneous movements in output or taxes.[e]

B. The other endogenous variables do not respond to the policy shock within the period. This could be motivated by sluggish responses of the other endogenous variables to shocks to the policy variable. This scheme involves constraining $b_{21}=b_{31}=0$, which is equivalent to ordering the policy variable last in the Cholesky ordering. For example, Bernanke and Blinder (1992) were the first to identify shocks to the federal funds rate as monetary policy shocks and used this type of identification.[f]

Several of the subsequent sections will discuss how these timing assumptions are not as innocuous as they might seem at first glance. For example, forward-looking behavior or superior information on the part of policymakers may invalidate these restrictions.

2.3.2 *Other Contemporaneous Restrictions*

Another more general approach (that nests the Cholesky decomposition) is what is known as a *structural VAR*, or SVAR, introduced by Blanchard and Watson (1986) and Bernanke (1986). This approach uses either economic theory or outside estimates to constrain parameters. Consider, for example, Blanchard and Perotti's (2002) identification of government spending and net tax shocks. Let Y_1 be net taxes, Y_2 be government spending, and Y_3 be GDP. They identify the shock to government spending using a Cholesky decomposition in which government spending is ordered first (ie, $b_{21}=b_{23}=0$). They identify exogenous shocks to net taxes by setting $b_{13}=2.08$, an outside estimate of the cyclical sensitivity of net taxes.[g] These three restrictions are sufficient to identify all of the remaining parameters and hence all three shocks.

2.3.3 *Narrative Methods*

Narrative methods involve constructing a series from historical documents to identify the reason and/or the quantities associated with a particular change in a variable. Friedman and Schwartz (1963) is the classic example of using historical information to identify policy shocks. Hamilton (1985) and Hoover and Perez (1994) used narrative methods to identify oil shocks. These papers isolated political events that led to disruptions in world

[e] To implement this identification using ordinary least squares (OLSs), one would simply regress government spending on p lags of all of the variables in the system and call the residual the government spending shock.

[f] To implement this identification using OLSs, one would regress the federal funds rate on contemporaneous values of the other variables in the system, as well as p lags of all of the variables, and call the residual the monetary policy shock.

[g] One way to implement the tax shock identification is to construct the variable $\eta_1-2.08\eta_3$ from the estimated reduced form residuals. One would then regress η_3 on η_1 and η_2, using $\eta_1-2.08\eta_3$ as the instrument for η_1. (Note that the assumption that $b_{21}=b_{23}=0$ identifies η_2 as ε_{2t}, which is uncorrelated with ε_{3t} by assumption) This regression identifies b_{31} and b_{32}. The residual is the estimate of ε_{3t}.

oil markets. Other examples of the use of narrative methods are Poterba's (1986) tax policy announcements, Romer and Romer's (1989, 2004) monetary shock series based on FOMC minutes, Ramey and Shapiro (1998) and Ramey's (2011a) defense news series based on *Business Week* articles, and Romer and Romer's (2010) narrative series of tax changes based on reading legislative documents.

Until recently, these series were used either as exogenous shocks in sets of dynamic single equation regressions or embedded in a Cholesky decomposition. For example, in the framework above, we could set Y_1 to be the narrative series and constrain $b_{12} = b_{13} = 0$. As a later section details, recent innovations have led to additional methods for incorporating these series.

A cautionary note on the potential of narrative series to identify exogenous shocks is in order. Some of the follow-up research has operated on the principle that the narrative alone provides exogeneity. It does not. Shapiro (1994) and Leeper (1997) made this point for monetary policy shocks. Another example is in the fiscal literature. A series on fiscal consolidations, quantified by narrative evidence on the expected size of these consolidations, is not necessarily exogenous. If the series includes fiscal consolidations adopted in response to bad news about the future growth of the economy, the series cannot be used to establish a causal effect of the fiscal consolidation on future output.

2.3.4 High-Frequency Identification

Research by Bagliano and Favero (1999), Kuttner (2001), Cochrane and Piazzesi (2002), Faust et al. (2004), Gürkaynak et al. (2005), Piazzesi and Swanson (2008), Gertler and Karadi (2015), Nakamura and Steinsson (2015), and others has used high-frequency data (such as news announcements around FOMC dates) and the movement of federal funds futures to identify unexpected Fed policy actions. This identification is also based in part on timing, but because the timing is so high frequency (daily or higher), the assumptions are more plausible than those employed at the monthly or quarterly frequency. As I will discuss in the foresight section later, the financial futures data are ideal for ensuring that a shock is unanticipated.

It should be noted, however, that without additional assumptions the unanticipated shock is not necessarily exogenous to the economy. For example, if the implementation does not adequately control for the Fed's private information about the future state of the economy, which might be driving its policy changes, these shocks cannot be used to estimate a causal effect of monetary policy on macroeconomic variables.

2.3.5 External Instruments/Proxy SVARs

The "external instrument," or "proxy SVAR," method is a promising new approach for incorporating external series for identification. This method was developed by Stock and Watson (2008) and extended by Stock and Watson (2012) and Mertens and Ravn (2013). This approach takes advantage of information developed from "outside" the VAR, such as series based on narrative evidence, shocks from estimated DSGE models, or high-frequency information. The idea is that these external series are noisy measures of the true shock.

Suppose that Z_t represents one of these external series. Then this series is a valid instrument for identifying the shock ε_{1t} if the following two conditions hold:

$$E[Z_t \varepsilon_{1t}] \neq 0 \tag{6a}$$

$$E[Z_t \varepsilon_{it}] = 0, \; i = 2, 3 \tag{6b}$$

Condition (6a) is the instrument *relevance* condition: the external instrument must be contemporaneously correlated with the structural policy shock. Condition (6b) is the instrument *exogeneity* condition: the external instrument must be contemporaneously uncorrelated with the other structural shocks. If the external instrument satisfies these two conditions, it can be used to identify the shock ε_{1t}.

The procedure is very straightforward and takes place with the following steps[h]:

Step 1: Estimate the reduced form system to obtain estimates of the reduced form residuals, η_t.

Step 2: Regress η_{2t} and η_{3t} on η_{1t} using the external instrument Z_t as the instrument. These regressions yield unbiased estimates of b_{21} and b_{31}. Define the residuals of these regressions to be ν_{2t} and ν_{3t}.

Step 3: Regress η_{1t} on η_{2t} and η_{3t}, using the ν_{2t} and ν_{3t} estimated in Step 2 as the instruments. This yields unbiased estimates of b_{12} and b_{13}.

As an example, Mertens and Ravn (2014) reconcile Romer and Romer's (2010) estimates of the effects of tax shocks with the Blanchard and Perotti (2002) estimates by using the Romer's narrative tax shock series as an external instrument Z to identify the structural tax shock. Thus, they do not need to impose parameter restrictions, such as the cyclical elasticity of taxes to output. As I will discuss in Section 2.4, one can extend this external instrument approach to estimating impulse responses by combining it with Jordà's (2005) method.

2.3.6 Restrictions at Longer Horizons

Rather than constraining the contemporaneous responses, one can instead identify a shock by imposing long-run restrictions. The most common is an infinite horizon long-run restriction, first used by Shapiro and Watson (1988), Blanchard and Quah (1989), and King et al. (1991). Consider the moving average representation of Eq. (3):

$$Y_t = C(L)\eta_t \tag{7}$$

[h] This exposition follows Merten and Ravn (2013, online appendix). See Mertens and Ravn (2013a,b) and the associated online appendices for generalizations to additional external instruments and to larger systems.

where $C(L) = [A(L)]^{-1}$. Combining Eq. (4b) with Eq. (7), we can write the Ys in terms of the structural shocks:

$$Y_t = D(L)\varepsilon_t \tag{8}$$

where $D(L) = C(L)H$. Suppose we wanted to identify a technology shock as the only shock that affects labor productivity in the long run. In this case, Y_1 would be the *growth rate* of labor productivity and the other variables would also be transformed to induce stationary (eg, first-differenced). Letting $D^{ij}(L)$ denote the (i,j) element of the D matrix and $D^{11}(1)$ denote the lag polynomial with $L=1$, we impose the long-run restriction by setting $D^{12}(1) = 0$ and $D^{13}(1) = 0$. This restriction constrains the unit root in Y_1 to emanate only from the shock that we are calling the technology shock. This is the identification used by Galí (1999).

An equivalent way of imposing this restriction is to use the estimation method suggested by Shapiro and Watson (1988). Let Y_1 denote the first difference of the log of labor productivity and Y_2 and Y_3 be the stationary transformations of two other variables (such as hours). Then, imposing the long-run restriction is equivalent to identifying the error term in the following equation as the technology shock:

$$Y_{1t} = \sum_{j=1}^{p} \omega_{11,j} Y_{1t-j} + \sum_{j=1}^{p-1} \omega_{12,j} \Delta Y_{2t-j} + \sum_{j=1}^{p-1} \omega_{13,j} \Delta Y_{3t-j} + \zeta_t \tag{9}$$

We have imposed the restriction by specifying that only the *first differences* of the other stationary variables enter this equation. Because the current values of those differences might also be affected by the technology shock, and therefore correlated with the error term, we use lags 1 through p of Y_2 and Y_3 as instruments for the terms involving the current and lagged values of those variables. The estimated residual is the identified technology shock. We can then identify the other shocks, if desired, by orthogonalizing the error terms with respect to the technology shock.

This equivalent way of imposing long-run identification restrictions highlights some of the problems that can arise with this method. First, identification depends on the relevance of the instruments. Second, it requires additional identifying restrictions in the form of assumptions about unit roots. If, for example, hours have a unit root, then in order to identify the technology shock one would have to impose that only the second difference of hours entered in Eq. (9).[i]

Another issue is the behavior of infinite horizon restrictions in small samples (eg, Faust and Leeper, 1997). Recently, researchers have introduced new methods that overcome these problems. Building on earlier work by Faust (1998) and Uhlig (2003, 2004), Francis

[i] To be clear, all of the Y variables must be trend stationary in this system. If hours have a unit root, then Y_2 must be equal to Δhours_t, so the constraint in Eq. (9) would take the form $\Delta^2\text{hours}_t$.

et al. (2014) identify the technology shock as the shock that maximizes the forecast error variance share of labor productivity at some finite horizon h. A variation by Barsky and Sims (2011) identifies the shock as the one that maximizes the *sum* of the forecast error variances up to some horizon h. See those papers for details on how to implement these methods.

2.3.7 Sign Restrictions

A number of authors had noted the circularity in some of the reasoning analyzing VAR specifications in practice. In particular, whether a specification or identification method is deemed "correct" is often judged by whether the impulses they produce are "reasonable," ie, consistent with the researcher's priors. Faust (1998) and Uhlig (1997, 2005) developed a new method to incorporate "reasonableness" without undercutting scientific inquiry by investigating the effects of a shock on variable Y, where the shock was identified by sign restrictions on the responses of *other* variables (excluding variable Y). Work by Canova and De Nicolo (2002) and Canova and Pina (2005) introduced other variations.

The sign restriction method has been used in many contexts, such as monetary policy, fiscal policy, and technology shocks. Recently, there have been a number of new papers on sign restrictions using Bayesian methods. For example, Arias et al. (2015b) propose methods involving agnostic priors in one dimension and Baumeister and Hamilton (2015) propose methods involving agnostic priors in another dimension. Amir Ahmadi and Uhlig (2015) combine sign restrictions with Bayesian factor-augmented VARs (FAVARs). See Stock and Watson (forthcoming) for more discussion of sign restrictions as an identification method.

2.3.8 Factor-Augmented VARs

A perennial concern in identifying shocks is that the variables included in the VAR do not capture all of the relevant information. The comparison of price responses in monetary VARs with and without commodity prices is one example of the difference a variable exclusion can make. To address this issue more broadly, Bernanke et al. (2005) developed the FAVARs based on earlier dynamic factor models developed by Stock and Watson (2002) and others. The FAVAR, which typically contains over 100 series, has the benefit that it is much more likely to condition on relevant information for identifying shocks. In most implementations, though, it still typically relies on a Cholesky decomposition. Amir Ahmadi and Uhlig's (2015) new method using sign restrictions in Bayesian FAVARs is one of the few examples that does not rely on Cholesky decompositions. One shortcoming of FAVAR methods is that all variables must be transformed to a stationary form, which requires pretesting and its concomitant problems (eg, Elliott, 1998; Gospodinov et al., 2013). See Stock and Watson (forthcoming) for an in depth discussion of dynamic factor models.

2.3.9 Estimated DSGE Models

An entirely different approach to identification is the estimated DSGE model, introduced by Smets and Wouters (2003, 2007). This method involves estimating a fully specified model (a new Keynesian model with many frictions and rigidities in the case of Smets and Wouters) and extracting a full set of implied shocks from those estimates. In the case of Smets and Wouters, many shocks are estimated including technology shocks, monetary shocks, government spending shocks, wage markup shocks, and risk premium shocks. One can then trace out the impulse responses to these shocks as well as do innovation accounting. Other examples of this method appear in work by Justiniano et al. (2010, 2011) and Schmitt-Grohe and Uribe (2012). Christiano et al. (2005) take a different estimation approach by first estimating impulse responses to a monetary shock in a standard SVAR and then estimating the parameters of the DSGE model by matching the impulse responses from the model to those of the data.

These models achieve identification by imposing structure based on theory. It should be noted that identification is less straightforward in these types of models. Work by Canova and Sala (2009), Komunjer and Ng (2011), and others highlights some of the potential problems with identification in DSGE models. On the other hand, this method overcomes some of the potential problems of unrestricted VARs highlighted by Fernandez-Villaverde et al. (2007).

2.4 Estimating Impulse Responses

Suppose that one has identified the economic shock through one of the methods discussed earlier. How do we measure the effects on the endogenous variables of interest? The most common way to estimate the impulse responses to a shock uses nonlinear (at horizons greater than one) functions of the estimated VAR parameters. In particular, estimation of the reduced form system provides the elements of the moving average representation matrix $C(L)=[A(L)]^{-1}$ in Eq. (7) and identification provides the elements of B_0. Recalling that $D(L)=C(L)H$, we write out $D(L)=D_0+D_1L+D_2L^2+D_3L^3+\cdots$, and denoting $D_h=[d_{ijh}]$, we can express the impulse response of variable Y_i at horizon $t+h$ to a shock to ε_{jt} as:

$$\frac{\partial Y_{i,t+h}}{\partial \varepsilon_{j,t}}=d_{ijh} \tag{10}$$

These d_{ijh} parameters are nonlinear functions of the reduced form VAR parameters.

If the VAR adequately captures the data-generating process, this method is optimal at all horizons. If the VAR is mispecified, however, then the specification errors will be compounded at each horizon. To address this problem, Jordà (2005) introduced a *local projection* method for estimating impulse responses. The comparison between his procedure and the standard procedure has an analogy with direct forecasting vs iterated forecasting (eg, Marcellino et al., 2006). In the forecasting context, one can forecast future values of a variable using either a horizon-specific regression ("direct" forecasting) or iterating on a one-period ahead estimated model ("iterated" forecasting). Jordà's method is analogous to the direct forecasting, whereas the standard VAR method is analogous to the iterated forecasting method. Chang and Sakata (2007) introduce a related method they call *long autoregression* and show its asymptotic equivalence to Jordà's method.

To see how Jordà's method works, suppose that ε_{1t} has been identified by one of the methods discussed in the previous section. Then, the impulse response of Y_i at horizon h can be estimated from the following single regression:

$$Y_{i,t+h} = \theta_{i,h} \cdot \varepsilon_{1t} + \text{control variables} + \xi_{t+h} \tag{11}$$

$\theta_{i,h}$ is the estimate of the impulse response of Y_i at horizon h to a shock ε_{1t}. The control variables need not include the other Ys as long as ε_{1t} is exogenous to those other Ys. Typically, the control variables include deterministic terms (constant, time trends), lags of the Y_i, and lags of other variables that are necessary to "mop up"; the specification can be chosen using information criteria. One estimates a separate regression for each horizon and the control variables do not necessarily need to be the same for each regression. Note that except for horizon $h=0$, the error term ξ_{t+h} will be serially correlated because it will be a moving average of the forecast errors from t to $t+h$. Thus, the standard errors need to incorporate corrections for serial correlation, such as a Newey–West (1987) correction.

Because the Jordà method for calculating impulse response functions imposes fewer restrictions, the estimates are often less precisely estimated and are sometimes erratic. Nevertheless, this procedure is more robust than standard methods, so it can be very useful as a heuristic check on the standard methods. Moreover, it is much easier to incorporate state dependence with this method (eg, Auerbach and Gorodnichenko, 2013).

One can extend the Jordà method in several ways that incorporates some of the new methodology. First, one can incorporate the advantages of the FAVAR method (see Section 2.3.8) by including estimated factors as control variables. Second, one can merge the insights from the external instrument/proxy SVAR method (see Section 2.3.5). To see this, modify Eq. (11) as follows:

$$Y_{i,t+h} = \theta_{i,h} \cdot Y_{1,t} + \text{control variables} + \zeta_{t+h} \tag{12}$$

where we have replaced the shock ε_{1t} with $Y_{1,t}$. As discussed earlier, an OLS regression of Y_i on Y_1 cannot capture the structural effect if Y_1 is correlated with ζ_{t+h}. We can easily deal with this issue, however, by estimating this equation using the external instrument Z_t as an instrument for $Y_{1,t}$. For example, if Y_i is real output and $Y_{1,t}$ is the federal funds rate, we can use Romer and Romer's (2004) narrative-based monetary shock series as an instrument. As I will discuss later, in some cases there are multiple potential external instruments. We can readily incorporate these in this framework by using multiple instruments for Y_1. In fact, these overidentifying restrictions can be used to test the restrictions of the model (using a Hansen's *J*-statistic, for example).

2.5 The Problem of Foresight

The problem of foresight presents serious challenges to, but also opportunities for, the identification of macroeconomic shocks.[j] There are two main foresight problems: (i) foresight on the part of private agents; and (ii) foresight on the part of policymakers. I will discuss each in turn.

It is likely that many changes in policy or other exogenous shocks are anticipated by private agents in advance. For example, Beaudry and Portier (2006) explicitly take into account that news about future technology may have effects today even though it does not show up in current productivity. Ramey (2011a) argues that the results of Ramey and Shapiro (1998) and Blanchard and Perotti (2002) differ because most of the latter's identified shocks to government spending are actually anticipated. Building on work by Hansen and Sargent (1991), Leeper et al. (2013) work out the econometrics of "fiscal foresight" for taxes, showing that foresight can lead to a nonfundamental moving average representation. The growing importance of "forward guidance" in monetary policy means that many changes in policy rates may be anticipated.

Consider the following example, based on Leeper et al. (2013), of a simple growth model with a representative household with log utility over consumption, discount factor β, and a production function $Y_t = A_t K_{t-1}^{\alpha}$, with $\alpha < 1$. The government taxes output Y at a rate τ_t and there are i.i.d. shocks, $\hat{\tau}_t$, to the tax rate relative to its mean τ. Shocks to technology, ε_{At}, are also i.i.d. Suppose that agents potentially receive news in period t of what the tax rate will be in $t+q$, so that $\hat{\tau}_t = \varepsilon_{\tau,t-q}$. If the shocks are unanticipated ($q=0$), the rule for capital accumulation is:

$$k_t = \alpha k_{t-1} + \varepsilon_{A,t}$$

which reproduces the well-known result that unanticipated i.i.d. tax rate shocks have no effect on capital accumulation. If the tax rate shock is anticipated two periods in advance ($q=2$), however, then optimal capital accumulation is:

$$k_t = \alpha k_{t-1} + \varepsilon_{A,t} - k\{\varepsilon_{\tau,t-1} + \theta \varepsilon_{\tau,t}\}$$

where $\theta = \alpha\beta(1-\tau) < 1$ and $k = (1-\theta)\dfrac{\tau}{1-\tau}$. Can we uncover the tax shocks by regressing capital on its own lags? No, we cannot. Because $\theta < 1$, this representation is not invertible in the current and past ks; we say that $\{\varepsilon_{\tau,t-j}\}_{j=0}^{\infty}$ is not *fundamental* for

[j] The general problem was first recognized and discussed decades ago. For example, Sims (1980a) states: "It is my view, however, that rational expectations is more deeply subversive of identification than has yet been recognized."

$\left\{k_{t-j}\right\}_{j=0}^{\infty}$. If we regress k_t on its own lags and recover the innovations, we would be recovering the discounted sum of tax news observed at date t and earlier, ie, "old" news. Adding lagged taxes to the VAR does not help.

Beaudry et al. (2015) develop a diagnostic to determine whether nonfundamentalness is quantitatively important. They argue that in some cases the nonfundamental representation is close to the fundamental representation.

The second foresight problem is foresight on the part of policymakers. Sometimes policymakers have more information about the state of the economy than private agents. If this is the case, and we do not include that information in the VAR, part of the identified shock may include the endogenous response of policy to expectations about the future path of macroeconomic variables. Consider the "price puzzle" in monetary VARs, meaning that some identified monetary policy shocks imply that a monetary contraction raises prices in the short run. Sims (1992) argued that the "price puzzle" was the result of typical VARs not including all relevant information for forecasting future inflation. Thus, the identified policy shocks included not only the exogenous shocks to policy but also the endogenous policy responses to forecasts of future inflation. In the fiscal context, governments may undertake fiscal consolidations based on private information about declining future growth of potential GDP. If this is not taken into account, then a finding that a fiscal consolidation lowers output growth may be confounding causal effects with foresight effects.

The principal methods for dealing with the problem of foresight are measuring the expectations directly, time series restrictions, or theoretical model restrictions. For example, Beaudry and Portier (2006) extracted news about future technology from stock prices; Ramey (2011a) created a series of news about future government spending by reading *Business Week* and other periodicals; Fisher and Peters (2010) created news about government spending by extracting information from stock returns of defense contractors; Poterba (1986) and Leeper et al. (2012) used information from the spread between federal and municipal bond yields for news about future tax changes; and Mertens and Ravn (2012) decomposed Romer and Romer's (2010) narrative tax series into one series in which implementation was within the quarter (unanticipated) and another series in which implementation was delayed (news). In the monetary shock literature, many papers use high-frequency financial futures prices to decompose the anticipated and unanticipated components of interest rates changes (eg, Rudebusch, 1998; Bagliano and Favero, 1999; Kuttner, 2001; Gürkaynak et al., 2005).

The typical way that news has been incorporated into VARs is by adding the news series to a standard VAR, and ordering it first. Perotti (2011) has called these "EVARs" for "Expectational VARs." Note that in general one cannot use news as an external instrument in Mertens and Ravn's proxy SVAR framework. The presence of foresight invalidates the interpretation of the VAR reduced form residuals as prediction errors, since the conditioning variables may not span the information set of forward-looking agents (Mertens and Ravn, 2013, 2014).

2.6 The Problem of Trends

Most macroeconomic variables are nonstationary, exhibiting behavior consistent with either deterministic trends or stochastic trends. A key question is how to specify a model when many of the variables may be trending. Sims et al. (1990) demonstrate that even when variables might have stochastic trends and might be cointegrated, the log levels specification will give consistent estimates. While one might be tempted to pretest the variables and impose the unit root and cointegration relationships to gain efficiency, Elliott (1998) shows that such a procedure can lead to large size distortions in theory. More recently, Gospodinov et al. (2013) have demonstrated how large the size distortions can be in practice.

Perhaps the safest method is to estimate the SVAR in log levels (perhaps also including some deterministic trends) as long as the imposition of stationarity is not required for identification. One can then explore whether the imposition of unit roots and cointegration lead to similar results but increase the precision of the estimates. For years, it was common to include a linear time trend in macroeconomic equations. Many analyses now include a broken trend or a quadratic trend to capture features such as the productivity slowdown in 1974 or the effect of the baby boom moving through the macroeconomic variables (eg, Perron, 1989; Francis and Ramey, 2009).

2.7 Some Brief Notes on Nonlinearities

In the previous sections, we have implicitly assumed that the relationships we are trying to capture can be well approximated with linear functions. There are many cases in which we believe that nonlinearities might be important. To name just a few possible nonlinearities, positive shocks might have different effects from negative shocks, effects might not be proportional to the size of the shock, or the effect of a shock might depend on the state of the economy when the shock hits.

A thorough analysis of nonlinearities is beyond the scope of this chapter, so I will mention only three items briefly. First, Koop et al. (1996) provide a very useful analysis of the issues that arise when estimating impulse responses in nonlinear models. Second, if one is interested in estimating state-dependent models, the Jordà (2005) local projection method is a simple way to estimate such a model and calculate impulse response functions. Auerbach and Gorodnichenko (2013) and Ramey and Zubairy (2014) discuss this application and how it relates to another leading method, smooth transition VARs.

The third point is a cautionary note when considering the possibility of asymmetries. Many times researchers posit that only positive, or only negative, shocks matter. For example, in the oil shock literature, it is common to assume that only oil price *increases* matter and to include a variable in the VAR that captures increases but not decreases.

Kilian and Vigfusson (2011) demonstrate the serious biases and faulty inference that can result from this specification. Their explanation is simple. Suppose Y is a linear function of X, where X takes on both negative and positive values. If one imposes the restriction that only positive values matter, one is in essence setting all of the negative values of X to zero. Figure 1 of Kilian and Vigfusson's (2011) paper demonstrates how this procedure that truncates on the X variable produces slope coefficients that are biased upward in magnitude. Thus, one would incorrectly conclude that positive Xs have a greater impact than negative Xs, even when the true relationship is linear. To guard against this faulty inference, one should always make sure that the model nests the linear case when one is testing for asymmetries. If one finds evidence of asymmetries, then one can use Kilian and Vigfusson's (2011) methods for computing the impulse responses correctly.

2.8 DSGE Monte Carlos

Much empirical macroeconomics is linked to testing theoretical models. A question that arises is whether shocks identified in SVARs, often with minimal theoretical restrictions, are capable of capturing the true shocks. Fernandez-Villaverde et al. (2007) study this question by comparing the state-space representation of a theoretical model with the VAR representation. They note that in some instances an invertibility problem can arise and they offer a method to check whether the problem is present.

Erceg et al. (2005) were perhaps the first to subject an SVAR involving long-run restrictions to what I will term a "DSGE Monte Carlo." In particular, they generated artificial data from a calibrated DSGE model and applied SVARS with long restrictions to the data to see if the implied impulse responses matched those of the underlying model.

This method has now been used in several settings. Chari et al. (2008) used this method to argue against SVARs' ability to test the real business cycle (RBC) model, Ramey (2009) used it to show how standard SVARs could be affected by anticipated government spending changes, and Francis et al. (2014) used this method to verify the applicability of their new finite horizon restrictions method. This method seems to be a very useful tool for judging the ability of SVARs to test DSGE models. Of course, like any Monte Carlo, the specification of the model generating the artificial data is all important.

3. MONETARY POLICY SHOCKS

Having discussed the definition of macroeconomic shocks and the leading methods for identifying them, I now turn to the first of the candidate shocks that will be discussed in detail: monetary policy shocks. In this section, I review the main issues and results from the empirical literature seeking to identify and estimate the effects of monetary policy shocks. I begin with a brief overview of the research before and after Christiano et al.'s (1999) *Handbook of Macroeconomics* chapter on the subject. I revisit Christiano, Eichenbaum and Evan's specification and then focus on two leading types of externally identified monetary policy shocks, Romer and Romer's (2004) narrative/Greenbook shock and Gertler and Karadi's (2015) recent high-frequency identification (HFI) shocks identified using federal funds futures. I focus on these two types of shocks in part because they both imply very similar effects of monetary policy on output, despite using different identification methods and different samples.

Before beginning, it is important to clarify that the "shocks" identified in the monetary shock literature are not always the empirical counterparts to the shocks from our theoretical models, as discussed in Section 2.1. Because monetary policy is typically guided by a rule, most movements in monetary policy instruments are due to the *systematic* component of monetary policy rather than to deviations from that rule.[k] We do not have many good economic theories for what a structural monetary policy shock should be. Other than "random coin flipping," the most frequently discussed source of monetary policy shocks is shifts in central bank preferences, caused by changing weights on inflation vs unemployment in the loss function or by a change in the political power of individuals on the FOMC. A few papers explicitly link the empirically identified shocks to shifts in estimated central bank preferences (eg, Owyang and Ramey, 2004; Lakdawala, 2015), but most treat them as innovations to a Taylor rule, with no discussion of their economic meaning.[l]

If many macroeconomists now believe that monetary policy shocks themselves contribute little to macroeconomic outcomes, why is there such a large literature trying to identify them? The reason is that we want to identify nonsystematic movements in monetary policy so that we can estimate *causal* effects of money on macroeconomic variables. As Sims (1998) argued in his response to Rudebusch's (1998) critique of standard VAR methods, we need instruments in order to identify key structural parameters. Analogous to the supply and demand framework where we need demand shift instruments to identify the parameters of the supply curve, in the monetary policy context we require deviations from the monetary rule to identify the response of the economy to monetary policy. Thus, much of the search for "shocks" to monetary policy is a search for instruments rather than for primitive macroeconomic shocks.

3.1 A Brief History Through 1999

The effect of monetary policy on the economy is one of the most studied empirical questions in all of macroeconomics. The most important early evidence was Friedman and Schwartz's path-breaking 1963 contribution in the form of historical case studies

[k] Friedman argued, however, that most fluctuations in monetary instruments before 1960 were due to nonsystematic components of monetary policy.

[l] Christiano et al. (1999) discuss a few other possibilities, such as measurement error in preliminary data (pp. 71–73).

and analysis of historical data. The rational expectations revolution of the late 1960s and 1970s highlighted the importance of distinguishing the part of policy that was part of a rule vs shocks to that rule, as well as anticipated vs unanticipated parts of the change in the policy variable. Sims (1972, 1980a,b) developed modern time series methods that allowed for that distinction while investigating the effects of monetary policy. During the 1970s and much of the 1980s, shocks to monetary policy were measured as shocks to the stock of money (eg, Sims, 1972; Barro, 1977, 1978). This early work offered evidence that (i) money was (Granger-) causal for income; and (ii) that fluctuations in the stock of money could explain an important fraction of output fluctuations. Later, however, Sims (1980b) and Litterman and Weiss (1985) discovered that the inclusion of interest rates in the VAR significantly reduced the importance of shocks to the money stock for explaining output, and many concluded that monetary policy was not important for understanding economic fluctuations.[m]

There were two important rebuttals to the notion that monetary policy was not important for understanding fluctuations. The first rebuttal was by Romer and Romer (1989), who developed a narrative series on monetary policy shocks in the spirit of Friedman and Schwarz's (1963) work. Combing through FOMC minutes, they identified dates at which the Federal Reserve "attempted to exert a contractionary influence on the economy in order to reduce inflation" (p. 134). They found that industrial production decreased significantly after one of these "Romer Dates." The Romer and Romer series rapidly gained acceptance as an indicator of monetary policy shocks.[n] A few years later, though, Shapiro (1994) and Leeper (1997) showed that Romer and Romer's dummy variable was, in fact, predictable from lagged values of output (or unemployment) and inflation. Both argued that the narrative method used by Romer and Romer did not adequately separate *exogenous* shocks to monetary policy, necessary for establishing the strength of the causal channel, from the *endogenous* response of monetary policy to the economy.[o]

The second rebuttal to the Sims and Litterman and Weiss argument was by Bernanke and Blinder (1992). Building on an earlier idea by McCallum (1983), Bernanke and Blinder turned the money supply vs interest rate evidence on its head by arguing that interest rates, and in particular the federal funds rate, were *the* key indicators of monetary policy. They showed that both in Granger-causality tests and in variance decompositions of forecast errors, the federal funds rate outperformed both M1 and M2, as well as the 3-month Treasury bill and the 10-month Treasury bond for most variables.

[m] Of course, this view was significantly strengthened by Kydland and Prescott's (1982) seminal demonstration that business cycles could be explained with technology shocks.

[n] Boschen and Mills (1995) also extended the Romer and Romer's dummy variables to a more continuous indicator.

[o] See, however, Romer and Romer's (1997) response to Leeper.

The 1990s saw numerous papers that devoted attention to the issue of the correct specification of the monetary policy function. These papers used prior information on the monetary authority's operating procedures to specify the policy function in order to identify correctly the shocks to policy. For example, Christiano and Eichenbaum (1992) used nonborrowed reserves, Strongin (1995) suggested the part of nonborrowed reserves orthogonal to total reserves, and Bernanke and Mihov (1998) generalized these ideas by allowing for regime shifts in the type of monetary instrument that is targeted.[p] Another issue that arose during this period was the "Price Puzzle," a term coined by Eichenbaum (1992) to describe the common result that a contractionary shock to monetary policy appeared to raise the price level in the short-run. Sims (1992) conjectured that the Federal Reserve used more information about future movements in inflation than was commonly included in the VAR. He showed that the price puzzle was substantially reduced if commodity prices, often a harbinger of future inflation, were included in the VAR.

Christiano et al.'s (1999) *Handbook of Macroeconomics* chapter "Monetary policy shocks: What have we learned and to what end?" summarized and explored the implications of many of the 1990s innovations in studying monetary policy shocks. Their benchmark model used a particular form of the Cholesky decomposition in which the first block of variables consisting of output, prices, and commodity prices was assumed not to respond to monetary policy shocks within the quarter (or month). They called this identification assumption the "recursiveness assumption." On the other hand, they allowed contemporaneous values of the first-block variables to affect monetary policy decisions. Perhaps the most important message of the chapter was the robustness of the finding that a contractionary monetary policy shock, whether measured with the federal funds rate or nonborrowed reserves, had significant negative effects on output. On the other hand, the price puzzle continued to pop up in some specifications.

3.2 Some Alternatives to the Standard Model

Not all research on monetary policy shocks has been conducted in the canonical time-invariant linear SVAR model. In this section, I briefly highlight some of the research that generalizes the linear models or uses completely different methods.

3.2.1 Regime-Switching Models

In addition to the switch between interest rate targeting and nonborrowed reserve targeting (discussed by Bernanke and Mihov, 1998), several papers have estimated regime-switching models of monetary policy. The idea in these models is that monetary policy is driven not just by shocks but also by changes in the policy parameters. In an early

[p] An important part of this literature was addressed to the "liquidity puzzle," that is, the failure of some measures of money supply shocks to produce a negative short-run correlation between the supply of money and interest rates.

contribution to this literature, Owyang and Ramey (2004) estimated a regime-switching model in which the Fed's preference parameters could switch between "hawk" and "dove" regimes. They found that the onset of a dove regime leads to a steady increase in prices, followed by decline in output after approximately a year. Primiceri (2005) investigated the roles of changes in systematic monetary policy vs shocks to policy in the outcomes in the last 40 years. While he found evidence for changes in systematic monetary policy, he concluded that they are not an important part of the explanation of fluctuations in inflation and output. Sims and Zha (2006a) also considered regime-switching models and found evidence of regime switches that correspond closely to changes in the Fed chairmanship. Nevertheless, they also concluded that changes in monetary policy regimes do not explain much of economic fluctuations.

3.2.2 Time-Varying Effects of Monetary Policy

In their summary of the monetary policy literature in their chapter in the *Handbook of Monetary Economics*, Boivin et al. (2010) focus on time variation in the estimated effects of monetary policy. I refer the reader to their excellent survey for more detail. I will highlight two sets of results that emerge from their estimation of a FAVAR, using the standard Cholesky identification method. First, they confirm some earlier findings that the responses of real GDP were greater in the pre-1979Q3 period than in the post-1984Q1 period.[q] For example, they find that for the earlier period, a 100 basis point increase in the federal funds rate leads to a decline of industrial production of 1.6% troughing at 8 months. In the later period, the same increase in the funds rate leads to a -0.7% decline troughing at 24 months. The second set of results concerns the price puzzle. They find that in a standard VAR the results for prices are very sensitive to the specification. Inclusion of a commodity price index does not resolve the price puzzle, but inclusion of a measure of expected inflation does resolve it in the post-1984:1 period. In contrast, there is no price puzzle in the results from their FAVAR estimation. Boivin et al. (2010) discuss various reasons why the monetary transmission mechanism might have changed, such as changes in the regulatory environment affecting credit and the anchoring of expectations.

Barakchian and Crowe (2013) estimate many of the standard models, such as by those by Bernanke and Mihov (1998), CEE (1999), Romer and Romer (2004), and Sims and Zha (2006b), splitting the estimation sample in the 1980s and showing that the impulse response functions change dramatically. In particular, most of the specifications estimated from 1988 to 2008 show that a positive shock to the federal funds rate *raises* output and prices in most cases.

Another source of time variation is state-dependent or sign-dependent effects of monetary shocks on the economy. Cover (1992) was one of the first to present evidence

[q] See, for example, Faust (1998), Barth and Ramey (2002), and Boivin and Giannoni (2006).

that negative monetary policy shocks had bigger effects (in absolute value) than positive monetary shocks. Follow-up papers such as by Thoma (1994) and Weise (1999) found similar results. Recent work by Angrist et al. (2013) finds related evidence that monetary policy is more effective in slowing economic activity than it is in stimulating economic activity. Tenreyro and Thwaites (forthcoming) also find that monetary shocks seem to be less powerful during recessions.

Olivei and Tenreyro (2007) estimate important seasonality in the effects of monetary shocks that is well explained by sticky wage models. They find that monetary shocks that take place in the first two quarters of the year have sizeable, but temporary, effects on output, whereas shocks that take place in the third and fourth quarters of the year have little effect on output. They explain these results with evidence on uneven staggering of labor contracts over the year: a shock that hits near the end of the year has little effect because the bulk of wage contracts is reset then, so wages can adjust immediately.

Since fall 2008, the federal funds rate has been near the zero lower bound. Thus, a key question that has arisen is how to measure shocks in light of this nonlinear constraint. Wu and Xia (2016) use a multifactor Shadow Rate Term Structure Model to estimate a shadow federal funds rate. This shadow rate can capture additional features, such as quantitative easing. Wu and Xia find that unconventional monetary policy has a noticeable impact on the macroeconomy.

3.2.3 Historical Case Studies

Given the important impact of Friedman and Schwartz's (1963) case study of monetary policy during the Great Depression, it is surprising that more case studies have not been conducted. Romer and Romer (1989)'s first narrative monetary analysis was designed to be a quasi-case study in the spirit of Friedman and Schwartz. Their dummy variable series was assigned to episodes in which the Fed decided to risk a recession in order to reduce inflation.

Velde (2009) presents one of the most striking case studies of monetary nonneutrality, based on an episode in 1724 France. In that year, the French government cut the money supply three times, resulting in a cumulative drop of 45%! The action was taken for a variety of reasons, such as long-term price targeting and worries that soldiers and creditors of the state were being hurt by the rise in prices during the previous 6 years. Velde finds that while prices on foreign exchange markets adjusted instantly, other prices adjusted slowly and incompletely and industrial output fell by 30%. The circumstances of that episode are unusually clean for a historical case study, so his evidence of monetary nonneutrality is quite compelling.

3.3 Main Identification Challenges

Several parts of Section 2 discussed some of the challenges to identification in general. Here, I review the issues that are particular important for the identification of monetary policy shocks.

3.3.1 The Recursiveness Assumption

A key assumption used by Christiano et al. (1999) was the "recursiveness assumption." Consider the trivariate model from Eq. (5) in the last section, rewritten here for convenience:

$$\begin{aligned}\eta_{1t} &= b_{12}\eta_{2t} + b_{13}\eta_{3t} + \varepsilon_{1t} \\ \eta_{2t} &= b_{21}\eta_{1t} + b_{23}\eta_{3t} + \varepsilon_{2t} \\ \eta_{3t} &= b_{31}\eta_{1t} + b_{32}\eta_{2t} + \varepsilon_{3t}\end{aligned} \tag{13}$$

CEE include more than three variables in the system, so we should think of each η_t as representing a block of variables: η_{1t} includes output, a general price index, and a commodities price index; η_{2t} is the federal funds rate; and η_{3t} contains a monetary stock measure such as M1 or M2, nonborrowed reserves, and total reserves. CEE interpret the equation for η_{2t} as the Fed's feedback rule and ε_{2t} as the monetary policy shock. They assume that current values of the η_{1t} enter the Fed's rule, so $b_{21} \neq 0$, but the money stock and reserves do not enter the rule, so $b_{23} = 0$. These are still not enough assumptions to identify the monetary policy shock because if the monetary policy shock can affect output, etc., within the period, η_{1t} will be correlated with ε_{2t} so we cannot use OLSs. CEE thus add the additional recursiveness assumptions that none of the η_{1t} variables (output and prices) is affected by the monetary policy shock or the monetary aggregates within the period, ie, $b_{12} = b_{13} = 0$. In practice, this is just a Cholesky decomposition generalized to blocks of variables. Since CEE focused only on the monetary policy shock, they did not need to make more assumptions to identify shocks within the first and third block.

It is important to emphasize, however, the importance of the recursiveness assumption for identification. All of CEE's results depend on setting $b_{12} = 0$, meaning that output and prices are not allowed to respond to changes in the federal funds rate within the period. Note that this assumption is at odds with some later estimated New Keynesian DSGE models. For example, Smets and Wouters' (2007) estimated model implies that output, hours, and inflation should respond immediately to the monetary policy shock (see figure 6 of their paper). The estimated DSGE model of CEE (2005) does not imply an immediate response, but only because they assume that no agents can react to the monetary policy shock within the period. They make this theoretical assumption because they estimate their model parameters to match the impulse responses of their VARs which identify the monetary policy shock with the recursiveness assumption.

Even research that develops external instruments typically uses the recursiveness assumption. For example, Romer and Romer (2004) develop a new measure of monetary policy shocks using narrative methods and Greenbook forecasts, but when they study the effects on output and prices, they impose the additional constraint that $b_{12} = b_{13} = 0$. They do so because they do not view their estimated shock as being pure, and thus also use the recursivity assumption as "exogeneity insurance." Coibion's (2012) generalization of the Romer and Romer procedures also imposes the constraint. Barakchian and

Crowe (2013) use HFI from fed funds futures, but nevertheless invoke the recursiveness assumption in their VARs. The typical FAVAR models, such as those by Bernanke et al. (2005) and Boivin et al. (2010), use the recursiveness assumption as well.

Some of the few papers that do not use the recursiveness assumption are those that use sign restrictions. Uhlig (1997, 2005), Faust (1998), Faust et al. (2004), Arias et al. (2015a), and Amir Ahmadi and Uhlig (2015) are able to avoid imposing the zero restriction associated with the recursiveness assumption by instead using sign restriction, also known as "set identification" or partial identification. For example, Uhlig (1997, 2005) imposes the restriction that contractionary monetary policy shocks cannot raise prices. Faust et al. (2004) constrain $-0.1 \leq b_{12} \leq 0$ for the output and price equations. The sign restriction papers can often yield confidence sets that imply possibly positive effects of contractionary monetary policy on output (eg, Uhlig, 2005; Faust et al., 2004; Amir Ahmadi and Uhlig, 2015).

In Section 3.4, I will investigate the importance of the recursiveness assumption in more detail.

3.3.2 Foresight Problems

Section 2.5 discussed how two types of foresight could create problems in identifying shocks and their effects. Both types of foresight are particularly important for monetary policy, and significant progress has been made recently both in appreciating their importance and in developing methods for addressing them.

The first type of foresight problem is foresight on the part of policymakers. As an illustration of the problem, suppose the Fed follows a simple policy rule:

$$ff_{\mathrm{t}} = \alpha_1 E_t(\Delta_h y_{t+h}) + \alpha_2 E_t(\Delta_h \pi_{t+h}) + \varepsilon_{ft} \tag{14}$$

where ff is the federal funds rate, y is the log output, and π is the inflation. Δ_h is the change in the variable from t to $t+h$. The Fed sets interest rates based on its expectations of the future path of output and inflation because it is aware of the lags in the effects of monetary policy. I have modeled this simply as expectations of changes from t to t+h, but the argument applies for more general notions of expectations about the path.

The usual SVAR specification assumes that the Fed's expectations about the future paths of output and inflation are adequately captured by the current and lagged values of the (typically) few macroeconomic variables included in the SVAR. This is a strong assumption. The idea that identified monetary shocks might be incorrectly mixing systematic responses to the Fed's expectations was first highlighted by Sims (1992), who argued that the price puzzle was due to the Fed observing more information on inflation. He advocated the incorporation of sensitive commodity prices to address the problem. However, this inclusion does not always get rid of the price puzzle. Bernanke et al.'s (2005) FAVARs are another method for incorporating more information. The FAVAR, which typically contains over one hundred series, has the benefit that it is much more likely to condition on relevant information for identifying shocks. The FAVAR method nonetheless relies on the assumption that linear combinations of publicly available series fully capture the Fed's expectations.

In a 2000 paper, Romer and Romer presented evidence suggesting that the Fed had superior information when constructing inflation forecasts compared to the private sector. To see the problem this asymmetric information presents, rewrite Eq. (14) as:

$$\begin{aligned} ff_t = {} & \alpha_1 E_t^p(\Delta_h y_{t+h}) + \alpha_2 E_t^p(\Delta_h \pi_{t+h}) + \alpha_1 \left[E_t^f(\Delta_h y_{t+h}) - E_t^p(\Delta_h y_{t+h})\right] \\ & + \alpha_2 \left[E_t^f(\Delta_h \pi_{t+h}) - E_t^p(\Delta_h \pi_{t+h})\right] + \varepsilon_{ft} \end{aligned} \tag{15}$$

In this equation, E_t^p denotes expectations based on private agent information and E_t^f denotes expectations based on the Fed's information. If information is symmetric and publicly available, then the two terms in square brackets will be zero and methods that incorporate sufficient amounts of public information should be able to identify the monetary policy shock ε_{ft} correctly. If, on the other hand, the Fed has superior information, the terms in brackets will not be zero and an SVAR or an FAVAR will produce an incorrectly identified monetary policy shock, $\widetilde{\varepsilon}_{ft}$ that consists of two components, the true shock as well as a component based on the informational superiority of the Fed:

$$\widetilde{\varepsilon}_{ft} = \varepsilon_{ft} + \alpha_1 \left[E_t^f(\Delta_h y_{t+h}) - E_t^p(\Delta_h y_{t+h})\right] + \alpha_2 \left[E_t^f(\Delta_h \pi_{t+h}) - E_t^p(\Delta_h \pi_{t+h})\right] \tag{16}$$

Barth and Ramey (2002) suggested that the problem might be corrected by controlling for Fed forecasts in VARs. They augmented their monetary VARs with Greenbook forecasts of inflation and output in order to determine whether controls for the Fed's superior information would make the price puzzle disappear in their early sample from 1965 to 1979. They found that even with the controls for the Greenbook forecasts, the price puzzle was still very strong in this early sample (see figure 7 of their paper).

Romer and Romer (2004) (R&R) combined the use of Greenbook forecasts with narrative methods to construct a new measure of monetary policy shocks. First, they derived a series of *intended* federal funds rate changes during FOMC meetings using narrative methods. Second, in order to separate the endogenous response of policy to information about the economy from the exogenous shock, they regressed the intended funds rate change on the current rate and on the Greenbook forecasts of output growth and inflation over the next two quarters. They then converted the estimated residuals based on the FOMC meeting frequency data to monthly and used them in dynamic regressions for output and other variables. They found very large effects of these shocks on output.

John Cochrane's (2004) NBER EFG discussion of the Romer and Romer paper highlights how their method can identify movements in monetary policy instruments that are exogenous to the error term of the model. If the Greenbook forecast of future GDP growth contains all of the information that the FOMC uses to make its decisions, then that forecast is a "sufficient statistic." Any movements in the target funds rate that are not predicted by the Greenbook forecast of future *output* can be used as an instrument to identify the causal effect of monetary policy on *output*. Analogously, any movements in the target funds rate that are not predicted by the Greenbook forecast of *inflation* can be used as an instrument to identify the causal effect of monetary policy on *inflation*. The idea is that if the Fed responds to a shock for reasons other than its effect on future output or future inflation, that response can be used as an instrument for output or inflation. Cochrane states the following proposition in his discussion:

> Proposition 1: To measure the effects of monetary policy on output, it is enough that the shock is orthogonal to output forecasts. The shock does not have to be orthogonal to price, exchange rate, or other forecasts. It may be predictable from time *t* information; it does not have to be a shock to the agent's or the Fed's entire information set (Cochrane, 2004).

Cochrane's conceptualization of the issue of identifying movements in monetary policy that are exogenous to the error term in the equation is an important step forward. Note, however, that what Cochrane calls a "shock" is not the same as the definition of shock that I use in this chapter. Cochrane's notion of a shock is not a primitive structural shock, but rather a useful instrument for estimating the effect of monetary policy on output, etc.

The possibility of asymmetric information between the Fed and the private sector leads to a further complication, though. If the Federal Reserve has superior information, then any action or announcement by the Fed presents a signal extraction problem for private agents. Private agents observe $\widetilde{\varepsilon}_{ft}$ in Eq. (14), but they know that it is composed of the true shock as well as the systematic component of the Fed's rule based on the Fed's informational advantage. The problem can easily be extended to include the possibility that $\widetilde{\varepsilon}_{ft}$ also contains time-varying inflation or output targets that are unobserved by the public. Gürkaynak et al. (2005) argue that their estimated negative effects of an unanticipated rise in the federal funds rate on long-term forward rates can be explained as the response to information revealed by the Fed action about inflation targets.

The second type of foresight problem is news about future policy actions. Campbell et al. (2012) argue that the Fed has been using forward guidance since the early 1990s. This means that many changes in the federal funds rate are in fact *anticipated* in advance. As discussed in Section 2.5 on foresight, foresight about future movements in policy variables can lead to a nonfundamental moving average representation. This would imply that standard VARs typically cannot be used to identify the shocks.

Fortunately, the monetary literature has developed excellent methods for identifying news shocks. As discussed in Section 2.3.4, research by many, such as Kuttner (2001), Cochrane and Piazzesi (2002), Gürkaynak et al. (2005), Piazzesi and Swanson (2008), Barakchian and Crowe (2013), Gertler and Karadi (2015), and Nakamura and

Steinsson (2015), has used the movements of federal funds and other interest rate futures in small windows around FOMC announcements to identify unexpected Fed policy actions. Exploiting the information in interest rate futures is an ideal way to construct "news" series. D'Amico and King (2015) study the effects of anticipated monetary policy by combining information on expectations, as in Campbell et al. (2012), with sign restrictions in an SVAR. In particular, they identify a monetary news shock by restricting the responses of the expected short term rate to move in the opposite direction of expected inflation and expected output.

3.4 Summary of Recent Estimates

Table 1 summarizes some of the main results from the literature on the impact of the identified monetary shock on output, the contribution of monetary shocks to output fluctuations, and whether the price puzzle is present. Rather than trying to be encyclopedic in listing all results, I have chosen leading examples obtained with the various identifying assumptions. In addition, I attempted to standardize the results by normalizing the peak of response of the federal funds rate to 100 basis points; this standardization does not control for differences in persistence of the response as I discuss later.

As Table 1, the standard Christiano et al. (1999) SVAR, the Faust et al. (2004) HFI, Uhlig's (2005) and Amir Ahmadi and Uhlig's (2015) sign restrictions, Smets and Wouters' (2007) estimated DSGE model, and Bernanke et al.'s (2005) FAVAR all produce rather small effects of monetary policy shocks. Also, most are plagued by the price puzzle to greater or lesser degree. On the other hand, Romer and Romer (2004), Coibion (2012), Barakchian–Crowe (2013), and Gertler–Karadi (2015) all find larger impacts of a given shock on output.

I will also summarize briefly the effects on other variables from some of the leading analyses. A particularly comprehensive examination for many variables is conducted by Boivin et al.'s (2010) with their FAVAR. Recall that they obtained different results for the pre- vs post-1980 period. For the period from 1984m1 to 2008m12, they found that a positive shock to the federal funds rate leads to declines in a number of variables, including employment, consumption expenditures, investment, housing starts, and capacity utilization.

3.5 Explorations with Three Types of Monetary Shocks

I now explore the robustness of the effects of monetary policy shocks using some of the new methods introduced in the literature to deal with both the foresight problems and the recursiveness assumption. For reference, I begin by estimating the classic Christiano et al.'s (1999) type of specification and then move on to the Romer and Romer (2004) shock and Gertler and Karadi's (2015) HFI shock.

Table 1 Summary of monetary policy shock effects on output and prices

Paper	Method, sample	Trough effect of 100 basis point funds peak	% of output explained by shock	Price puzzle?
Christiano et al. (1999)–FFR identification	SVAR, 1965q3–1995q3	−0.7% at 8 quarters	4.4% at 2 years	Yes, but very small
Faust et al. (2004)	HFI, 1991m2–2001m7	−0.6% at 10 months		
Romer and Romer (2004)	Narrative/Greenbook 1970m1–1996m12	−4.3% at 24 months	Major part	No, but prices do not change for 22 months
Uhlig (2005)	Sign restrictions, 1965m1–1996m12	Positive, but not statistically different from 0	5–10% at all horizons.	No (by construction)
Bernanke et al. (2005)	FAVAR, 1959m1–2001m7	−0.6% at 18 months	5% at 5 years	Yes
Smets–Wouters (2007)	Estimated DSGE model, 1966Q1–2004Q4	−1.8 at 4 quarter trough	10% at 1 year (trough)	No
Boivin et al. (2010)	FAVAR, 1962m1–1979m9, 1984m1–2008m12	−1.6% at 8 months in early period −0.7% at 24 months in later period		Only in the early period
Coibion (2012)	"Robust" Romer–Romer methods, 1970m1–1996m12	−2% at 18 months	"Medium" part	Yes, sometimes
Barakchian–Crowe (2013)	HFI, Romer hybrid VAR, 1988m12–2008m6	−5% at 23 months	50% at 3 years	Yes
Gertler–Karadi (2015)	HFI-proxy SVAR, 1979m7–2012m6 (1991m1–2012m6 for instruments)	−2.2% at 18 months	?	No
Amir Ahmadi and Uhlig (2015)	Bayesian FAVAR with sign restrictions, 1960m2–2010m6	−1.3% at 9 months	7% at 24 months	No (by construction)

3.5.1 The Christiano et al. (1999) Benchmark

Christiano et al. (1999) presented estimates based on shocks identified using the recursiveness assumption and showed robust results that were generally consistent with conventional views on the effect of monetary shocks. Here I study how the results change when the sample is updated.

I estimate a specification similar to Christiano et al.'s (1999) specification but use Coibion's (2012) macroeconomic variables for the first block. In particular, I use monthly data and include log industrial production, the unemployment rate, the log of the CPI, and the log of a commodity price index in the first block. The second block consists of the federal funds rate. The third block consists of the logs of nonborrowed reserves, total reserves, and M1. Thus, the innovation to the federal funds rate (orthogonal to contemporaneous values of the first block variables and lags of all of the variables) is identified as the monetary policy shock.

Fig. 1 shows the estimated impulses for this SVAR. The solid black line and shaded areas are the point estimates and 90% bootstrap confidence bands for the system estimated over CEE's sample period, 1965m1–1995m6. The responses look like classic effects of monetary policy shocks. The Federal funds rate jumps up temporarily but then falls back to 0 by 6 months. This temporary blip in the funds rate, however, sets off a prolonged recession. Industrial production begins to fall in the next month and troughs 21 months later. Unemployment rises and peaks around 23 months later. Both unemployment and industrial production return to normal after 4 years. Prices rise slightly for the first few months, but then follow a steady path down, settling at the new lower level after 4 years. Nonborrowed reserves, total reserves, and M1 fall and then recover after 3 years. Nonborrowed reserves display some unusual oscillations, though. For the most part, these responses look very similar to the ones shown in figure 3 of CEE (1999).

The blue short dashed lines in the same figures show the responses for the sample from 1983m1 to 2007m12. The sample stops in 2007 both to exclude the financial crisis and for the practical reason that nonborrowed reserves became negative starting in 2008. The results change dramatically and imply that increases in the federal funds rate lower the unemployment rate. These results echo those of Barakchian and Crowe (2013), who show that the leading specifications imply expansionary effects in the sample from 1988 through 2007.

The red long-dashed lines show the results of a simplified model for the sample 1983m1–2007m12. This model omits M1, nonborrowed reserves, and total reserves. In this specification, there is still a small amount of expansionary effect on output and unemployment at the beginning, but then the more standard contractionary effects take hold. Prices never fall, however.

Table 2 shows various measures of the importance of monetary policy shocks for industrial production in CEE's specification. The first column shows the trough effect on output of a shock that raises the funds rate to a peak of 100 basis points. Even in CEE's original sample, the effects are very modest, less than a −0.5% fall. When estimated over the period

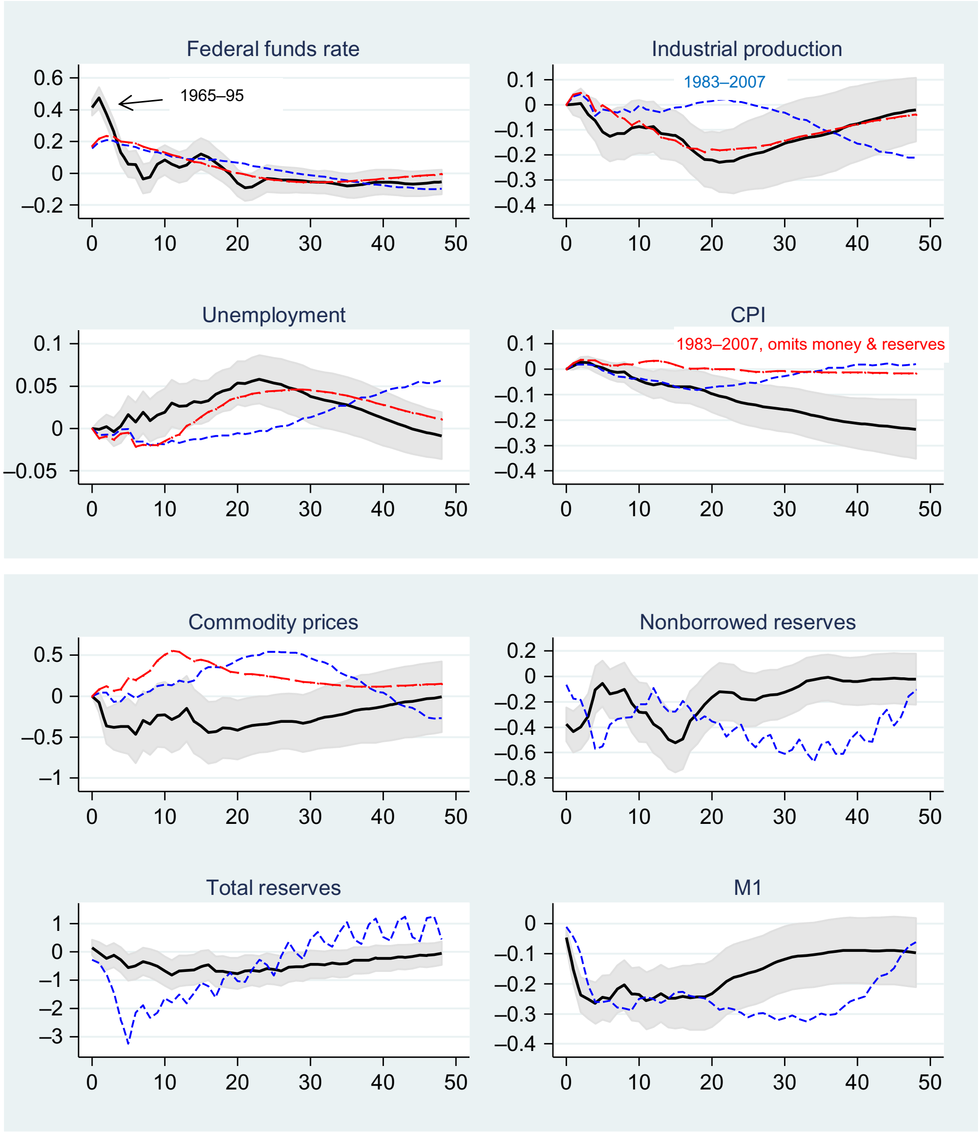


Fig. 1 Christiano et al. (1999) identification. 1965m1–1995m6 full specification: *solid black lines*; 1983m1–2007m12 full specification: *short dashed blue* (*dark gray* in the print version) *lines*; 1983m1–2007m12, omits money and reserves: *long-dashed red* (*gray* in the print version) *lines*. Light gray bands are 90% confidence bands.

1959 through 2007, the effects are less than half. The other columns show the forecast error variance decompositions at 24 months. These indicate that monetary policy shocks account for less than 7% of the forecast error variance in the original sample and less than 1% in the longer sample. A reasonable interpretation of the decline in the contribution of monetary shocks to output volatility is improved, less erratic monetary policy.

Table 2 Effects of Monetary Policy Shocks on Industrial Production: My Estimates

Method and sample	Trough effect of 100 basis point funds peak (%)	Forecast error variance decompositions (%) 24 months
CEE: 1965m1–1996m6	−0.48	6.6
CEE: 1959m1–2007m12	−0.20	0.5
R&R VAR: 1969m3–1996m12	−1.38	8.8
R&R VAR: 1969m3–2007m12	−0.83	2.7
R&R, Jordà method: 1969m3–1996m12	−0.83	
R&R, Jordà method, no recursiveness assumption: 1969m3–1996m12	−0.90	
Gertler–Karadi, proxy SVAR: 1990m1–2012m6	−2.2	
Gertler–Karadi Jordà method: 1990m1–2012m6	−1, but then rises to +4	

Notes: See text for the description of the CEE (Christiano et al., 1999), R&R (Romer and Romer, 2004) VARs, and Gertler–Karadi proxy SVARs.

3.5.2 Greenbook/Narrative Identification of Shocks

Next, I consider the effects of the shocks identified by Romer and Romer (2004). As discussed in Section 3.3.2, Romer and Romer (2004) (R&R) sought to overcome the problem of superior Federal Reserve information by regressing the federal funds target rate on the Greenbook forecasts at each FOMC meeting date and using the residual as the monetary policy shock. They find much larger effects of monetary policy than CEE do. Coibion (2012) explores many possible reasons for the differences and provides very satisfactory and revealing answers. In particular, he finds that R&R's main results, based on measuring the effect of their identified shock using a single dynamic equation, are very sensitive to the inclusion of the period of nonborrowed reserves targeting, 1979–82, and the number of lags (the estimated impact on output is monotonically increasing in the number of lags included in the specification). In addition, their large effects on output are linked to the more persistent effects of their shock on the funds rate. In contrast, R&R's hybrid VAR specification, in which they substituted their (cumulative) shocks for the federal funds rate (ordered last) in a standard VAR, produces results implying that monetary policy shocks have "medium" effects. Coibion (2012) goes on to show that the hybrid model results are consistent with numerous other specifications, such as GARCH estimates of Taylor Rules (as suggested by Hamilton, 2010; Sims–Zha, 2006a) and time-varying parameter models as in Boivin (2006) and Coibion and Gorodnichenko (2011). Thus, he concludes that monetary policy shocks have "medium" effects. In particular, a 100 basis point rise in the federal funds rate leads industrial production to fall 2–3% at its trough at around 18 months.

In my explorations, I use the Coibion version of the R&R hybrid VAR, a monthly VAR with the log of industrial production, the unemployment rate, the log of the CPI, the log of a commodity price index in the first block, and the cumulative Romer and Romer shock replacing the federal funds rate. This specification uses the recursiveness assumption as well, placing the funds rate last in the ordering and thus assuming that the monetary shock cannot affect the macroeconomic variables within the month.

Fig. 2A shows the results, with the solid lines and confidence bands estimated using the original R&R shocks on the original R&R sample of 1969m3–1996m12. These results also match the classic monetary policy effects. Output falls within a month or two, while unemployment rises. Prices remain constant until around 9 months, when they fall steadily until they bottom out during the 4th year after the shock. A qualitative difference with the CEE results is that the response of the federal funds rate is more persistent in the R&R VAR.

The short dashed blue lines show the responses based on the sample from 1983m1 to 2007m12. I constructed new R&R shocks by reestimating their FOMC meeting regression for the later sample, using the updates by Wieland and Yang (2015). I converted these shocks to monthly and then used them in the VAR estimated over the same later sample. The results are similar to those found by Barakchian and Crowe (2013): contractionary monetary policy shocks appear to be expansionary.

The long-dashed red lines show the responses based on the sample from 1969m3 to 2007m12. The R&R shock is based on reestimating their FOMC meeting regression for the entire sample. The results for the full sample look more like those for the original R&R sample, but with more muted effects on output and unemployment.

Rows 3 and 4 of Table 2 show the trough effects and variance decompositions for the R&R VAR. In their original sample, the trough effect on output is −1.38, which is substantially larger than the results using CEE.[r] The forecast error variance decomposition implies that monetary policy shocks account for 9% of the variance for horizons at 24 months.[s] As with most monetary shock specifications, however, the effects are considerably less if we include more recent periods in the sample.

An odd feature of the impulse responses shown in Fig. 2A is the robust rebound of industrial production after 2 years of recession. The peak of industrial production at 48 months is the same magnitude as the trough at 13 months. One possible explanation is that

[r] My numbers are slightly different from those of Coibion's for the original sample because he normalized by the impact effect on the funds rate rather than the peak response of the funds rate. Emi Nakamura has suggested that it might be better to compare the integral of the output response to the integral of the funds rate response because this measure incorporates persistence. I found that this measure sometimes behaved oddly because of the tendency of some of the variables to oscillate around zero.

[s] Neither Romer and Romer (2004) nor Coibion (2012) conducted forecast error variance decompositions. Their claim of large or "medium" effects was based on comparing the actual paths of output to the predicted paths implied by the estimated monetary policy shocks.

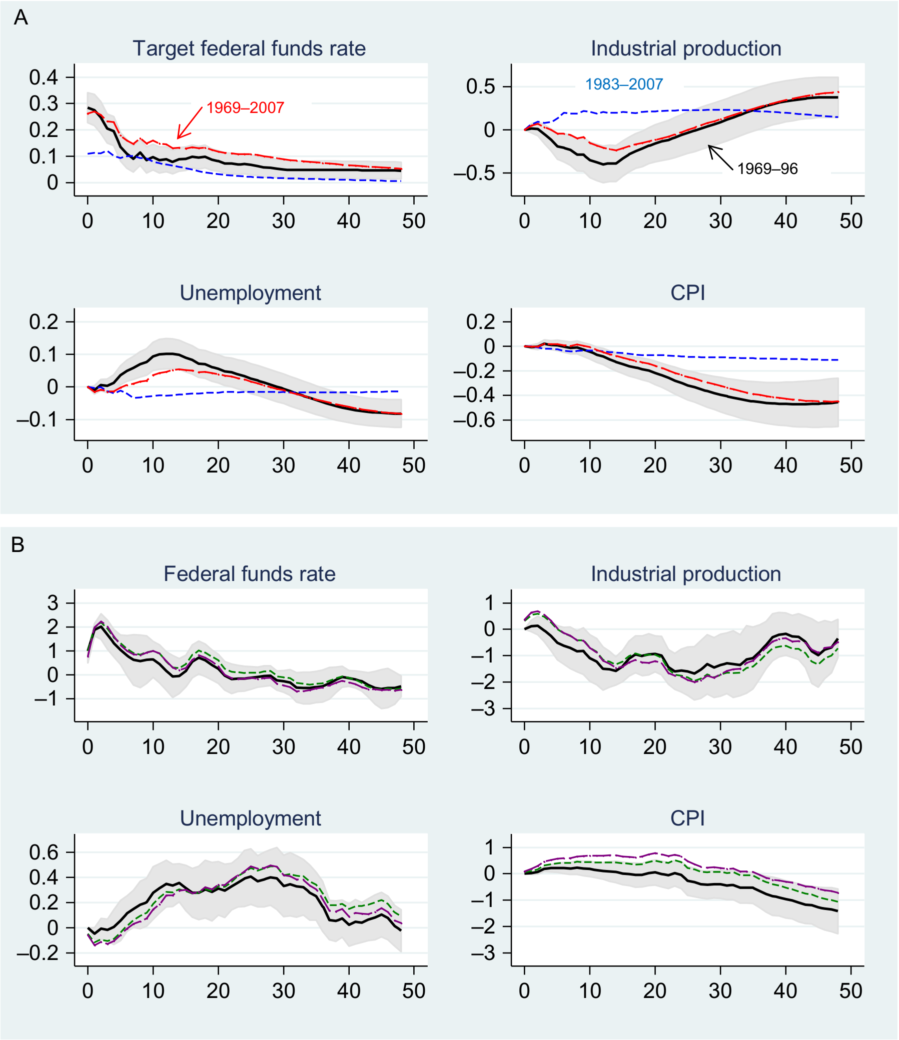


Fig. 2 Romer and Romer monetary shock. (A) Coibion VAR 1969m3–1996m12: *solid black lines*; 1983m1–2007m12: *short dashed blue* (*dark gray* in the print version) *lines*; 1969m3–2007m12: *long-dashed red* (*gray* in the print version) *lines*. (B) Jordà local projection method, 1969m3–1996m12 recursiveness assumption: *solid black lines*; no recursiveness assumption: *short dashed green* (*dark gray* in the print version) *lines*; no recursiveness assumption, FAVAR controls: *long dashed purple* (*dark gray* in the print version) *lines*. Light gray bands are 90% confidence bands.

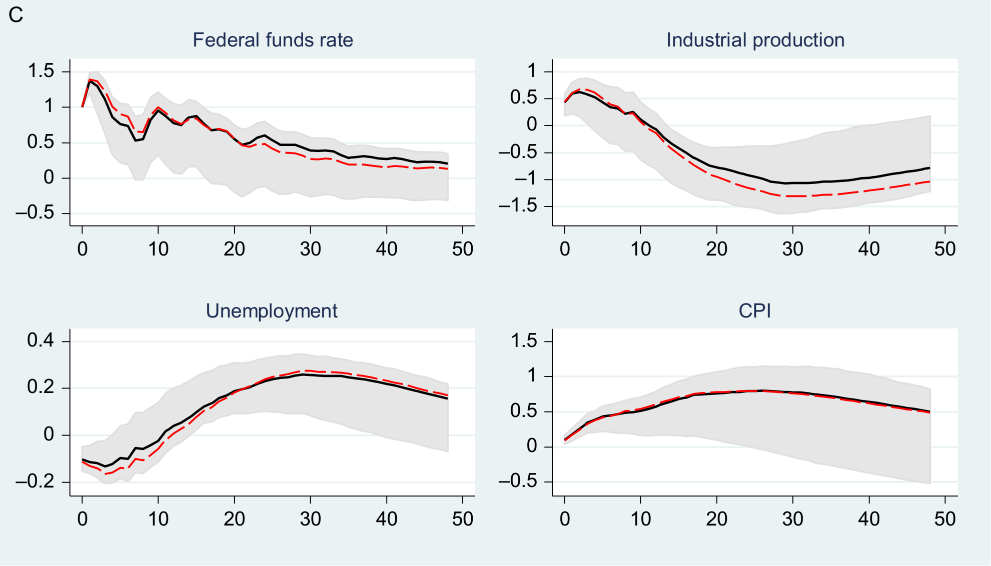


Fig. 2—Cont'd (C) Proxy SVAR, 1969m3–1996m12: *solid black lines*; 1969m3–2007m12: *long-dashed red* (*gray* in the print version) *lines*.

misspecification of the VAR is distorting the estimated impulse responses. One way to assess this hypothesis is to use Jordà's (2005) local projection method. As discussed in Section 2.4, the Jordà method puts fewer restrictions on the impulse responses. Rather than estimating impulse responses based on nonlinear functions of the reduced form parameters, the Jordà method estimates regressions of the dependent variable at horizon $t+h$ on the shock in period t and uses the coefficient on the shock as the impulse response estimate.

To investigate the results of this less restrictive specification, I estimate the following series of regressions:

$$z_{t+h} = \alpha_h + \theta_h \cdot \text{shock}_t + \text{control variables} + \varepsilon_{t+h} \tag{17}$$

The z is the variable of interest. The control variables include two lags of the R&R shock, the federal funds rate, the log of industrial production, the unemployment rate, the log of the CPI, and the log of the commodity price index.[t] In addition, to preserve the recursiveness assumption, I include *contemporaneous* values of the log of industrial production, unemployment rate, and the logs of the two price indices. The coefficient θ_h gives the response of z at time $t+h$ to a shock at time t. As discussed in Section 2, ε_{t+h} will be serially correlated, so the standard errors must incorporate a correction, such as Newey–West.

Fig. 2B shows the impulse responses estimated using the Jordà method on R&R's original sample 1969m3–1996m12. The relevant impulse responses are indicated by

[t] The point estimates are similar if more lags are included.

the black solid lines. While the responses are somewhat more erratic, they display more coherent dynamics. In particular, rather than the swing from recession to boom for industrial production implied by the VAR estimates, the Jordà estimates imply a more persistent decline in output (and rise in unemployment) that slowly returns to normal.

As discussed earlier, the R&R VAR still imposes the recursiveness assumption. If one believes that the Greenbook forecasts incorporate all relevant information used by the Fed, then one does not need to impose the additional CEE recursiveness assumption. To determine the effect of removing the restriction that output and prices cannot respond to the shock within the month, I reestimate the Jordà regressions omitting the contemporaneous values of all variables other than the R&R shock. The results of this estimation are shown by the green short dashed lines in Fig. 2B. Many aspects of the responses are similar to those obtained with the recursiveness assumptions. However, there are several key differences. First, the estimates imply that a shock that raises the funds rate is initially expansionary: industrial production rises and the unemployment rate falls for the first several months, and the point estimates are statistically different from zero (not shown). Second, there is a pronounced price puzzle for the first 2 years, and most of those estimates are statistically different from zero.

One possible explanation for these puzzles is a failure of the Greenbook forecasts to capture all of the information the Federal Reserve uses. To examine whether expanding the information set helps eliminate the price puzzle, I reestimate the nonrecursive Jordà model with the following alternative controls: two lags each of the R&R shock, the federal funds rate, the dependent variable, and updates of Boivin et al.'s (2010) five FAVAR factors.[u] The results are shown as the long dashed purple (*dark gray* in the print version) line in Fig. 2B. In this case, the price puzzle is even worse and the initial expansionary effects on output and unemployment are no better. Thus, including FAVAR-type factors does not reproduce the results obtained using the recursiveness assumption.

The proxy SVAR is another method that can be used to relax the recursiveness assumption. Kliem and Kriwoluzky (2013) use this method to reconcile VAR monetary shocks with the Romers' narrative shocks. They do not, however, explore effects on output, prices, or other variables. To investigate the results using this method, I estimate the reduced form of Coibion's system with the federal funds rate substituted for the cumulative R&R shock and with R&R's monetary policy shock as an external instrument following Stock and Watson's (2012) and Mertens and Ravn's (2013) proxy SVAR method (see Section 2 for a description).

Fig. 2C shows the results for the original sample (1969–1996) and the full sample from 1969 through 2007. The shaded areas are 90% confidence bands using Mertens and Ravn's methods for the original sample estimates. In both samples, a shock to monetary policy raises the federal funds rate, which peaks at 1.4% by the month after the shock and

[u] I am indebted to Shihan Xie for providing her updates of the Boivin et al. estimated factors.

falls slowly to 0% thereafter. The response of industrial production (shown as the red long-dashed line in Fig. 2) is different from the one obtained using the hybrid VAR. In particular, industrial production now rises above normal for about 10 months and then begins falling, hitting a trough at about 29 months. Normalized by the funds rate peak, the results imply that a shock that raises the funds rate to a peak of 100 basis points first raises industrial production by 0.5% at its peak a few months after the shock and then lowers it by −0.9% by 29 months. The unemployment rate exhibits the same pattern in reverse. After a contractionary monetary policy shock, it falls by 0.1 percentage points in the first year and then begins rising, hitting a peak of about 0.2 percentage points at month 30. The behavior of the CPI shows a pronounced, statistically significant prize puzzle.

In sum, relaxing the recursiveness assumption imposed by Romer and Romer's hybrid VAR leads to several puzzles. A contractionary monetary policy shock is now expansionary in its first year and the price puzzle is very pronounced.

The most obvious explanation for these results is that the FOMC responds to more information than even the Greenbook forecast and FAVAR factors capture, and making the R&R shock orthogonal to current output and prices (ie, the recursiveness assumption) helps cleanse the shock of these extra influences. However, this means that even with the R&R shock, one is forced to make the recursiveness assumption, which does not have a solid economic basis. As discussed earlier, leading New Keynesian models, such as Smets and Wouters (2007), imply immediate effects of monetary policy shocks on output and prices.

This exploration highlights the importance of additional restrictions imposed in standard monetary models, as well as the importance of the sample period when estimating the effects of monetary shocks. Without the additional recursiveness assumption, even narrative methods can produce puzzling results. Furthermore, as highlighted by Barakchian and Crowe (2013), many of the methods that produce classic monetary shock results in samples through the mid-1990s produce puzzles when estimated over later samples. In particular, contractionary monetary shocks seem to have expansionary effects in the first year and the price puzzle is pervasive. A plausible explanation for the breakdown in results in the later sample is an identification problem: because monetary policy has been conducted so well in the last several decades, true monetary policy shocks are rare. Thus, it is difficult to extract meaningful monetary shocks that are not contaminated by problems with foresight on the part of the monetary authority.

3.5.3 HFI Shocks

As discussed in the previous sections, numerous papers have used HFI methods to deal with possible foresight about monetary policy changes. Part of the literature focuses only on the effects on interest rates and asset prices (eg, Krishnamurthy and Vissing-Jorgensen, 2011; Hanson and Stein, 2015). Nakamura and Steinsson (2015) link their estimated interest rate changes to output effects by calibrating a New Keynesian model. The strength of the effect, however, depends crucially on the assumed intertemporal elasticity of substitution. For this reason, it is also important to estimate direct links in the data as well.

A recent paper by Gertler and Karadi (2015) combines HFI methods with proxy SVAR methods to investigate the effects on macroeconomic variables. They have two motivations for using these methods. First, they seek to study the effect of monetary policy on variables measuring financial frictions, such as interest rate spreads. The usual Cholesky ordering with the federal funds rate ordered last imposes the restriction that no variables ordered earlier respond to the funds rate shocks within the period. This is clearly an untenable assumption for financial market rates. Second, they want to capture the fact that over time the Fed has increasingly relied on communication to influence market beliefs about the future path of interest rates (forward guidance).

In the implementation, Gertler and Karadi estimate the residuals using monthly data from 1979 to 2012, but then execute the proxy SVAR from 1991 to 2012 since the instruments are only available for that sample. Fig. 3A replicates the results from Gertler and Karadi's baseline proxy SVAR for figure 1 of their paper.[v] This system uses the 3-month ahead fed funds futures (ff4_tc) as the shock and the 1-year government bond rate as the policy instrument. The other variables included are log of industrial production, log CPI, and the Gilchrist and Zakrajšek (2012) excess bond premium spread. The results show that a shock raises the 1-year rate, significantly lowers industrial production, does little to the CPI for the first year, and raises the excess bond premium. In order to put the results on the same basis as other results, I also estimated the effect of their shock on the funds rate. The results imply that a shock that raises the federal funds rate to a peak of 100 basis points (not shown) lowers industrial production by about −2%.

I explore the robustness of the results by estimating the effects of their shocks in a Jordà local projection framework. The control variables are two lags of the shock itself, the interest rate on 1-year government bonds, industrial production, the CPI, and the Gilchrist–Zakrajsek (2012) excess bond premium spread. I do not include current values of these other variables, so I am not imposing the recursiveness assumption.

Fig. 3B shows the results. The impulse responses look very different from those using the proxy SVAR method. The interest rate rises more slowly, but then remains high for a much longer time. Output does not respond for a year, but then rises. Prices respond little for the first 30 months, but then finally fall.

I then conducted some further investigations of the Gertler-Karadi shock. Several features emerge. First, the shock is not zero mean. The mean is −0.013 and is statistically different from zero. Second, it is serially correlated; if I regress it on its lagged value, the coefficient is 0.31 with a robust standard error of 0.11. This is not a good feature since it is

[v] The only difference is that I used 90% confidence intervals to be consistent with my other graphs.

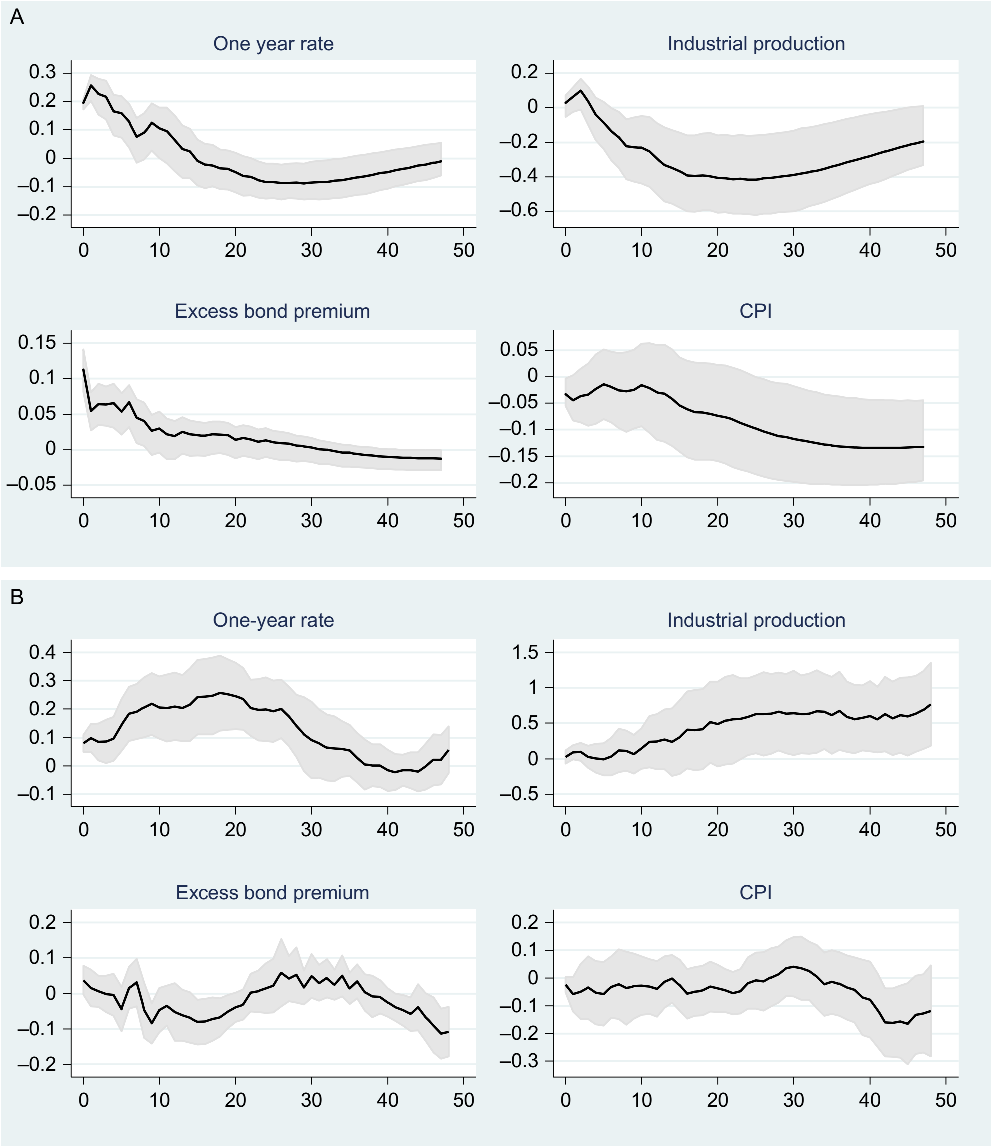


Fig. 3 Gertler–Karadi's monetary shock. (A) Gertler–Karadi's monetary proxy SVAR, VAR from 1979m7 to 2012m6, instrument from 1991m1 to 2012m6. (B) Gertler–Karadi monetary shock, Jordà 1990m1–2012m6. Light gray bands are 90% confidence bands.

supposed to capture only unanticipated movements in interest rates. In my local projection framework implementation, I include lagged values of the shock, so my procedure purges the shock of this serial correlation. I discovered that the serial correlation is induced by the method that Gertler and Karadi use to convert the announcement day shocks to a monthly series.[w] Third, if I used FOMC-frequency data to regress the shock on all of the Greenbook variables that the Romers used to create their shock, the R-squared of the regression is 0.21 and I can reject that the coefficients are jointly zero with a p-value of 0.027.[x] Thus, the Gertler–Karadi variable is predicted by Greenbook projections. Gertler and Karadi also worried about this issue, but they performed a robustness check based only on the *difference* between private forecasts and Greenbook forecasts. They found a much lower R-squared (see their table 4). When they use their purged measure, they find greater falls in industrial production. I explored the effect of using a version of their measure that was (i) orthogonal to the Romer Greenbook variables; and (ii) converted to a monthly basis the same way that the Romer's converted their data, in the Jordà framework. The results (not shown) were still puzzling.

Why does the Jordà method give such different estimates from the proxy SVAR? One possible explanation is the different method and sample used to estimate the impulse response function. Gertler and Karadi's impulse responses functions are constructed as nonlinear functions of the reduced form VAR parameters estimated on data from 1979 through 2012; the Jordà method estimates are for the 1991–2012 sample and are direct projections rather than functions of reduced form VAR parameters. Since the estimates of the impact effects on industrial production are near zero for both methods, the entire difference in the impulse responses is due to the differences in the dynamics implied by Gertler and Karadi's reduced form VAR parameter estimates. A second possible explanation for the difference is that the rising importance of forward guidance starting in the mid-1990s means that the VAR underlying the proxy SVAR is misspecified. As discussed in Section 2.5, anticipations of future policy actions can lead to the problem of a nonfundamental moving average representation. Gertler and Karadi's fed funds futures variable captures news well, but they do not include it directly in the SVAR; they only use it as an instrument.

3.6 Summary of Monetary Shocks

When Christiano et al. (1999) wrote their Handbook chapter, they provided what became a benchmark framework for identifying monetary policy shocks and tracing their effects. As long as the recursiveness assumption was incorporated, the results were quite robust. Since then, the literature has incorporated new methods and faced new challenges. Researchers now take instrument identification and relevance much more seriously when estimating monetary policy shocks. New methods, such as FAVARs and Greenbook forecasts, have improved the conditioning set for estimating monetary policy shocks. Structural VARs (SVARs), sign restrictions, and regime-switching models have provided alternatives to the usual Cholesky decomposition. Moreover, new measures of

[w] See footnote 11 of Gertler and Karadi (2015) for details.

[x] I am indebted to Peter Karadi for sharing with me the announcement date series.

monetary shocks have been developed using rich external data, such as narrative data, Greenbook projections, and high-frequency information from financial markets. Recently published work using shocks estimated with external data results in similar conclusions. In particular, Coibion's (2012) reconciliation of the Romer results with the VAR results suggests that a 100 basis point rise in federal funds rate lowers industrial production by about −2% at 18 months. Those results are based on data from 1969 through 1996. Gertler and Karadi's (2015) research uses HFI from fed funds futures and external instruments/proxy SVAR methods to find very similar results for a later sample.

This rosy reconciliation picture disappears, however, when the specifications are subjected to some robustness tests. My explorations have highlighted several potential issues, some of which were already noted in the literature. First, the original Christiano et al. (1999) specification, as well as many other specifications, does not hold up well in later samples. Second, lifting the recursiveness assumption can lead to estimates that imply expansionary effects of contraction monetary policy in the short run. Third, one needs to be very careful when estimating models in samples where anticipation effects may be important. For example, it is not clear that HFI shocks should be used as external instruments for otherwise standard VARs.

How should we interpret these results? I would argue that the most likely reason for the breakdown of many specifications in the later sample is simply that we can no longer identify monetary policy shocks well. Monetary policy is being conducted more systematically, so true monetary policy shocks are now rare. It is likely that what we now identify as monetary policy shocks are really mostly the effects of superior information on the part of the Fed, foresight by agents, and noise. While this is bad news for econometric identification, it is good news for economic policy.

What, then, are we to conclude about the output effects of monetary shocks? I would argue that the best evidence still remains the historical case studies, such as Friedman and Schwarz, and the times series models estimated on samples that exclude recent decades. Of course, one worries that the structure of the economy may have changed in the last few decades, but we simply do not have enough information to produce estimates with any great certainty. Monetary policy can have big effects, but it is likely that monetary shocks are no longer an important source of macro instability.

4. FISCAL SHOCKS

This section reviews the main identification methods and summarizes existing results from the empirical literature seeking to identify and estimate the effects of fiscal policy shocks. It also presents some new results comparing several leading identified shocks.

In contrast to a monetary policy shock, a fiscal shock is a more straightforward economic concept. Because the legislative and executive branches of government often make tax and spending decisions based on concerns that are orthogonal to the current state of the macroeconomy, the notion of regularly occurring fiscal policy shocks is more plausible than regularly occurring monetary policy shocks.

Measuring the empirical effects of changes in government spending and taxes on aggregate GDP and its components was an active research area for a number of decades. The large Keynesian models of the 1960s included fiscal variables and numerous academic papers estimated their effects in behavioral equations. For several decades afterward, though, research on the aggregate effects of tax and spending shocks experienced a lull, punctuated by only a few papers. Most empirical research on shocks during the 1980s, 1990s, and 2000s instead focused on monetary policy. With the onset of the Great Recession and the zero lower bound, however, research energy immediately shifted to the effects of fiscal policy. The recent literature has built on and extended the strides made by the few authors working on the topic during the long dormant period.

The following sections will discuss some of the literature since 1990 that has sought to analyze the effects of fiscal shocks. I will begin by considering government spending shocks and then discuss tax shocks.

4.1 Government Spending Shocks

In this section, I discuss shocks to government spending. When I use the term *government spending*, I mean government *purchases*, ie, G in the NIPA identity. In common parlance, however, government spending typically refers to government *outlays*, which include both government purchases and transfer payments. Economists usually consider transfer payments to be negative taxes. Thus, I will include a discussion of transfer payments in the section on the effects of tax shocks.

4.1.1 Summary of Identification Methods

Many of the identification methods summarized in Section 2 are used in the literature that analyzes the effects of shocks to government spending. These methods include SVARs with contemporaneous restrictions, sign restrictions, medium-horizon restrictions, narrative methods, and estimated DSGE models.

Perhaps the first example of what looks like a VAR-type analysis of the effects of fiscal shocks is Rotemberg and Woodford's (1992) analysis of the effects of military spending and employment on macroeconomic variables. Their purpose was to provide evidence in favor of their countercyclical markup model, showing that a "demand" shock would lead to countercyclical markups. To do this, they estimated systems with military spending, military employment, and a macroeconomic variable of interest (such as private value added and private hours worked). They included lags of the variables in the system, but restricted the VAR so that there was no feedback of the macroeconomic variables onto the military variables. In their system, identification was achieved as follows. To identify government purchases shocks that were exogenous to the economy, they followed Hall (1980, 1986) and Barro (1981) who argued that defense spending is driven by military events rather than by macroeconomic events. To identify *unanticipated* shocks, they regressed the military variables on their own lags and used the residuals. This identification strategy assumes that all relevant information for predicting military spending and employment is contained in lags of military spending and employment. They showed that their identified shocks to defense spending raised real wages.

In a paper analyzing the effects of sectoral shifts in the presence of costly mobility of capital across sectors, Ramey and Shapiro (1998) used narrative techniques to create a dummy variable capturing major military buildups. They read through *Business Week* in order to isolate the political events that led to the buildups in order to create a series that was exogenous to the current state of the economy. They also used this narrative approach to ensure that the shock was unanticipated. They stated: "We believe this approach gives a clearer indicator of unanticipated shifts in defense spending than the usual VAR approach, since many of the disturbances in the VAR approach are due solely to timing effects on military contracts and do not represent unanticipated changes in military spending" (Ramey and Shapiro, 1998, p. 175). Ramey and Shapiro (1998) estimated the effects of "war dates" by regressing each variable of interest on current values and lags of the war dates and lags of the left-hand side variable. A number of follow-up papers embedded the war dates in VARs, ordered first in the Cholesky decomposition, creating "EVARs", a term coined by Perotti (2011). These papers include Edelberg et al. (1999) and Burnside et al. (2004). Most applications typically found that while government spending raised GDP and hours, it lowered investment, consumption, and real wages. Most of these papers did not specifically estimate a multiplier, though one can typically back out the implied multiplier from the impulse responses.

In contrast, Blanchard and Perotti (2002) used an SVAR to identify both government purchases and tax shocks. They assumed that government purchases were predetermined within the quarter, and identified the shock to government purchases using a standard Cholesky decomposition with government spending ordered first. They found that government purchases shocks raised not only GDP but also hours, consumption, and real wages. Follow-up work, such as by Fatás and Mihov (2001), Perotti (2005), Pappa (2009) and Galí et al. (2007), found similar results. Mountford and Uhlig (2009) used sign restrictions and found only weak effects on GDP and no significant effect on consumption.

In Ramey (2011a), I sought to reconcile why the war dates were producing different results from the SVARs that used Cholesky decompositions. I argued that most government spending is anticipated at least several quarters in advance, so that the standard SVAR method was not identifying unanticipated shocks. In support of this idea, I showed that the shocks from an SVAR were indeed Granger-caused by the Ramey and Shapiro (1998) war dates. To create a richer narrative variable to capture the "news" part of government spending shocks, I read *Business Week* starting in 1939 and created a quantitative series of estimates of changes in the expected present value of government purchases, caused by military events. I then embedded the news series in a standard VAR, with the news ordered first in the Cholesky decomposition. In that work, I found results that were broadly consistent with the estimates based on the simple war dates.

In follow-up work, Owyang et al. (2013) and Ramey and Zubairy (2014) extended the military news series back to 1889. The military news variable tends to have low instrument relevance for samples that begin after the Korean War, though. In Ramey (2011a), I augmented my analysis by also considering shocks that were orthogonal to professional forecasts of future government purchases. Fisher and Peters (2010) created an alternative series of news based on the excess returns of defense contractor stocks for the period starting in 1958. Recent work by Ben Zeev and Pappa (forthcoming) uses the medium-horizon identification methods of Barsky and Sims (2011) to identify news shocks to defense spending from a time series model. In particular, Ben Zeev and Pappa identify defense spending news as a shock that (i) is orthogonal to current defense spending; and (ii) best explains future movements in defense spending over a horizon of 5 years.

All of these measures of anticipations have weaknesses, though. First, because they are associated with military events, there are likely confounding effects (eg, rationing, price controls, conscription, patriotic increases in labor supply). Second, as I show below, some of the shocks suffer from low first-stage *F*-statistics in some samples, indicating that they might not be relevant instruments for estimating multipliers.

Thus, there are two main differences in the shocks identified across these two classes of models. First, the SVAR shocks are more likely to be plagued by foresight problems. As I discussed in Section 2, this problem of foresight can be a serious flaw in SVARs. Second, the news alternatives are not rich enough in some subsamples and there may be confounding influences.

A more structural way to identify shocks to government purchases is through an estimated DSGE model. For example, one of the shocks identified by Smets and Wouters (2007) is a government purchases shock. Cogan et al. (2010) also estimate government spending multipliers in the context of an estimated DSGE model.

4.1.2 Summary of the Main Results from the Literature

Typically, the literature on government spending has sought to answer one or both of two main questions: (1) Are the empirical results consistent with theoretical DSGE models? (2) What are the government spending multipliers?

Let us begin by considering results that shed light on the first question. Most versions of standard neoclassical theory and standard new Keynesian theory predict that a rise in government purchases (financed with deficits or lump-sum taxes and not spent on public infrastructure, etc.) should raise GDP and hours, but should decrease consumption and real wages. Whether investment initially rises or falls depends on the persistence of the increase in government spending. It is only when one adds extra elements, such as rule-of-thumb consumers and off-the-labor supply behavior of workers that one can produce rises in consumption and real wages in a model (eg, Galí et al., 2007).

Both SVARs and EVARs that use a news variable produce qualitative similar results for some variables. For example, both typically estimate an increase in GDP and hours and a fall in investment (at least after the first year) in response to a positive government spending shock. In contrast, the SVAR typically implies a rise in consumption and real wages, whereas the EVAR predicts a fall in consumption and real wages.

The second question the literature seeks to answer is the size of "the" government purchases multiplier. Unfortunately, most estimates are not for pure deficit-financed multipliers since most rises in government spending are accompanied by a rise in distortionary taxes, typically with a lag. This caveat should be kept in mind in the subsequent discussion of multiplier estimates.

One might assume that SVARs produce bigger multipliers since they predict increases in consumption. They do not. In Ramey (2013), I compared the effects of government spending on private spending, ie, GDP minus government spending, of several shocks based on the various identification methods. If the government spending multiplier is greater than unity, then private spending must increase. I found that all of the shocks lowered private spending, but that the Blanchard–Perotti (2002) shocks lowered it more, implying a smaller multiplier.

The estimated DSGE models of Smets and Wouters (2007) and Cogan et al. (2010) produce results that are close to the neoclassical model. In both cases, a shock to government spending lowers consumption and results in multipliers below unity.

In my survey of the literature on multipliers in Ramey (2011b), I found that most estimates of the government spending multiplier in aggregate data were between 0.8 and 1.2. The only multipliers that were larger were (1) those estimated on states or regions; and (2) some of those estimated allowing state dependence. As suggested in my survey, and as shown formally by Nakamura and Steinsson (2014) and Farhi and Werning (2012), the link between estimates of multipliers in a fiscal union (eg, across US states or regions) for aggregate multipliers is not entirely clear. Most of the cross-state or cross-region studies look at the effect of federal spending on a locality. Unfortunately, because constant terms or time fixed effects are included, these regressions difference out the effects of the financing, since taxes that finance federal spending are levied at the national level.[y] This explains why multipliers on federal spending at the state level will be higher than the aggregate multipliers. I will discuss the issue of state dependence in more detail momentarily.

[y] A notable exception is the paper by Clemens and Miran (2012), which identifies exogenous variation in state-level spending. Interestingly, they find multipliers around 0.5, which is closer to those found at the national level.

Since writing that survey, I realized that there were two potential biases in the way that many researchers calculated their multiplier, and as a result, many reported estimates are not comparable. First, many researchers followed Blanchard and Perotti's (2002) lead and calculated multipliers by comparing the *peak* output response to the *initial* government spending impact effect. While comparing values of impulse responses at peaks or troughs is a useful way to compare impulse responses, it is not a good way to calculate a multiplier. As argued by Mountford and Uhlig (2009), Uhlig (2010), and Fisher and Peters (2010), multipliers should instead be calculated as the *integral* (or present value) of the output response divided by the *integral* government spending response. The integral multipliers address the relevant policy question because they measure the cumulative GDP gain relative to the cumulative government spending during a given period. In many cases, Blanchard and Perotti's ratio of peak-to-impact method gives a higher number for the multiplier than the integral method. Second, most researchers estimating VARs use logarithms of variables. To convert the estimates to multipliers, they usually multiply the estimates by the sample mean of GDP to government spending ratio. As Owyang et al. (2013) point out, this can lead to serious biases in samples with significant trends in the GDP to government spending ratio. In the few cases where I have been able to adjust the estimates of multipliers to be integral multipliers, I have found that the multipliers are often below one.

With this additional caveat in mind, Table 3 shows a summary of a few of the estimates of multipliers for government purchases. Even with the variety of ways of calculating multipliers from the estimated impulse response functions, most of the estimates are from 0.6 to 1.5.

A number of researchers and policy makers have suggested that multipliers may be state dependent. Auerbach and Gorodnichenko (2012) use a smooth transition vector autoregression model and find evidence of larger multipliers in recessions. Ramey and Zubairy (2014) use the Jordà (2005) method (also used by Auerbach and Gorodnichenko, 2013 in a panel of countries) and find little evidence of state dependence, based on recessions, elevated unemployment rates, or the zero lower bound. They argue that their different finding is not so much due to the underlying parameter estimates but rather due to the additional assumptions that Auerbach and Gorodnichenko (2012) made when transforming those estimates into multipliers.

Most of the studies I have summarized focus on government purchases that do not involve infrastructure spending. The reason for the paucity of research on infrastructure spending is the difficulty of identifying shocks to infrastructure spending, particularly in the United States. The US highway system was an important part of government purchases starting in the late 1950s through the early 1970s. The problem with identifying the aggregate effects is that most of the spending was anticipated once the highway bill was passed in 1956. Most of the credible analyses have used clever indirect methods or used variation in cross-state expenditures. Fernald (1999) provides very strong evidence for a causal effect of the highway system on productivity by showing its greater effect on

Table 3 Summary of government spending multiplier estimates for the aggregate United States

Study	Sample	Identification	Implied spending multiplier
Barro (1981), Hall (1986), Hall (2009), Barro–Redlick (2011)	Annual historical samples	Use military spending as instrument for government spending	0.6–1
Rotemberg–Woodford (1992)	Quarterly, 1947–1989	Residuals from regression of military spending on own lags and lags of military employment	1.25
Ramey–Shapiro (1998), Edelberg et al. (1999), Burnside et al. (2004)	Quarterly, 1947 to the late 1990s or 2000s	Ramey–Shapiro dates, which are based on narrative evidence of anticipated military buildups	0.6–1.2, depending on sample and whether calculated as cumulative or peak
Blanchard–Perotti (2002)	Quarterly, 1960–1997	SVARS, Cholesky decomposition with *G* ordered first	0.9–1.29, calculated as peak multipliers
Mountford–Uhlig (2009)	Quarterly, 1955–2000	Sign restrictions on an SVAR	0.65 for a deficit-financed increase in spending
Bernstein and Romer (2009)	Quarterly	Average multipliers from FRB/US model and a private forecasting firm model	Rising to 1.57 by the 8th quarter
Cogan et al. (2010)	Quarterly, 1966–2004	Estimated Smets–Wouters model	0.64 at peak
Ramey (2011a,b)	Quarterly, 1939–2008 and subsamples	VAR using shocks to the expected present discounted value of government spending caused by military events, based on narrative evidence	0.6 −1.2, depending on sample
Fisher–Peters (2010)	Quarterly, 1960–2007	VAR using shocks to the excess stock returns of military contractors	1.5 based on cumulative effects
Auerbach and Gorodnichenko (2012)	Quarterly, 1947–2008	SVAR that controls for professional forecasts, Ramey news Key innovation is regime-switching model	Expansion: −0.3 to 0.8 Recession: 1–3.6 (uses a variety of ways to calculate multipliers)
Ben Zeev and Pappa (forthcoming)	Quarterly, 1947–2007	Shock that (i) is orthogonal to current defense spending; and (ii) best explains future movements in defense spending over a horizon of 5 years	2.1 based on integral multiplier at 6 quarters

vehicle-intensive industries. These estimates do not directly inform us about aggregate effects, though. Leduc and Wilson (2013) identify news shocks about highway spending in a panel of US states using the arrival of new information about institutional formula factors. However, as discussed earlier, the multipliers they find cannot be converted to aggregate multipliers.

Gechert (2015) conducts a meta-analysis of 104 studies of multiplier effects across a variety of countries, including many different types of analyses from reduced form empirical to estimated DSGE models. With the caveat that the context and experiment varies across studies, Gechert finds that public spending multipliers are close to one, while public investment multipliers are around 1.5.

4.1.3 Explorations with Several Identified Shocks

I now study the effects of several of the leading identified government spending shocks in the Jordà local projection framework. This exploration is useful not only for gauging the robustness of the results to local projection methods but also for comparing the effects of the identified shocks using the same data, same specification, and the same way to calculate multipliers. Thus, any differences in results will be due to the identification methods rather than differences in data or implementation.

The three main shocks I study are (i) the shock identified using Blanchard and Perotti's (2002) method (which simply orders government spending first in a Cholesky decomposition); (ii) my narrative military news shock, updated in Ramey and Zubairy (2014); and (iii) Ben Zeev and Pappa's (forthcoming) defense news shock identified using Barsky and Sims' (2012) medium-run horizon method.[z] I also comment briefly on results using Fisher and Peters' (2010) military contractor excess returns.

In all cases, I use the following data transformations and functional forms. The first set of transformations is intended to facilitate the direct estimation of multipliers in order to avoid ad hoc transformation of estimates based on logs, as discussed by Owyang et al. (2013). One can use either the Hall (2009) and Barro–Redlick (2011) transformation or a Gordon–Krenn (2010) transformation. The Hall–Barro–Redlick transformation constructs variables as $(X_{t+h}-X_{t-1})/Y_{t-1}$, where X is the NIPA variable deflated by the GDP deflator and Y_{t-1} is real GDP before the shock hits in period t. The Gordon–Krenn transformation divides all NIPA variables by "potential GDP," estimated as an exponential trend. Both methods give similar results. I follow the Gordon–Krenn procedure, fitting log real GDP to a quadratic trend.[aa] Thus, the NIPA variables are

[z] I estimated the Blanchard–Perotti shock using logarithms of real government spending, GDP, and taxes and four lags. One could instead estimate it directly in the regression using the Gordon-Krenn transformed variables. Ben Zeev and Pappa kindly provided me with estimates of their shock.

[aa] One could use the CBO estimate of real potential GDP instead. I found, however, that when I used the CBO estimate, the implied multipliers were noticeably smaller than when I used Hall–Barro–Redlick or Gordon–Krenn with either a quadratic or a quartic log trend.

transformed to be $z_t = X_t / Y_t^*$, where Y_t^* is the estimated trend in real GDP. The impulse responses using this transformation look qualitatively similar to those using log levels, but often have more narrow confidence bands.

The non-NIPA variables are transformed as follows. The average tax rate is federal current receipts divided by nominal GDP. The hours variable is the log of total hours per capita, where the total population is used in the denominator. Wages are given by the log of nominal compensation in the business sector, deflated by the price deflator for private business. The real interest rate is the 3-month Treasury bill rate minus the rate of inflation calculated using the GDP deflator.

The equation used to estimate the impulse responses for each variable z at each horizon h is given by

$$z_{t+h} = \alpha_h + \theta_h \cdot \text{shock}_t + \varphi_h(L) y_{t-1} + \text{quadratic trend} + \varepsilon_{t+h} \tag{18}$$

where z is the variable of interest, *shock* is the identified shock, y is a vector of control variables, and $\varphi_h(L)$ is a polynomial in the lag operator. All regressions include two lags of the shock (to mop up any serial correlation), transformed real GDP, transformed real government purchases, and the tax rate. Regressions for variables other than real GDP, government purchases, and tax rates also include two lags of the left-hand side variable. The coefficient θ_h gives the response of z at time $t+h$ to the shock at time t.

As discussed earlier, the correct way to calculate a multiplier is as the integral under the impulse response of GDP divided by the integral of the impulse response of government spending. We could compute the multiplier using the following multistep method: (1) estimate Eq. (18) for GDP for each horizon and sum the coefficients θ_h up to some horizon H; (2) estimate Eq. (18) for government purchases for each horizon and sum the coefficients θ_h up to some horizon H; and (3) construct the multiplier as the answer from step (1) divided by the answer from step (2). Estimating the standard error, however, requires some ingenuity, such as estimating all of the regressions jointly in a panel estimation.

Alternatively, we can easily estimate the multiplier estimate and its standard error in one step if we cumulate the variables and reformulate the estimation problem as an instrumental variables (IV) estimation. In particular, we can estimate the following equation:

$$\sum_{i=0}^{h} z_{t+i} = \beta_h + m_h \cdot \sum_{i=0}^{h} g_{t+i} + \chi_h(L) y_{t-1} + \text{quadratic trend} + \nu_{t+h} \tag{19}$$

where the dependent variable is the *sum* of real GDP (or other NIPA variable) from t to $t+h$ and the government spending variable is the *sum* of the government purchases variable. We use the identified shock as the *instrument* for the sum of government purchases. The estimated coefficient, m_h, is the multiplier for horizon h. There are several advantages to this one-step IV method. First, the standard error of the multiplier is just the standard

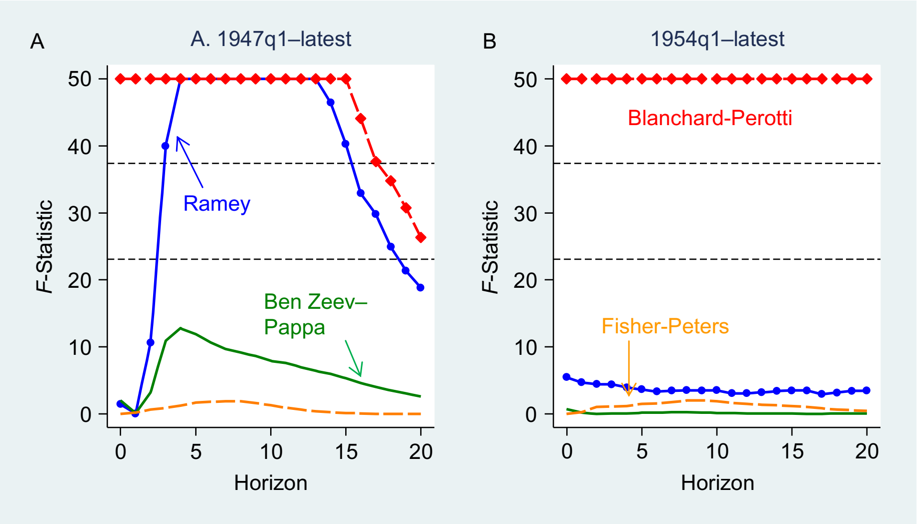


Fig. 4 First-stage *F*-statistics for government spending shocks. Note: The *F*-statistics are based on the regression of the sum of government spending from *t* to *t*+*h* on the shock at *t*, plus the lagged control variables. Values above 50 have been capped at 50. The *horizontal dashed lines* are the Montiel Olea and Pflueger (2013) 5% worst case bias (*upper line*) and 10% worst case bias (*lower line*) thresholds.

error of the coefficient m_h.[bb] Second, the shock can have measurement error as long as the measurement error is not correlated with any measurement error in government spending. Third, formulating the estimation as an IV problem highlights the importance of instrument relevance.

Thus, I first consider how relevant each of the identified shocks is as an instrument for the integral of government spending. Stock et al. (2002) argue that the first-stage *F*-statistic should be above 10 for the IV estimates to be reliable, but their threshold applies only to first-stage regressions with serially uncorrelated error terms. Fortunately, follow-up work by Montiel Olea and Pflueger (2013) constructs thresholds for cases with serial correlation. For the first stage of Eq. (19), the thresholds are 23 for the 10% level and 37 for the 5% level.[cc] *F*-Statistics below those thresholds indicate a possible problem with instrument relevance.

Fig. 4 shows the first-stage *F*-statistics for the sum of government purchases on each identified shock for each horizon up to 20 quarters. Values above 50 have been capped at 50 for ease of viewing. The graph at the left shows the results for the sample starting in 1947 and the graph on the right shows the results for the sample starting in 1954, after the

[bb] Because of the serial correlation in the errors, any procedures for estimating standard errors should use methods that account for serial correlation.

[cc] These F-statistics and thresholds were derived using Pflueger and Wang's (2015) "weakivtest" Stata module.

Korean War. Several results emerge. First, the Blanchard–Perotti (BP) identified shock always has very high *F*-statistics. This is not surprising because the shock is equal to the portion of government spending not predicted by four lags of government spending, GDP, and taxes. Second, the Fisher–Peters defense contractor excess returns shock has very low *F*-statistics for all horizons and both samples, indicating that the return variable is not a good instrument for government spending. Third, the Ramey and Ben Zeev–Pappa (BZP) news shocks have low relevance for short horizons, but this is to be expected since those shocks capture news about *future* government spending. Fourth, in the full sample at horizons beyond three quarters, the Ramey news shock has *F*-statistics above the Montiel–Pflueger thresholds, whereas the BZP news shock *F*-statistics lie below them and range between 8 and 13. Fifth, in the sample that excludes the Korean War, all of the *F*-statistics are very low except for the Blanchard–Perotti shock. Thus, the BP shock surpasses the relevance safety threshold for all horizons in both samples, the Ramey news shock does so for the full sample for horizons at four to 20 quarters, while the BZP shock may have relevance problems at most horizons and the Fisher–Peters shock always has very low relevance. I thus exclude the Fisher–Peters shock from the rest of the analysis.

Fig. 5 shows the impulse responses estimated using Eq. (18), with estimates normalized across specifications to have the same peak in government purchases. The scales in the graphs of the NIPA variables should be interpreted as dollars; ie, a rise in government purchases that peaks at $1 leads to rise in GDP that peaks at 75 cents. The scales of the other graphs are in percentage points. The confidence bands are 90% bands based on Newey–West corrections of standard errors. They do not, however, take into account that two of the shocks are generated regressors.

Consider first the upper left graph in Fig. 5. Both the Ramey and BZP news variables imply similar paths of government purchases, with little effect for the first few quarters rising to a peak around five quarters after the shock. In contrast, the BP shock leads to an immediate rise in government spending. The graph in the top right shows that in response to all three shocks, GDP jumps immediately. The response of GDP is greatest for the BZP shock, but GDP begins to fall back to normal even before government purchases have hit their peak.

The tax rate series is simply federal receipts divided by GDP. This variable can rise either because of tax legislation or because higher GDP pushes more households into higher tax brackets. According to the estimates, tax rates begin to rise immediately for the BZP shock but only gradually for the Ramey news shock. Tax rates gradually fall after BP shock. Real interest rates (measured as the 3-month T-bill rate minus inflation) fall after a news shock, but rise slightly after a BP shock. Examination of the responses of the components of the real interest rate (not shown) reveals that the fall is due to both a drop in the nominal interest rate and a rise in inflation. As explained by Ramey (2011a), the rise in inflation is in large part driven by the spike up in prices at the beginning of the

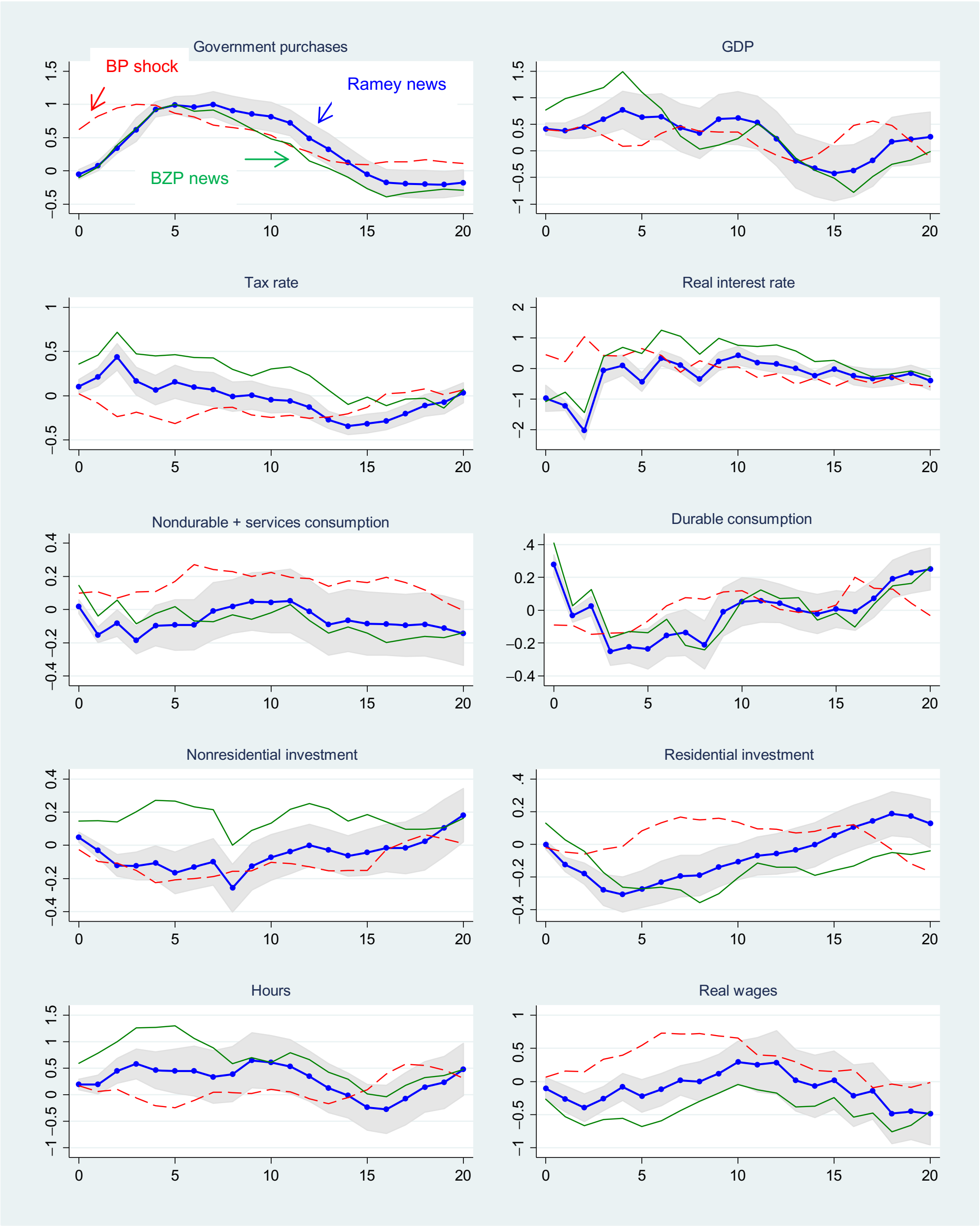


Fig. 5 Comparison of the effects of government spending shocks (*BP*: Blanchard–Perotti; *BZP*: Ben Zeev–Pappa). Light gray bands are 90% confidence bands.

Table 4 Multiplier estimates (HAC standard errors in parenthesis)

Horizon (in quarters)	Blanchard–Perotti	Ramey news	Ben Zeev–Pappa news
0	0.65 (0.24)	−7.53 (7.26)	−7.37 (5.85)
4	0.37 (0.23)	1.37 (0.33)	2.91 (1.13)
8	0.39 (0.32)	0.80 (0.25)	1.41 (0.61)
12	0.39 (0.44)	0.77 (0.27)	1.24 (0.71)
16	0.40 (0.58)	0.60 (0.36)	1.10 (1.01)
20	0.44 (0.63)	0.69 (0.48)	1.17 (1.46)

Notes: Multipliers estimated using Eq. (19). All regressions also include two lags of the shock (to mop up any serial correlation), real GDP (divided by potential GDP), real government purchases (divided by potential GDP), the tax rate, and a quadratic trend.

Korean War: with recent memories of WWII, firms thought that price controls were coming and raised their prices in advance.

Consider now four components of the national income accounts, shown in the middle graphs in Fig. 5. Nondurable plus services consumption falls after a Ramey news shock, responds little after a BZP shock, but rises after a BP shock. In Ramey (2009), I show using simulations of a DSGE model that one can estimate a rise in consumption if one treats an anticipated shock as an unanticipated shock. I argue that the rise in consumption after a BP shock can be explained by this type of identification problem.

Durable consumption spikes up initially and then falls after the two news shocks. As in the cast of prices, this initial spike on the arrival of news is driven mostly by the response of consumers to the beginning of the Korean War in 1950: with recent memories of WWII, consumers worried that rationing of durable goods was imminent so they hurried out to buy durable goods. Nonresidential investment rises in response to the BZP shock, but falls in response to both the Ramey news and BP shock. Residential investment falls in response to the two news shocks, but rises after a year in response to the BP shock.

Finally, both news shocks imply a rise in hours and a fall in the real wage, while the BP shock predicts very little response of hours, but a rise in the real wage.

Table 4 shows the estimated multipliers for various horizons.[dd] The impact multipliers for the two news shocks are negative because output jumps up, while government spending falls slightly. For the next few quarters, the multipliers for the two news shocks are large because output responds immediately to news of future government spending which has not yet fully transpired. Once government spending has risen to its peak, the implied multipliers using the Ramey news shock are just below unity, whereas those using the BZP news shock are above unity. For example, the BZP news shock multiplier is 1.4 at 8 quarters and 1.1 at 16 quarters. The responses in Fig. 5 reveal that the reason for the larger multiplier after a BZP shock is the large rise in nonresidential investment.

[dd] These estimates are based on the one-step method shown in Eq. (19).

It should be noted, though, that the BZP multipliers are estimated rather imprecisely as evidenced by the standard errors. This is one manifestation of the possible low instrument relevance of the BZP news shock. On the other hand, the multipliers implied by the Blanchard and Perotti shock are all low, most below 0.5. However, the estimates are not precise enough to reject a multiplier of either 0 or 1 at standard significance levels.

I now consider what each shock implies about the contribution to the forecast error variance of output. Although one can calculate forecast error variances using the estimated local projection coefficients, I found that the shares sometimes exceeded 100%. Thus, for present purposes I calculate the forecast error variance in a standard VAR with the shock, log government spending, log real GDP, and log taxes. The shock is ordered first and four lags are included, along with a quadratic trend.

Table 5 shows the forecast error variance decompositions of each of the three identified government spending shocks for government spending and output. Despite having the lowest contribution for government spending, the BZP shock has the highest contribution for output, but it is still 13% or below. The BP and Ramey news contributions tend to be 5% of below. Thus, none of the three shocks is an important contributor to the variance of output.

To summarize, most of the aggregate analyses find government spending multipliers between 0.6 and 1.5. The BP shock leads to smaller multipliers, but does imply that government spending leads to rises in consumption and real wages along with GDP and hours. In contrast, both the Ramey news and BZP news shocks lead to falls in real wages. Both news shock lead to an initial spike in durable consumption (due to the consumer fears of rationing), followed by a decline. The BZP shock produces a temporary blip in nondurable consumption, but then it falls to 0. The Ramey news shock implies a decline nondurable consumption. None of the methods suggests that government spending shocks explain an important part of GDP fluctuations.

Table 5 Shock contribution to the forecast error variance of government spending and output

	Blanchard–Perotti		Ramey news		Ben Zeev–Pappa news	
Horizon (in quarters)	Government spending	Output	Government spending	Output	Government spending	Output
0	100.0	5.5	1.0	2.2	1.4	5.6
4	96.2	3.3	31.8	2.6	14.0	10.1
8	90.5	2.9	50.2	2.9	27.0	12.6
12	86.5	2.5	50.5	2.5	29.8	12.1
16	83.1	2.4	46.7	2.4	29.4	11.8
20	80.2	2.3	43.0	2.2	28.7	11.7

Notes: Based on standard VAR with the shock, log output, log government spending, log taxes, and a quadratic trend. The shock is ordered first and four lags of the variables are included.

4.2 Tax Shocks

I now turn to the literature on tax shocks. Taxes were often an important component of the big Keynesian econometric models of the 1960s. The public finance literature has analyzed many details of the effects of taxes. In this section, I will focus on estimates of the effects of taxes in the macroeconomic literature since the 1990s.

4.2.1 Unanticipated Tax Shocks

4.2.1.1 Summary of the Literature

Macroeconomists have used both estimated DSGE models and time series models to estimate the effects of taxes. One of the first systematic analyses of macroeconomic tax effects in an estimated DSGE model was by McGrattan (1994). She extended the Kydland and Prescott (1982) model to include government consumption, labor income taxes, and capital income taxes and estimated the parameters using maximum likelihood. She found that the role of technology in business cycle fluctuations was much reduced, 41% rather than Kydland and Prescott's 75% estimate. She found that shocks to government consumption accounted for 28% of the forecast error variance of output, labor income tax shocks for 27%, and capital income tax shocks for 4%.

Among time series approaches, Blanchard and Perotti (2002) used an SVAR approach in which they identified tax shocks by imposing the elasticity of net taxes to GDP estimated from other studies. Returning to the discussion of the simple trivariate model from Section 2, consider the following system:

$$\begin{aligned}\eta_{1t} &= b_{12}\eta_{2t} + b_{13}\eta_{3t} + \varepsilon_{1t}\\ \eta_{2t} &= b_{21}\eta_{1t} + b_{23}\eta_{3t} + \varepsilon_{2t}\\ \eta_{3t} &= b_{31}\eta_{1t} + b_{32}\eta_{2t} + \varepsilon_{3t}\end{aligned} \tag{20}$$

where η_{1t} is the reduced form residual of net taxes, η_{2t} is the reduced form residual of government spending, and η_{3t} is the reduced form residual of GDP. Blanchard and Perotti (2002) identify the shock to government spending using a Cholesky decomposition in which government spending is ordered first (ie, $b_{21} = b_{23} = 0$). They identify exogenous shocks to net taxes by setting $b_{13} = 2.08$, an outside estimate of the cyclical sensitivity of net taxes. These three restrictions are sufficient to identify all of the remaining parameters and hence all three shocks. Blanchard and Perotti's estimated "impact multiplier" was -0.78. Their impact multiplier was calculated as the *trough* of GDP relative to the *initial* shock to taxes.

Mountford and Uhlig (2009) use sign restrictions to identify tax and spending shocks. Their results imply a multiplier of -5 at 12 quarters for a deficit-financed tax cut, when the multiplier is calculated as the ratio of the present value of the impulse response functions. In order to compare their results to Blanchard and Perotti, they also calculate "impact multipliers," meaning the value of the GDP response at a certain quarter to the initial shock impact on the fiscal variable. They find that whereas the Blanchard and Perotti method implies a peak-to-impact multiplier of −1.3 at quarter 7, Mountford and Uhlig's results imply a peak-to-impact multiplier of −3.6.

In the context of the Blanchard and Perotti (2002) model, Caldara and Kamps (2012) demonstrate how the estimated multiplier depends crucially on their assumption about the elasticity of net tax revenue to GDP. Particularly important is their demonstration of how a small change in the assumed cyclical elasticity parameter can result in large changes in the estimated tax multiplier; to wit, this seems to be a case of a "multiplier multiplier" on the assumed elasticity. Caldara and Kamps (2012) propose a new method, which involves imposing probability restrictions on the output elasticities of taxes and spending. When they implement this method, they find peak-to-impact multipliers of −0.65 for tax shocks and peak-to-impact multipliers greater than unity for government spending shocks.

Barro and Redlick (2011) construct a new series of average marginal tax rates using IRS data and analyze its effects in a system that also considers government spending in annual data extending back to 1917. In their baseline specification, they find that an increase in the average marginal tax rate of 1 percentage point lowers GDP by 0.5%. Their calculations indicate a tax multiplier of −1.1.

Romer and Romer (2010) (R&R) use narrative methods to identify tax shocks. Based on presidential speeches and congressional reports, they construct several series of legislated tax changes and distinguish those series based on the motivation for enacting them. They argue that tax changes motivated by a desire to pay down the deficit or long-run growth considerations can be used to establish the causal effect of tax changes on output. When they estimate their standard dynamic single equation regression of output growth on its lags and on current and lagged values of the "exogenous" tax changes, they obtain estimates implying tax multipliers of −2.5 to −3 at 3 years. Leigh et al. (2010) use a similar narrative method to study fiscal consolidations across countries.[ee] Cloyne (2013) uses this method to identify exogenous tax shocks in the United Kingdom.

Favero and Giavazzi (2012) embed the R&R series in a somewhat restricted VAR and find smaller multipliers. In a series of papers, Mertens and Ravn (2011b, 2012, 2013, 2014) exploit the R&R narrative tax information in a way that significantly expands our understanding of the effects of tax shocks on the economy. I will focus on several of their contributions in this section and discuss the others in the next section. First, Mertens and Ravn (2011b, 2012) split the Romer and Romer series into anticipated vs unanticipated shocks based on the delay between the passing of the legislation and the implementation of the legislation. R&R had timed all of their shocks to coincide with the implementation rather the legislation. I will discuss the findings using unanticipated shocks here and discuss the findings using anticipated shocks below. Second, in their 2013

[ee] The Leigh et al. attempts to address measurement concerns in the very large literature on the effects of fiscal consolidations across countries, perhaps best exemplified by Giavazzi and Pagano (1990, 1996), Alesina and Perotti (1995, 1997), and Alesina and Ardagna (1998, 2010).

paper, Mertens and Ravn (2013) decomposed the unanticipated parts of the R&R series into personal income tax changes and corporate income tax changes and showed the differences in the two types of cuts on the economy. In their 2014 paper, Mertens and Ravn (2014) reconciled the Blanchard and Perotti SVAR estimates with the narrative estimates by introducing the proxy SVAR method discussed in detail in previous sections.

As discussed in Section 2.3.5, Mertens and Ravn's (2014) proxy SVAR provides a new method for identifying shocks using external instruments. In particular, they regress the reduced form residual of GDP, η_{3t}, from Eq. (20) on the reduced form residual of taxes, η_{1t}, using the R&R shock as an instrument. This leads to an unbiased estimate of b_{31} (since it is assumed that η_{2t} is the structural shock to government spending, which is uncorrelated with the other structural shocks). We can then use the estimated residual from that regression as one of the instruments in the regression of η_{1t} on η_{2t} and η_{3t}. This regression identifies b_{12} and b_{13}. When they implement their method, they estimate $b_{13} = 3.13$ with a 95% confidence band of (2.73, 3.55). Thus, their results suggest that Blanchard and Perotti's preset estimate of $b_{13} = 2.08$ is too low. Setting the output elasticity of tax revenue too low results in estimated tax shocks that include a reverse causality component (ie, there is a positive correlation between the cyclical components of taxes and output because of the positive causal effect of output on tax revenues). This is also an excellent illustration of Caldara and Kamps' (2012) insight on the link between the assumed structural tax elasticity and the estimated multipliers. Table 6 shows various tax multiplier estimates from the literature.

Mertens and Ravn (2013) split the unanticipated Romer shocks into changes in personal income tax rates vs corporate income tax rates. They find that cuts in either tax rate have positive effects on output, employment, hours, and the tax base. Interestingly, a cut in the corporate tax rate does not decrease corporate tax revenues because the corporate income tax base responds so robustly. Personal income tax cuts tend to raise consumption and investment more than corporate income tax cuts do. See figures 2, 9, and 10 of Mertens and Ravn (2013) for more detail.

Oh and Reiss (2012) highlight the importance of transfers in the stimulus packages adopted in response to the Great Recession. They formulate a heterogeneous agent model and explore the predicted multipliers on transfers. There has been, however, very little empirical work on the multipliers associated with government transfers.[ff] A major challenge has been identifying exogenous movements in transfers. Hausman (2016) studied the large veteran's bonus of 1936, equaling 2% of GDP, and found that it led to immediate effects on consumption spending. His calculations suggest that it led to faster GDP growth in 1936, but followed by a quick reversal in growth in 1937. Romer and Romer (2016) study the effects of changes in Social Security benefit payments in aggregate US

[ff] There is a large literature on the effects of various transfers on individual household consumption and saving. However, these estimates do not translate directly to aggregate multipliers.

Table 6 Summary of some tax multiplier estimates for the aggregate Unites States

Study	Main sample	Identification	Implied tax multiplier
Evans (1969)	Quarterly, 1966–1974	Based on estimates of equations of Wharton, Klein–Goldberger, and Brookings models	−0.5 to −1.7, depending on horizon, type of tax, and model
Blanchard–Perotti (2002)	Quarterly, 1960–1997	Assumed output elasticities in an SVAR. "Taxes" are actually taxes less transfers	−0.78 to −1.33 (peak to impact)
Mountford–Uhlig (2009)	Quarterly, 1955–2000	Sign restrictions on a VAR. Use same variables as BP	−5 for a tax increase that reduces the deficit
Romer–Romer (2010)	Quarterly, 1947–2007	Legislated tax changes driven by an inherited government budget deficit or to promote future growth, based on narrative evidence	−3, based on peak effect. Romer–Romer (2009) show that these tax shocks do not raise government spending significantly, so these are close to pure tax shocks
Barro–Redlick (2011)	Annual, 1917–2006 and subsamples	Average marginal income tax rate	−1.1
Favero and Giavazzi (2012)	Quarterly, 1950–2006	Romer–Romer shocks embedded in an SVAR	−0.5
Caldara and Kamps (2012)	Quarterly, 1947–2006	SVAR using outside elasticities	−0.65 (peak to impact)
Mertens–Ravn (2014)	Quarterly, 1950–2006	Proxy SVAR using Romer–Romer unanticipated shocks	−3 at 6 quarters

data. They find very rapid responses of consumption to permanent changes in benefit, but the results dissipate within a few months. Moreover, there is no clear evidence of effects on aggregate output or employment.

Gechert (2015) conducts a meta-analysis of various types of multipliers. He finds that tax and transfer multiplier tend to be around 0.6–0.7.

4.2.1.2 Further Explorations

I now investigate the Mertens and Ravn (2014) reconciliation of the tax results in more detail. To do this, I first use Mertens and Ravn's (2014) specification, data, and sample. The specification is a trivariate SVAR with federal government spending, output, and federal tax revenue, all in real per capita logarithms.[gg] The SVAR includes four lags of the variables in addition to a quadratic trend and a dummy variable for the second quarter of 1975 (following Blanchard and Perotti, 2002). The tax shock is Mertens and Ravn's unanticipated shocks extracted from the R&R narrative, demeaned as they describe.

Fig. 6A shows the impulse responses for tax revenue and output from their proxy SVAR using their programs.[hh] The results show that a positive R&R tax shock that has an impact effect on tax revenues equal to 1% of GDP raises tax revenue for several quarters and then lowers it below zero (though not statistically different). Output falls significantly on impact and troughs around -3 after a year. The magnitude of the results is similar to those found by R&R (2010) with their entire exogenous series.

My further investigation reveals some potentially problems with instrument relevance, though. The first-stage regression of tax revenue on the unanticipated tax shock (controlling for the lags of the other variables in the VAR) has an *F*-statistic of 1.6 (based on robust standard errors), which suggests a possible problem with instrument relevance.[ii] Stock and Watson (2012) also noticed problems with first-stage *F*-statistics of some of these external instruments in their dynamic factor model.[jj] Of course, the feedback from GDP to tax revenues is a potential complication. An exogenous tax increase should raise revenue, holding GDP constant; however, the decline in GDP exerts a downward effect on tax revenues. Thus, perhaps it is better to think of the R&R tax shock as an instrument for tax *rates*. Ideally, one would use statutory rates, since the actual rate paid is partly endogenous (since a change in income can push an entity into a different tax bracket).

[gg] Blanchard and Perotti actually used *net* taxes, meaning taxes less transfers. I follow Mertens and Ravn and use taxes. One could augment the system to include transfers as a fourth variable and use Romer and Romer's (2014) narrative-based transfer shock series as an external instrument.

[hh] This is the same as Mertens and Ravn's (2014) figure 4 with the signs reversed to examine the effect of a tax increase.

[ii] These additional results are based on the same data definitions and specification as Mertens and Ravn (2014), but on revised data. The results are similar if I use their original data.

[jj] See Lundsford (2015) and Montiel Olea et al. (2015) for discussions of instrument relevance in the external instruments framework.

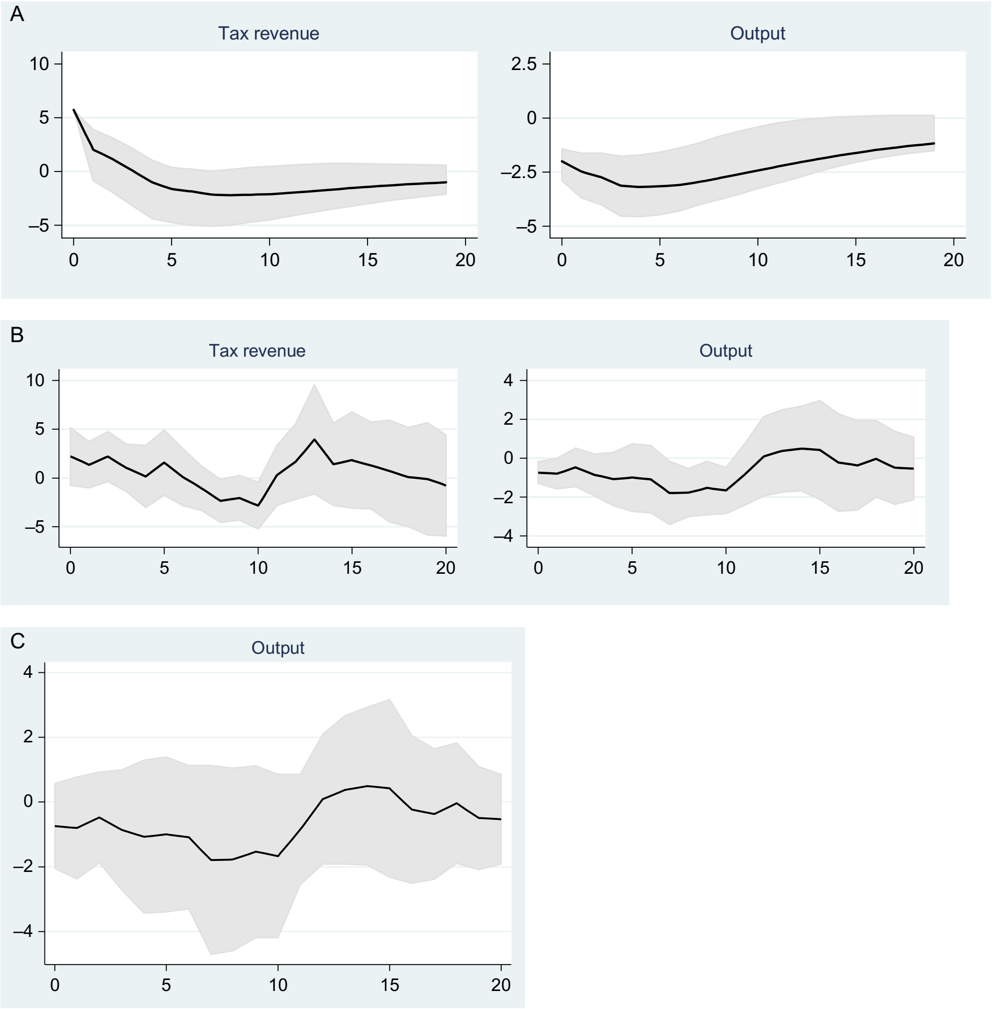


Fig. 6 Effects of unanticipated Romer tax shock, trivariate VAR, 1950q1–2006q4. (A) Mertens–Ravn (2014) proxy SVAR. (B) Jordà local projection, reduced form. (C) Jordà local projection, IV regression of output on tax revenue. Light gray bands are 90% confidence bands.

I do not have those data, so I simply construct an average tax rate as federal tax revenues divided by nominal GDP. I then estimate the first-stage regression described earlier with the average tax rate substituted for the log of taxes. The *F*-statistic on the R&R shock in this regression is 3.2, twice as high as the previous case but still well below the threshold for instrument relevance.

With the caveats about instrument relevance in mind, I further explore the robustness of Mertens and Ravn's (2014) results by estimating impulse responses using the Jordà local projection method and the Romer tax shock. I first estimate the reduced forms.

Table 7 Tax shock contribution to the forecast error variance of output

Horizon (in quarters)	Romer-Romer unanticipated	Leeper et al. (2012) anticipated tax series
0	1.6	0.4
4	0.4	5.7
8	0.5	4.8
12	1.1	4.4
16	1.8	4.3
20	2.1	4.3

Notes: Based on standard VAR with the shock, log output, log government spending, log taxes, and a quadratic trend. The shock is ordered first and four lags of the variables are included.

As discussed earlier, this involves regressing the dependent variable at $t+h$ on the shock at t, controlling for lags of other variables. To be consistent with Mertens and Ravn's specification, I use the same lags and variables in their proxy SVAR. Fig. 6B shows the impulse responses from the reduced form. Tax revenue increases in response to the shock initially and then falls below normal. In response to the tax shock, output falls on impact and then declines further to about −2 at 2 years, before beginning to recover. The confidence bands are wider, both because the Jordà method imposes fewer restrictions on the dynamics and because they incorporate the uncertainty about the impact of the tax shock on tax revenue. However, the point estimates for output for the first few years are broadly consistent with both Romer and Romer's (2010) original results and Mertens and Ravn's (2014) proxy SVAR.[kk] Table 7 shows the forecast error variance decomposition based on a standard VAR.[ll] Unanticipated tax shocks appear to account for very little of the forecast error variance of output.

As Mertens and Ravn (2014) note, however, external instruments tend to have measurement error, so they should not be used directly in an SVAR. A natural way to adjust for this in the Jordà setup is to estimate things as an IV regression, as discussed in Section 2. Thus, in a second specification I regress output at $t+h$ on the change in tax revenue at t, instrumented with the unanticipated part of the Romer tax shock, also controlling for the same variables as in the proxy SVAR (four lags of output, tax revenue, and government spending, as well as the deterministic terms). Fig. 6C shows the estimated impulse response of output for this specification. The point estimates for these results look very similar to those for output in Fig. 6B. The difference is that the confidence intervals are

[kk] If I use the Jordà method on the Romer's original specification and tax shock, I obtain results that are very close to theirs. This is as one would expect since they do not calculate impulses from a VAR.

[ll] As discussed in Section 4.1.3, although one can calculate forecast error variances using the estimated local projection coefficients, I found that the shares sometimes exceeded 100%. Thus, for present purposes I calculate the forecast error variance in a standard VAR with the shock, log government spending, log real GDP, and log taxes. The shock is ordered first and four lags are included, along with a quadratic trend.

very wide, always encompassing zero. Moreover, when I test whether the integral of the response for the first 12 quarters is different from zero, I cannot reject that it is zero.[mm]

To summarize, the literature on the effects of tax shocks has employed numerous methods, such as SVARs with calibrated elasticities, narrative approaches, and sign restrictions. Mertens and Ravn's (2014) reconciliation of some of the various approaches tends to support Romer and Romer's (2010) large estimated elasticities. My robustness checks suggest that while there might be a problem with instrument relevance, less restrictive ways to estimate impulse responses also generally support Romer and Romer's (2010) estimates of tax multipliers of −2 to −3.

4.2.2 News About Future Tax Changes

4.2.2.1 Summary of the Literature

Theory predicts that anticipated tax changes should have very different effects from unanticipated tax shocks (eg, Yang, 2005). If agents know that tax rates will increase in the future, they should respond by intertemporally substituting taxable activity into the present. Moreover, as discussed in Section 2, foresight about future tax changes can lead to identification problems in a standard SVAR. I will now review some recent results on the effects of anticipated tax changes on aggregate outcomes and provide some new results.

Mertens and Ravn (2011b, 2012) explore the effects of anticipated tax changes by splitting the Romers' narrative tax shock series into anticipated vs unanticipated shocks based on the delay between the passing of the legislation and the implementation of the legislation. Romer and Romer had timed all of their shocks to coincide with the implementation rather the legislation. Mertens and Ravn argue that the response of macroeconomic variables should be very different for anticipated vs unanticipated shocks.

Mertens and Ravn separate out the tax changes that were legislated more than 90 days before they were implemented. Because there are not a large number of these kinds of tax changes and because the lags between legislation and implementation vary significantly, Mertens and Ravn preserve the degrees of freedom in their estimation by combining various anticipated tax shocks according to the number of quarters left before implementation. Thus, their study does not trace out the effect of "news" per se; rather, it is more similar to an event study of the behavior of variables before and after the tax changes are implemented. Mertens and Ravn (2011b, 2012) estimate that anticipated and unanticipated tax shocks together account for 20% of the historical variation in output at business cycle frequencies. Particularly interesting is their finding that the so-called Volcker recession was in fact mostly caused by the Reagan tax cuts. The legislation was passed in summer 1981, but the actual tax cuts were phased in between 1982 and 1984. Mertens and Ravn's estimates imply that most of the decline in output from the second half of 1981 through 1982 was due to the negative output effects of anticipated future tax cuts.

[mm] Reducing the number of lags or control variables changes the results little.

Leeper, Richter, and Walker (2012) (LRW) construct an alternative measure of expected tax changes based on the spread between federal bonds and municipal bonds. They use their new series to inform their theoretical model but do not estimate effects of shocks to their series directly from the data. In the unpublished supplement to their 2013 *Econometrica* paper, Leeper et al. (2013) investigate the effect of their measure on output and show that expectations of a future tax increase raise output when the news arrives.

4.2.2.2 Further Explorations

I now explore the effects of several of the leading identified tax news shocks. Fig. 7 reproduces Mertens and Ravn (2011a,b) estimates of the effects of Romer tax shocks that were anticipated. Quarter 0 is the date of the implementation, negative quarters are quarters

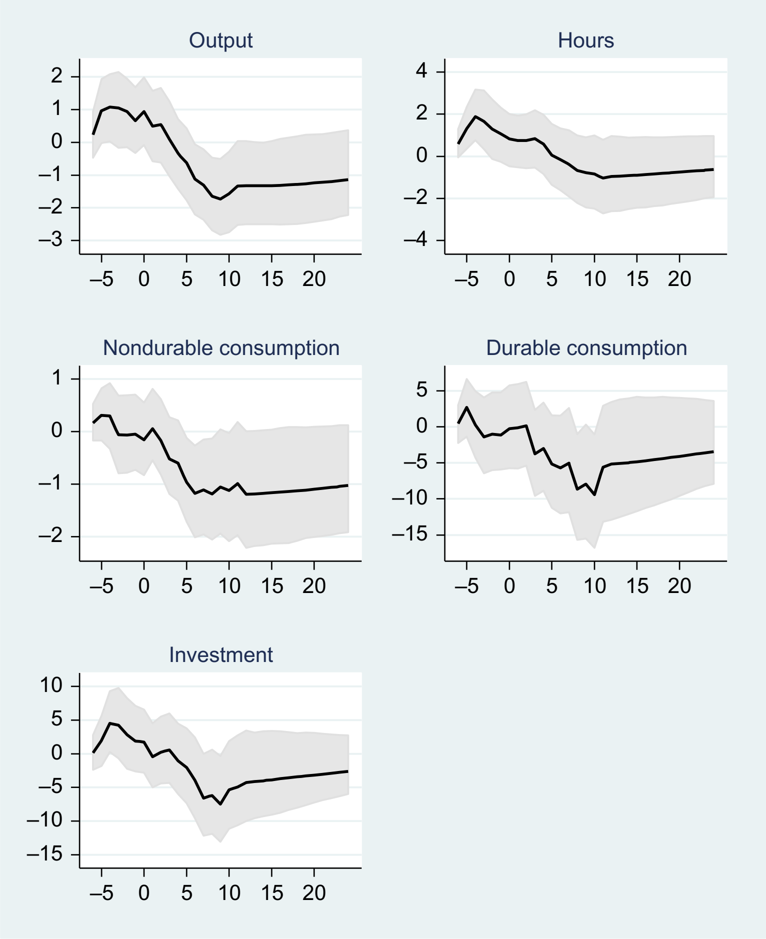


Fig. 7 Effects of news of future tax increases, Mertens–Ravn estimates based on Romer–Romer narrative, 1950q1–2006q4. Light gray bands are 90% confidence bands.

between the arrival of the news and before the implementation, and positive quarters are after the implementation. The graphs show clear evidence of anticipation effects and intertemporal substitution. Most variables, including output, hours, investment, and durable goods consumption expenditures, are higher than average in the interval between the announcement of a tax increase and its actual implementation. After implementation, all variables fall below normal, including nondurable consumption. Thus, the behavior of the data is very consistent with the theory.

To see how the results compare to Mertens and Ravn's results, I analyze the effects of Leeper et al.'s (2012) measure of average expected future tax rates from 1 to 5 years forward (AFTR15). Using a Jordà local projection, I estimate several sets of regressions at each horizon. In particular, I regress the endogenous variable of interest at $t+h$ on AFTR15 in period t, as well as on four lags of AFTR15, four lags of the endogenous variable and four lags of the average federal tax rate (total federal receipts divided by GDP). Because I do not orthogonalize the shock with respect to current values of any of the other variables, this identification scheme is similar to the one used by Leeper et al. (2013), where they order the tax news first in the Cholesky decomposition.

Fig. 8 shows the estimated responses to "news" that future tax rates will rise. The results are quite similar to those of Mertens and Ravn's results, even though the tax news variable is from a completely difference source and the model is estimated as responses to news rather than as an event study around the implementation. Output, hours, and investment start rising when news arrives at period 0 that tax rates will increase in the interval between 1 and 5 years. The variables remain high for a while and then fall below normal after a year or so.

Table 7 shows the forecast error variance decomposition for the LRW measure of expected tax changes. These shocks appear to account for more of the variance of output than the unanticipated tax changes, but still less than 6%.

In sum, some of the strongest and most robust findings in the fiscal literature are those associated with news about future tax changes. Expectations that future tax rates will increase lead to boom now followed by "busts." This is perhaps some of the strongest evidence that "news" can drive economic fluctuations.

4.3 Summary of Fiscal Results

In this section, I have summarized some of the main methods and findings concerning the effects of fiscal shocks. For both government spending and taxes, the methods that use external narrative series tend to find bigger effects on output than the more traditional SVAR method. For both government spending and taxes, anticipation effects are found to be very important.

Some of the literature has studied the effects of government spending and tax shocks jointly and made statements about "which" multiplier is larger. Some find larger

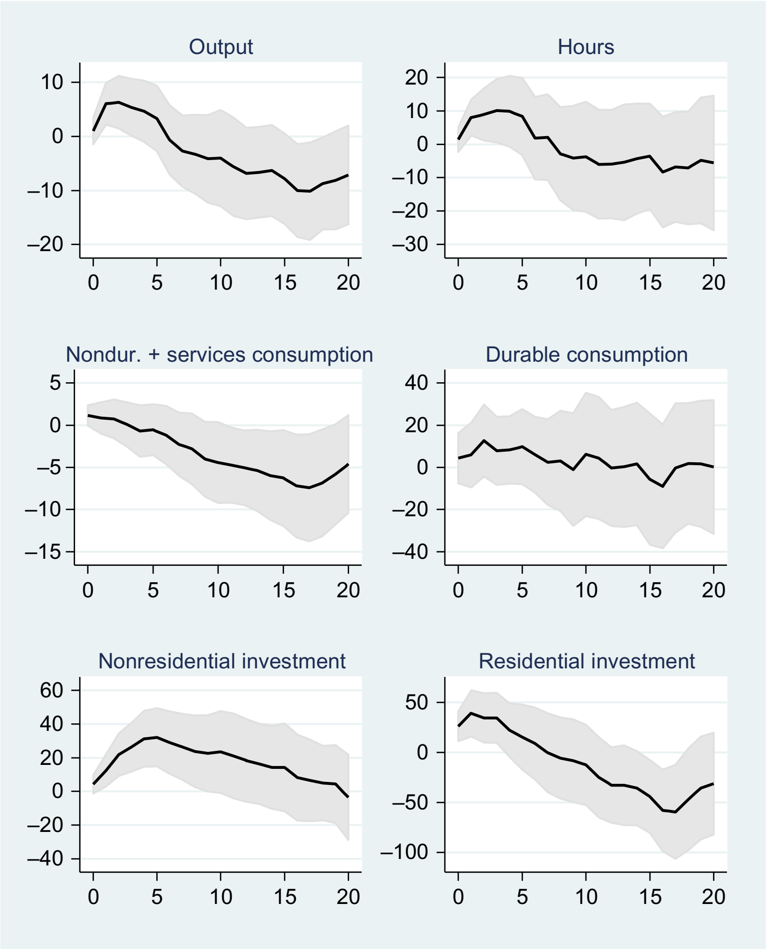


Fig. 8 Effect of news of future tax increase, Leeper et al. (2012) measure, Jordà local projection, 1954q1–2005q4. Light gray bands are 90% confidence bands.

government spending multipliers, and others find larger tax multipliers. My assessment is that the existing methods do not yield precise enough and robust enough estimates to be able to make this comparison.

5. TECHNOLOGY SHOCKS

Technology shocks are the most important type of nonpolicy shocks. In this section, I review the literature on technology shocks and present some new results comparing various shocks from the literature. I discuss both the classic unanticipated technology shocks and news about future changes in technology. I also distinguish between neutral and investment-specific technology (IST) shocks.

5.1 Neutral Technology Shocks

In 1982, Kydland and Prescott (1982) demonstrated the (then) surprising result that one could produce business cycle movements of key variables from a DSGE growth model beset by only one type of shock: variations in the growth rate of exogenous total factor productivity (TFP). To be concrete, consider the following aggregate production function:

$$Y_t = A_t F(L_t, K_t) \tag{21}$$

where Y_t is output, A_t is TFP, L_t is labor, and K_t is the capital stock at the beginning of period t. Neutral technology shocks, or TFP shocks, are shocks to the process driving A_t.

Several empirical regularities supported Kydland and Prescott's (1982) hypothesis. First, Solow (1957) showed that 87% of the growth in average labor productivity from 1909 to 1949 was due to TFP growth. If TFP growth was so important for growth, why not also for business cycles? Second, at the time that Kydland and Prescott published their article, a long-standing stylized fact was the procyclicality of labor productivity. In fact, this stylized fact was a problem for Keynesian "aggregate demand" explanations of business cycles, since diminishing returns would predict countercyclical labor productivity. Typically, the aggregate demand-driven business cycle literature had to resort to stories of labor hoarding or increasing returns to explain the procyclicality of labor productivity.

In follow-up work, Prescott (1986) used the Solow residual as his measure of exogenous TFP and used the standard deviation of that series along with his model to argue that the bulk of business cycle fluctuations could be explained by technology shocks. Beginning in the 1990s, though, several new results emerged that cast doubt on using the Solow residual as an exogenous technological progress for the purpose of business cycle analysis. First, Evans (1992) showed that variables such as money, interest rates, and government spending Granger-caused the Solow residual. Second, Hall (1988, 1990) developed a generalization of the Solow residual framework that relaxed the assumptions of competition and constant returns to scale. This framework demonstrated how endogenous components could enter the Solow Residual. Third, a number of papers, such as Shapiro (1993), Burnside et al. (1995), and Basu and Kimball (1997), used proxies such as the workweek of capital, electricity, or average hours to adjust the Solow residual for variations in the utilization of labor and/or capital. They found that much of the procyclicality of the Solow residual disappeared once it was adjusted.

Two approaches called into question whether technology shocks even led to business cycle-like movements. Galí (1999) and Basu et al. (2006) used different methods but both found results, suggesting that a positive technology shock led to a decline in labor inputs, such as hours. Both of these analyses assumed that all technology shocks were neutral technology shocks. I will discuss each of the approaches with the follow-up work in turn.

Galí (1999) used long-run restrictions to identify neutral technology shocks. He argued that a standard RBC model predicted that technology shocks were the only shocks that could have permanent effects on labor productivity. As discussed in

Section 2.3.6, Galí (1999) estimated a bivariate VAR with labor productivity and hours (or employment) and imposed the long-run restriction that technology shocks were the only shocks that could have a permanent effect on labor productivity. Francis and Ramey (2005) derived additional long-run restrictions from the theory and used them as an overidentification test and found that one could not reject the overidentifying restrictions. Francis and Ramey (2006) constructed new historical data for the United States and extended the analysis back to 1889. They found that a positive technology shock raised hours in the pre-WWII period but lowered them in the post-WWII period. They explained the switch with the difference in the serial correlation properties of productivity. In the early period, an identified technology shock raised productivity immediately, whereas in the later period an identified technology shock raised productivity more gradually. This gradual rise in the later period provides an incentive to reduce hours worked in the short run in anticipation of higher productivity in the long run.

Galí (1999) and Francis and Ramey (2005) both assumed that both (log) labor productivity and hours had a unit root and that their first differences were stationary. As Section 2.3.6 demonstrated, imposing long-run restrictions also requires the imposition of assumptions on stationarity. Christiano et al. (2003) argued that it makes no sense to model hours per capita as having a unit root since it is bounded above and below. They showed that if instead one assumes that hours are stationary and then impose the Galí long-run restriction, a positive technology shock leads to a rise in hours worked. Fernald (2007) noted the structural break in labor productivity growth, and when he allowed for that feature of the data, he found that hours fell after a positive technology shock. Francis and Ramey (2009) argued that the baby boom led to low-frequency movements in both labor productivity growth and hours worked per capita and that failure to correct for these led to the positive correlations found by Christiano et al. When they corrected for demographics, they found that a positive technology shock led to a decrease in hours. Gospodinov et al. (2013) discuss various econometric issues that arise in this setting with low-frequency movements.

Building on ideas of Uhlig (2003), Francis et al. (2014) introduced a new method of imposing long-run restrictions that overcame many of these problems. They identify the technology shock as the shock that maximizes the forecast error variance share of labor productivity at some finite horizon *h*. Using that scheme, they find that their identified technology leads to a fall in hours worked. They estimate that technology shocks contribute 15–40% of the forecast error variance of output at business cycle horizons. A variation by Barsky and Sims (2011) identifies the technology shock as the one that maximizes the *sum* of the forecast error variances up to some horizon *h*.

Several papers have questioned Galí's (1999) basic identifying assumption that technology shocks are the only shocks that have a long-run effect on labor productivity. Uhlig (2004) argues that capital taxation and shifts in preferences involving "leisure in the workplace" can also have long-run effects on labor productivity. He also introduces a "medium-run" identification procedure that anticipates the procedures discussed earlier. He finds that the impact effect on hours is zero and that there is a small rise afterward. Mertens and Ravn (2011a) include the Romer and Romer (2010) exogenous tax shocks in a vector error correction model and find that once taxes are controlled and cointegration is allowed, a positive TFP shock raises hours in the short run. They also find that technology shocks account for 50–55% of the forecast error variance of output.

Basu et al. (2006) found that technology shocks were contractionary using a completely different method. Employing theoretical insights from Basu and Kimball (1997), they adjusted the annual Solow residual for utilization using hours per worker as a proxy. When they examined shocks to this purged Solow residual, they found that positive shocks to technology led to a decline in hours worked. Fernald (2014) has now constructed a quarterly version of this utilization-adjusted TFP series.[nn]

Alexopoulos (2011) identified technology shocks by creating an entirely new data series for measuring technology. Meticulously collecting and counting book publications for several types of technologies, she constructed several annual series on new technologies. She found that these series were not Granger-caused by standard macroeconomic variables. Using her new series in VARs, she found that a positive technology raises output and productivity. Contrary to the findings of Galí (1999) and Basu et al. (2006), she estimated that a positive shock to technology raises output, though the effect is weak.

Table 8 summarizes some of the estimates of the contribution of TFP shocks to output fluctuations at business cycle frequencies, based on approaches that identify technology shocks in time series models.

Numerous papers have identified technology shocks through estimated DSGE models. McGrattan (1994) estimated a neoclassical DSGE model with technology and fiscal shocks. Smets and Wouters (2007) estimated a New Keynesian DSGE model using Bayesian methods in order to explore the effects of various shocks. They incorporate a number of different shocks in the model, including neutral technology shocks, IST shocks (discussed in the next section), monetary shocks, government spending shocks, markup shocks, and risk premium shocks. Their estimates imply that a positive neutral technology shock lowers hours worked. Justiniano et al. (2010, 2011) also estimate a New Keynesian model, but incorporate also investment-specific shocks and MEI shocks. Schmitt-Grohe and Uribe (2012) estimate a DSGE model which allows all of their shocks to have an unanticipated component and a "news," or unanticipated, component. Miyamoto and Nguyen (2015) extend their estimation method by including series on survey expectations in the estimation. I will discuss these papers in more detail in the sections on IST shocks and news. Blanchard et al. (2013) estimate a DSGE model allowing for both "news" and "noise." Table 9 summarizes the estimates from DSGE models

[nn] The series is regularly updated and made available by John Fernald at http://www.frbsf.org/economic-research/economists/jfernald/quarterly_tfp.xls.

Table 8 Estimated importance of technology shocks in SVAR models

Study	Method	Type	News?	% of output explained
Galí (1999), Francis–Ramey (2005)	Long-run restrictions, hours in first differences	TFP	No	Very little
Christiano et al. (2004)	Long-run restrictions, hours in levels	TFP	No	31–45% for horizons up to 20 quarters
Christiano et al. (2004)	Long-run restrictions, hours in first differences	TFP	No	1–17% for horizons up to 20 quarters
Basu et al. (2006)	Utilization and effort adjusted TFP	TFP	No	17–40% from 1 to 3 years
Beaudry and Portier (2006)	Short-run or long-run restrictions	TFP	Yes	50%
Fisher (2006)	Long-run restrictions involving both labor productivity and investment goods prices	TFP	No	32% at 12 quarters (see papers for more details)
Fisher (2006)	Long-run restrictions involving both labor productivity and investment goods prices	IST	No	26% (49%) at 12 quarters in early (late) sample
Mertens and Ravn (2010)	Long-run restrictions, cointegration, include taxes	TFP	No	50–55% at business cycle frequencies
Barsky and Sims (2011)	Medium-horizon restrictions	TFP	Yes	9–43% for horizons up to 24 quarters.
Barsky and Sims (2011)	Medium-horizon restrictions	TFP	No	6–20% for horizons up to 24 quarters
Francis et al. (2014)	Medium-horizon restrictions	TFP	No	15–40% for horizons up to 32 quarters.
Francis et al. (2014)	Long-run restrictions	TFP	No	40–55% for horizons up to 32 quarters
Ben Zeev and Khan (2015)	Medium-horizon restrictions	IST	Yes	73% at 8 quarters
Ben Zeev and Khan (2015)	Medium-horizon restrictions	IST	No	Very little
Ben Zeev and Khan (2015)	Medium-horizon restrictions	TFP	No	10% at 8 quarters

Notes: TFP denotes neutral total factor productivity technology, IST denotes investment-specific technology, and MEI denotes marginal efficiency of investment.

Table 9 Estimated importance of technology shocks in DSGE models

Study	Model features	Type	News?	% of output explained at business cycle frequencies
Prescott (1986)	Calibrated neoclassical DSGE model	TFP	No	75%
McGrattan (1994)	Neoclassical model with distortionary taxes and government spending	TFP	No	41%
Greenwood et al. (2000)	Calibrated DSGE model, technology identified with relative price of investment	IST	No	30%
Smets and Wouters (2007)	New Keynesian model with many types of shocks	TFP	No	15–30% from horizon 1–10 quarters
Justiniano et al. (2011)	New Keynesian model with many types of shocks	TFP	No	25%
Justiniano et al. (2011)	New Keynesian model with many types of shocks	IST	No	0%
Justiniano et al. (2011)	New Keynesian model with many types of shocks	MEI	No	60%
Schmitt-Grohe and Uribe (2012)	Distinguishes unanticipated vs anticipated, TFP vs investment specific, no sticky prices	TFP	No	25%
Schmitt-Grohe and Uribe (2012)	Distinguishes unanticipated vs anticipated, TFP vs investment specific, no sticky prices	TFP	Yes	3%
Schmitt-Grohe and Uribe (2012)	Distinguishes unanticipated vs anticipated, TFP vs investment specific, no sticky prices	IST	No	21%
Schmitt-Grohe and Uribe (2012)	Distinguishes unanticipated vs anticipated, TFP vs investment specific, no sticky prices	IST	Yes	7%
Khan and Tsoukalas (2012)	New Keynesian model, distinguishes unanticipated vs anticipated	TFP	No	24%
Khan and Tsoukalas (2012)	New Keynesian model, distinguishes unanticipated vs anticipated	MEI	No	47%
Khan and Tsoukalas (2012)	New Keynesian model, distinguishes unanticipated vs anticipated	IST	No	1.2%
Khan and Tsoukalas (2012)	New Keynesian model, distinguishes unanticipated vs anticipated	TFP + MEI + IST	Yes	1.4%
Miyamoto and Nguyen (2013)	Extends Schmitt–Grohe–Uribe analysis by using data on expectations	TFP	No	19%
Miyamoto and Nguyen (2013)	Extends Schmitt–Grohe–Uribe analysis by using data on expectations	TFP	Yes	7%
Miyamoto and Nguyen (2013)	Extends Schmitt–Grohe–Uribe analysis by using data on expectations	IST	No	27%
Miyamoto and Nguyen (2013)	Extends Schmitt–Grohe–Uribe analysis by using data on expectations	IST	Yes	12%

Notes: TFP denotes neutral total factor productivity technology, IST denotes investment-specific technology, and MEI denotes marginal efficiency of investment.

of the contribution of various types of technology shocks to output fluctuations at business cycle frequencies.

5.2 Investment-Related Technology Shocks

Greenwood et al. (1988) were the first to examine in a DSGE model Keynes' idea that shocks to the marginal efficiency of investment (MEI) could be a source of business cycle volatility. In follow-up work, Greenwood et al. (2000) used a calibrated DSGE model to examine the importance of IST change in business cycles. They used the relative price of new equipment to identify the process driving IST shocks and concluded that these shocks could account for 30% of business cycle volatility.

Fisher (2006) extended Galí's (1999) analysis of neutral technology shocks by incorporating additional data and restrictions that separately identify neutral and IST shocks. In particular, he assumed that only IST shocks affect the relative price of investment goods in the long run and only neutral technology and IST technology shocks affect labor productivity in the long run. Because of some sample instability, he estimated his model over two periods: 1955q1–1979q2 and 1982q3–2000q4. He found that both technology shocks together accounted for a substantial shared of the forecast error variance of output, 60% at 12 quarters in the early sample, 83% in the later sample.

Justiniano et al. (2010, 2011) estimate a New Keynesian DSGE model with a variety of unanticipated shocks. Justiniano et al. (2011) distinguish between IST shocks and MEI shocks. To be concrete, consider simplified versions of two equations from their DSGE model:

$$I_t = \Psi_t Y_t^I \tag{22a}$$

$$K_{t+1} = (1-\delta)K_t + \mu_t I_t \tag{22b}$$

I_t is the production of investment goods and Ψ_t denotes the rate of transformation of final goods, Y_t^I, into investment goods. Ψ_t is IST which, according to their model, should be equal to the inverse of the relative price of investment goods to consumption goods. K_{t+1} is the level of capital at the beginning of period $t+1$, δ is the depreciation rate, and μ_t is the rate of transformation between investment goods and installed capital, or the *MEI*. Previous research, such as by Greenwood et al. (2000) and Fisher (2006), had not distinguished IST from MEI and had assumed their product was equal to the inverse of the relative price of investment goods. Justiniano et al. (2011) estimate that (unanticipated) MEI shocks contribute 60% of the variance of output at business cycle frequencies.

Schmitt-Grohe and Uribe (2012), Khan and Tsoukalas (2012), and Miyamoto and Nguyen (2015) estimate DSGE models that incorporate both TFP and IST shocks. An important focus of their estimation is the distinction between unanticipated technology changes and news about future changes, so I will discuss their work in the next section on news.

Although there is a wide range of results, a general pattern that emerges is that when models include IST and/or MEI shocks, they tend to explain a significant portion of the variation in output at business cycle frequencies.

5.3 News About Future Technology Changes

Both Pigou (1927) and Keynes (1936) suggested that changes in expectations about the future may be an important driver of economic fluctuations. Beaudry and Portier (2006) reignited interest in the idea by providing time series evidence that news about future productivity could explain half of output fluctuations over the business cycle. Furthermore, their estimates implied that hours and output rose when the news arrived, thus creating business cycle-type comovements. They identified news shocks using two methods; both involved identifying shocks that moved stock prices immediately, but affected productivity only with a lag. Beaudry and Lucke (2010) and Kurmann and Otrok (2013) used other techniques to reach similar conclusions. More recently, however, Barsky and Sims (2011) and Barsky et al. (2014) have used medium-run restrictions and series on consumer confidence to identify news shocks and found that news shocks did not generate business cycle fluctuations. In particular, hours fell when news arrived. Fisher (2010), Kurmann and Mertens (2014), and Forni et al. (2014) have highlighted problems with Beaudry and Portier's identification method. For example, Kurmann and Mertens (2014) show that the larger VECM in Beaudry and Portier's (2006) paper is not identified. Forni et al. (2014) argue that the small-scale SVARs are affected by the "nonfundamentalness" problem discussed in Section 2.5. Thus, the empirical work based on time series identification is in flux. Beaudry and Portier (2014) present a comprehensive summary of the literature.

I would add that another potential problem with Beaudry and Portier's (2006) method for identifying TFP news shocks is the implicit assumption they make about stock prices. They assume that the future profits from the TFP shock will show up in current stock prices. It is not clear that this assumption holds for major innovations. Greenwood and Jovanovic (1999) and Hobijn and Jovanovic (2001) present theory and evidence that major technological innovations (such as information technology) actually lead to a temporary *decline* in stock market values because they lower the value of the existing technology. Revolutionary innovations usually arise in private companies and claims to future dividend streams only show up in stock prices after the initial public offerings. Thus, we should not necessarily see positive effects of news about future TFP on stock prices.

Ben Zeev and Khan (2015) identify both unanticipated IST shocks and IST news shocks. To do this, they extend Barsky and Sims (2011) medium-horizon restriction method for identifying news and employ Fisher's (2006) assumptions linking IST and the relative price of investment goods. They find that IST news shocks explain 73% of the forecast error variance of output at a horizon of eight quarters. They show that the IST shocks originally identified by Fisher (2006) were a combination of the unanticipated IST shocks and news about IST. Ben Zeev and Khan's paper thus corroborates Fisher's finding that IST shocks are the major source of fluctuations, but goes on to show that it is the news part that is the most important.

Another strategy for identifying news is through estimation of a DSGE model, as pioneered by Schmitt-Grohe and Uribe (2012). This approach achieves part of its identification by assuming particular lags between the arrival of news and the realization of the change. Schmitt-Grohe and Uribe (2012) allow for a variety of unanticipated and news shocks for variables such as TFP, IST, and wage markups. They estimate that all news variables combined (including nontechnology shocks such as wage markup shocks) account for 50% of output fluctuations according to their estimates. An extension by Miyamoto and Nguyen (2015) uses actual survey forecasts for the expectations variables.

Khan and Tsoukalas (2012) estimate a New Keynesian DSGE model with both IST and MEI shocks and allow for both unanticipated changes and news shocks. They find that unanticipated MEI shocks contribute an important part of the variance of output (47%), but that technology news shocks are not important at all. Nontechnology news shocks do, however, contribute to the variance decomposition of hours. In particular, wage markup shocks account for over 40% of the variance of hours. Thus, their results on the importance of unanticipated technology shocks contrast with those of Schmitt-Grohe and Uribe (2012), but their results on the importance of news about wage markups are consistent with their findings. The estimates of the importance of technology news are summarized in Table 9.

5.4 Explorations with Estimated Technology Shocks

I now study the relationship between some of the leading shocks and explore the effects of a few of them in the Jordà local projection framework. I reestimate the Galí (1999), Christiano et al. (2003) (CEV), and Beaudry and Portier (2006) systems using updated data. In each case, I used a simple bivariate system. Both the Galí and CEV shocks use long-run restrictions, with the former assuming a unit root in hours per capita and the latter assuming a quadratic trend in hours. I use Beaudry and Portier's shock estimated with the short-run restriction; ie, it is the shock to stock prices that does not affect TFP on impact; the correlation of this shock with their shock estimated using long-run restrictions is 0.97. The Fernald shocks are simply the growth rate of Fernald's (2014) utilization-adjusted TFP for the aggregate economy or for the investment goods sector. The rest of the estimated shocks were kindly provided by Francis et al. (2014) (medium-horizon restrictions), Barsky and Sims (2011) (medium-run restrictions, consumer confidence), Justiniano et al. (2011) (estimated DSGE model), Ben Zeev and Khan (2015) (SVAR with medium-run restrictions), and Miyamoto and Nguyen (2015) (estimated DSGE model with forecast data). The joint sample is 1955q2–2006q4, except for the TFP news sample, which is limited to 1961q1–2006q4 by the Barsky and Sims shock availability. Correlations between subsets of shocks that are available over longer samples are similar to those reported for the joint sample. Table 10 shows the correlations, broken

Table 10 Correlation of various estimated technology shocks (sample is 1955q2–2006q4)

A. Unanticipated TFP shocks

	gali_tfp	cev_tfp	jf_tfp	ford_tfp	bzk_tfp	jpt_tfp	mn_tfp_p	mn_tfp_s
gali_tfp	1.00							
cev_tfp	0.62	1.00						
jf_tfp	0.68	0.42	1.00					
ford_tfp	0.75	0.62	0.62	1.00				
bzk_tfp	0.67	0.78	0.54	0.63	1.00			
jpt_tfp	0.68	0.69	0.53	0.54	0.63	1.00		
mn_tfp_p	0.17	0.16	0.20	0.28	0.08	0.16	1.00	
mn_tfp_s	0.52	0.59	0.47	0.52	0.58	0.62	0.10	1.00

B. TFP news shocks

	bp_news	bs_news	mn_p_n4	mn_p_n8	mn_s_n4	mn_p_n8
bp_news	1.00					
bs_news	0.25	1.00				
mn_p_n4	0.08	0.12	1.00			
mn_p_n8	0.05	0.00	0.29	1.00		
mn_s_n4	0.04	−0.04	0.53	−0.14	1.00	
mn_p_n8	0.05	0.00	0.29	1.00	−0.14	1.00

C. Unanticipated IST or MEI shocks

	jf_ist	bzk_ist	jpt_mei	jpt_ist	mn_ist_p	mn_ist_s
jf_ist	1.00					
bzk_ist	0.17	1.00				
jpt_mei	−0.27	0.05	1.00			
jpt_ist	0.19	0.49	−0.01	1.00		
mn_ist_p	0.03	0.31	0.17	0.20	1.00	
mn_ist_s	−0.13	0.11	0.27	0.14	−0.06	1.00

D. IST news shocks

	bzk_news	mn_p_n4	mn_p_n8	mn_s_n4	mn_s_n8
bzk_news	1.00				
mn_p_n4	0.15	1.00			
mn_p_n8	0.02	0.18	1.00		
mn_s_n4	0.12	0.07	0.12	1.00	
mn_s_n8	0.08	0.01	0.02	0.28	1.00

Abbreviations: *bp*, Beaudry-Portier; *bs*, Barsky–Sims; *bzk*, Ben Zeev and Khan; *cev*, Christiano, Eichenbaum, Vigfusson; *ford*, Francis, Owyang, Roush, DiCecio; *gali*, Gali; *ist*, investment-specific technology; *jf*, John Fernald; *jpt*, Justiniano, Primiceri, Tambolotti; *mei*, marginal efficiency of invest; *mn*, Miyamoto and Nguyen; *n4*, news with 4 quarter lead; *n8*, news with 8 quarter lead; *_p*, permanent; *_s*, stationary; *tfp*, total factor productivity.

down according to whether the shock is to TFP or IST (or MEI) and whether it is unanticipated or is news.

Table 10A shows the results for unanticipated TFP shocks, which have received the most attention. Most of the shocks have a correlation above 0.6 with the shock estimated using Galí's (1999) method. The exception is the Miyamoto and Nguyen (2015) permanent TFP shock. It is surprising that the Miyamoto and Nguyen stationary TFP shock has a higher correlation than the permanent TFP shock, since the Galí method only identifies permanent TFP shocks.

Table 10B shows news shocks about TFP. The correlation between Beaudry and Portier's (2006) shock estimated using short-run restrictions and Barsky and Sims' (2011) shock estimated using medium-horizon restrictions is only 0.25. The correlations of both of those SVAR-based shocks with the Miyamoto and Nguyen (2015) DSGE-based shocks are essentially 0.

Table 10C shows correlations of various estimates of unanticipated IST or MEI shocks. The correlations between the various estimates are quite low. For example, the correlation between Fernald's utilization-adjusted TFP for the investment goods sector and Justiniano et al.'s (2011) IST shock is only 0.19 and is −0.27 for Justiniano et al.'s MEI shock. The highest correlation of 0.49 is between Justiniano et al.'s IST shock and Ben Zeev and Khan's (2015) IST shock. The higher correlation is not surprising since both methods associate the IST shock with the inverse of the relative price of equipment.

Table 10D shows correlations of various estimates of IST news shocks. There is essentially no correlation between Ben Zeev and Khan's (2015) SVAR-based estimates and Miyamoto and Nguyen's (2015) DSGE-estimated shocks.

If we were simply trying to develop instruments for estimating structural parameters, it would not matter if various instruments had low correlation.[oo] If, however, we are trying to estimate shocks in order to determine their importance for macroeconomic fluctuations, a low correlation across various estimates is troubling. The large number of low correlations across methods and the widely varying results reported across papers suggest that we are still far from a consensus on the nature and importance of technology shocks. The problem is not one of lack of consensus of estimated DSGE vs SVAR methods. Even within each class of method, the results vary widely, as evidenced in Tables 8 and 9.

Moreover, many of the estimated shocks do not satisfy the property that they are unanticipated or news. Table 11 shows the *p*-values from two sets of tests. The first one tests for serially correlation of the shock by regressing the shock on its own two lags and testing their joint significance. The SVAR-estimated shocks do well on this test, but quite a few of the DSGE-estimated shocks fail this test. The second set of tests is for Granger causality (Granger, 1969). To conduct these tests, I augment this regression with two lags each of log real GDP, log real consumption, and log real stock prices. I chose

[oo] Sims (1998) made this argument in his discussion of Rudebusch's (1998) monetary shock critique.

Table 11 Tests for serial correlation and Granger causality

	p-Value test for significance of own lags	*p*-Value test for Granger causality
gali_tfp	0.986	**0.020**
cev_tfp	0.986	**0.000**
jf_tfp	0.718	**0.001**
jf_ist	0.644	**0.000**
ford_tfp	0.991	0.855
bp_tfp_news_sr	0.999	0.910
bs_tfp_news	0.834	0.935
bzk_ist_news	0.724	**0.049**
bzk_ist	0.981	0.740
bzk_tfp	0.949	0.992
jpt_tfp	0.101	**0.000**
jpt_mei	**0.006**	**0.000**
jpt_ist	0.941	0.854
mn_tfp_p	0.133	**0.000**
mn_ist_p	**0.000**	0.287
mn_tfp_s	**0.010**	**0.008**
mn_ist_s	**0.000**	**0.024**
mn_tfp_p_n4	**0.000**	**0.001**
mn_tfp_p_n8	**0.000**	**0.087**
mn_ist_p_n4	**0.000**	0.924
mn_ist_p_n8	**0.000**	**0.076**
mn_tfp_s_n4	**0.098**	0.134
mn_tfp_s_n8	0.353	0.783
mn_ist_s_n4	**0.000**	0.497
mn_ist_s_n8	**0.000**	**0.052**

Notes: The tests for serial correlation are conducted by regressing the shock on its own two lags and testing the joint significance. The tests for Granger causality are conducted by regressing the shock on its own two lags, as well as two lags of log real GDP, log real consumption, and log real stock prices. The test is on the joint significance of the lags of the three additional variables. P-values less than 0.1 are indicated in bold.

Abbreviations: *bp*, Beaudry–Portier; *bs*, Barsky–Sims; *bzk*, Ben Zeev and Khan; *cev*, Christiano, Eichenbaum, Vigfusson; *ford*, Francis, Owyang, Roush, DiCecio; *gali*, Gali; *ist*, investment-specific technology; *Jpt*, Justiniano, Primiceri, Tambolotti; *mei*, marginal efficiency of invest; *mn*, Miyamoto and Nguyen;*n4*, news with 4 quarter lead; *_p*, permanent; *_s*, stationary; *tfp*, total factor productivity.

consumption and stock prices because those variables have forward-looking components to them. Half of the shocks fail this test. Of course, the Galí and CEV shocks were estimated using a simple bivariate model. Had I augmented those systems with these variables, the shocks would have passed the tests by construction. The Francis et al., Beaudry and Portier, and Ben Zeev and Kahn shocks pass this test, as do about half of the DSGE-estimated shocks.

Next, I explore the effects of a few of the estimated shocks on several macroeconomic variables in a Jordà local projection framework. To do this, I estimate the following series of regressions:

$$z_{t+h} = \alpha_h + \theta_h \cdot \text{shock}_t + \varphi_h(L)y_{t-1} + \text{quadratic trend} + \varepsilon_{t+h} \tag{23}$$

The z is the variable of interest. The control variables include two lags each of the shock (to mop up any serial correlation in the shocks), log real GDP per capita, log real stock prices per capita, log labor productivity (equal to real GDP divided by total hours worked), and the dependent variable. The coefficient θ_h gives the response of z at time $t+h$ to a shock at time t. As discussed in Section 2, ε_{t+h} will be serially correlated, so the standard errors must incorporate a correction, such as Newey–West.

Fig. 9 shows the responses of real GDP, labor productivity hours, stock prices, consumption, and nonresidential investment to three different measures of unanticipated TFP shocks: the Francis, Owyang, Roush, and DiCecio (2014) (FORD) measure, which

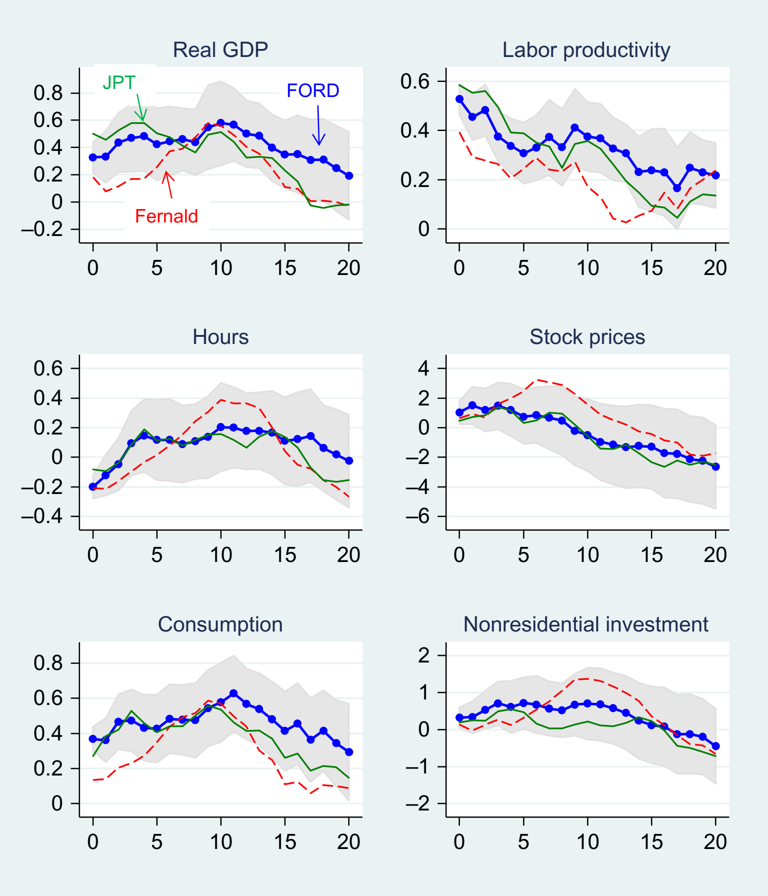


Fig. 9 Effects of TFP shock, Jordà local projection, various samples. Francis, Owyang, Roush, DiCecio (FORD): *blue lines with circles* (*black* in the print version); Fernald utilization-adj TFP: *dashed red* (*gray* in the print version) *lines*; Justiniano, Primiceri, Tambalotti (JPT) DSGE TFP: *solid green* (*gray* in the print version) *lines*. Light gray bands are 90% confidence bands.

uses medium-run restrictions; Fernald's (2014) utilization-adjusted TFP growth, and Justiniano, Primiceri, and Tambalotti's (2011) (JPT) estimate from their DSGE model. The responses to the FORD and JPT shocks are quite similar: GDP, labor productivity, stock prices, and consumption all jump immediately and significantly. Hours fall for a few quarters before rising. The Fernald shock implies a more hump-shaped response of GDP, hours, stock prices, consumption, and investment. Labor productivity rises immediately but then returns to normal at around 16 quarters. The Fernald shock also shows an initial decline in hours before they rise.

Fig. 10 shows the effects of the Ben Zeev and Khan (2015) IST news shock. Recall that this shock is an extension of Fisher's (2006) method, estimated using the Barsky–Sims (2011) medium-horizon method combined with information on relative prices of investment. This shock appears to generate a classic business cycle pattern. GDP, hours, stock

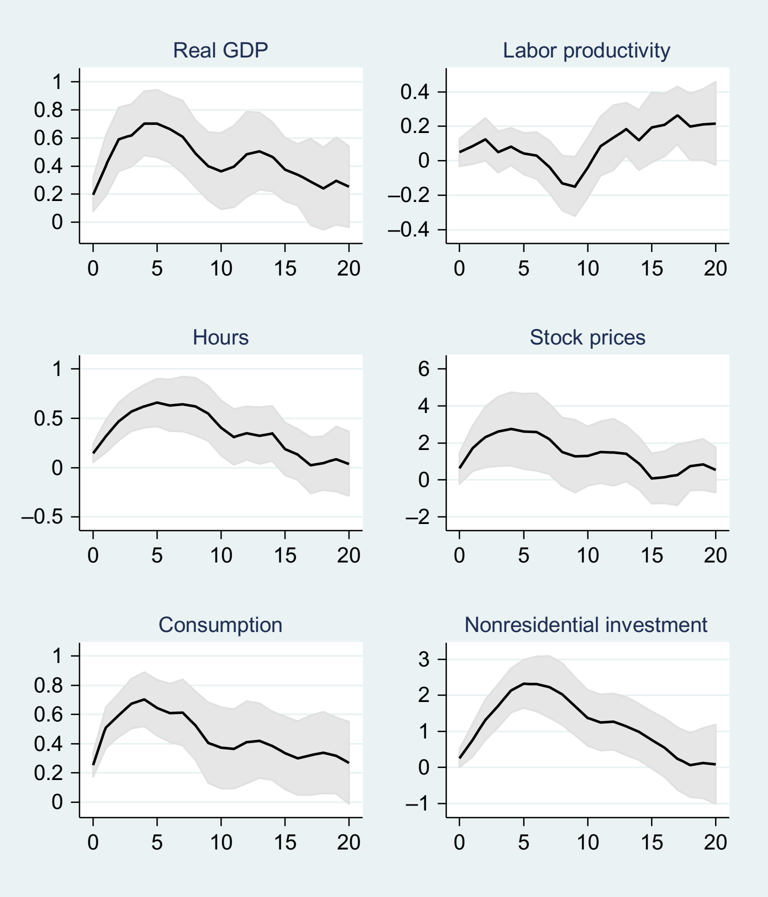


Fig. 10 Effects of news of investment-specific technology shocks, Ben Zeev–Khan (2015) measure, Jordà local projection, 1952q1–2012q1. Light gray bands are 90% confidence bands.

prices, consumption, and nonresidential investment increase with a prolonged hump shape. Labor productivity does nothing for about six quarters, falls around nine quarters, and then rises.

Fig. 11 shows the effects of JPT's MEI shock, estimated from their DSGE model. While this shock leads to temporary rises (for a year or less) in real GDP, labor productivity, consumption, and nonresidential investment, it leads to a fall in stock prices, which is puzzling.

Tables 12A and 12B show the forecast error variance decompositions for these five shocks for both output and hours. The decompositions are calculated from a standard VAR with the shock, and log per capita values of real GDP, hours, stock prices, consumption, and nonresidential investment. Although some of the unanticipated TFP shocks can account for up to 16% of output, none accounts for much of the variance

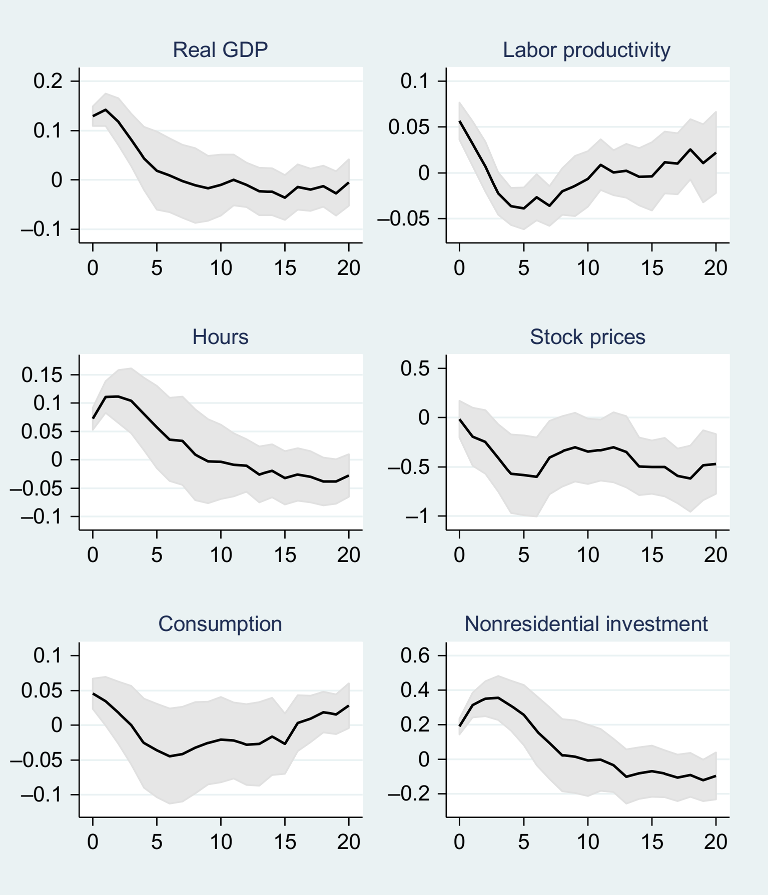


Fig. 11 Effects of marginal efficiency of investment shock, Justiniano et al. (2011) measure, Jordà local projection, 1954q3–2009q1. Light gray bands are 90% confidence bands.

Table 12A TFP shock contribution to the forecast error variance of output and hours

	FORD TFP		Fernald TFP		JPT TFP	
Horizon (in quarters)	**Output**	**Hours**	**Output**	**Hours**	**Output**	**Hours**
0	16.2	10.5	6.1	10.5	28.2	1.0
4	13.1	2.0	2.0	2.4	15.1	0.9
8	14.3	1.9	2.8	1.3	15.9	1.6
12	14.3	1.6	3.1	1.2	16.3	1.6
16	14.0	1.5	3.1	1.5	16.0	1.6
20	13.7	1.5	3.0	2.0	15.7	1.9

Notes: These results are based on a standard VAR with the shock, output, hours stock prices, consumption, and nonresidential investment. The shock is ordered first. Four lags are included, along with a quadratic trend.

Table 12B Investment-related technology shock contribution to the forecast error variance of output and hours

	Ben Zeev–Khan IST news		JPT MEI	
Horizon (in quarters)	**Output**	**Hours**	**Output**	**Hours**
0	7.8	6.9	49.6	26.4
4	33.2	31.3	19.8	20.9
8	36.8	38.5	11.9	12.1
12	36.8	38.8	11.4	10.5
16	36.4	37.9	11.3	10.1
20	35.9	36.8	11.1	9.8

Notes: These results are based on a standard VAR with the shock, output, hours stock prices, consumption, and nonresidential investment. The shock is ordered first. Four lags are included, along with a quadratic trend.

of hours. In contrast, the Ben Zeev and Khan IST news shock accounts for well over a third of the forecast error variance of both output and hours. JPT's MEI shock accounts for large fractions on impact, 50% for output and 26% for hours, but the effects dissipate fairly quickly.

5.5 Summary of Technology Shocks

The literature investigating the effects of technology shocks has moved far beyond the simple Solow residual. Various methods have been introduced to deal with changes in measured TFP that are due to variable utilization. Moreover, the literature has moved beyond neutral technology shocks to recognize the potential importance of IST shocks and MEI shocks. In addition, recent contributions have investigated the importance of news shocks.

My analysis shows, however, that some of the estimated shocks are not highly correlated with other versions. Moreover, many of the shocks are serially correlated or Granger-caused by other variables. This suggests that more research needs to be done to refine these shock measures. One of the shocks that seem to be promising both for generating business-cycle comovement and for contributing significant amounts to the variance of output is the shock that captures news about IST change.

6. ADDITIONAL SHOCKS

So far, this chapter has focused on only three types of shocks—monetary, fiscal, and technology shocks. There are numerous other candidates for potentially important macroeconomic shocks. Here, I will briefly review a few.

One obvious additional candidate for a macroeconomic shock is oil shocks. Hamilton (1983) argued that oil supply shocks were a major driver of economic fluctuations. Since then, a large literature has examined the effects of oil supply shocks. One of the major themes of the literature is the changing estimated effects of oil price shocks, identified by ordering oil prices first in a linear VAR. In particular, after the 1970s oil price changes seemed to have smaller effects. One potential explanation is asymmetries. Several analyses, such as by Davis and Haltiwanger (2001) and Hamilton (2003), argued that oil price increases have larger effects than oil price decreases. Subsequent research, however, found that there was not strong evidence of asymmetry (eg, Kilian and Vigfusson, 2011). A second potential explanation for the changing effects of oil supply shocks is that the oil price increases in earlier periods were accompanied by price controls, which led to many distortions (Ramey and Vine, 2011). When they constructed an implicit cost of oil that incorporated a proxy for the distortion costs, they did not find much evidence of changing effects. A third explanation was by Kilian (2009) and was a critique of standard identification methods. He argued that many of the changes in oil prices are driven now driven by demand shocks, not supply shocks, so a standard Cholesky decomposition with oil prices ordered first does not properly identify oil supply shocks. Stock and Watson's (forthcoming) chapter in this *Handbook* uses oil shocks as a case study of their methods. They find that oil supply shocks, identified using Kilian's (2009) method, do not account for a significant fraction of the forecast error variance of output.

Credit shocks are another possible candidate for a macroeconomic shock. There is huge literature analyzing the importance of credit and credit imperfections in economic fluctuations and growth. Most of this literature focuses on credit as an important propagation and amplification mechanism (eg, "the credit channel" of monetary policy), rather than as an important independent source of shocks. Gilchrist and Zakrajšek's (2012) recent analysis of the effects of innovations to their new excess bond premium variable can be interpreted as an analysis of credit market shocks. They showed that innovations to the excess bond premium that were orthogonal to the current state of the economy had significant effects on macroeconomic variables. They interpret a negative "shock" to this variable as signaling a reduction in the effective risk-bearing capacity of the financial sector.

The role of uncertainty in the business cycle has received heightened attention recently. In addition to standard firm-level uncertainty and financial uncertainty, recent work has suggested a possible role for policy uncertainty. More research needs to be done to untangle uncertainty as an endogenous propagation mechanism vs uncertainty as an independent source of macroeconomic shocks.

Labor supply shocks are yet another possible source of macroeconomic shocks. It is well known that cyclical variations in the "labor wedge" are an important component of business cycles. Shapiro and Watson (1988) estimated an SVAR with long-run restrictions and found that labor supply shocks were the dominant driver of business cycles. In estimated DSGE models with many shocks, wage markup shocks are often found to play an important role. This is particularly the case for *news* about wage markups. For example, both Khan and Tsoukalas (2012) and Schmitt-Grohe and Uribe (2012) find that wage markup news shocks account for 60% of the variance share of hours. A key question is whether exogenous shocks to the labor market are an important part of fluctuations or whether we are accidentally identifying an internal propagation mechanism as an exogenous shock.

7. SUMMARY AND CONCLUSIONS

This chapter has summarized the new methods and new findings concerning macroeconomic shocks and their propagation. Identification is particularly challenging in macroeconomics because researchers ask questions for which *dynamics* are all important, *general equilibrium* effects are crucial, and *expectations* have powerful effects.

The literature has made substantial progress in thinking seriously about identification of shocks since the early days of Cholesky decompositions. It now exploits new data sources, such as narrative records, survey expectations, and high-frequency financial data, combines theory with extra data series (eg, the relative price of investment goods), and incorporates that information in estimated DSGE models and SVARs.

The introduction to this chapter posed the question: Are we destined to remain forever ignorant of the fundamental causes of economic fluctuations? I would argue "no." Although we still have far to go, substantial progress has been made since Cochrane (1994) asked that question.

In support of my answer, I offer the following forecast error variance decomposition that combines some of the leading shocks I have discussed in this chapter. I specify a VAR that contains the shocks as well as some macroeconomic variables. In particular, it contains (in this order) Ben Zeev and Pappa's (forthcoming) military news shock, Leeper et al. (2012) news about future taxes from bond prices, the Romer and Romer's (2010) unanticipated tax shocks (as constructed by Mertens and Ravn, 2012), Francis et al.'s (2014) medium-horizon restriction TFP shock, Ben Zeev and Khan's (2015)

IST news shock, and Justiniano et al.'s (2011) MEI shock. The macroeconomic variables include the logs real per capita values of GDP and total hours, as well as the log of commodity prices and the GDP deflator. Ordered last is the federal funds rate. Four lags are used and a quadratic trend is included.

Table 13 shows the forecast error variance decomposition of log real GDP per capita and log hours per capita. Because of data limitations on some of the shocks, the sample starts after the Korean War. It should not be surprising, then, that the government spending shocks are not very important. The tax news shocks contribute a small amount, less than 10%. The unanticipated tax shocks are unimportant.

Which shocks are important? The most important for both output and hours is Ben Zeev and Khan's (2015) news about IST change. This variable contributes an important part of the forecast error variance of both output and hours. For example, at 8 quarters the contribution to hours is 40%. The 90% confidence interval (not shown in the table) is (25, 54). Justiniano et al.'s (2011) MEI shock contributes 42% on impact, with a 90% confidence interval of (34, 50), but this falls to 24% by 1 year. If we associate the innovations to the federal funds rate with monetary policy shocks, then monetary policy shocks contribute up to 8% of the variance of output, but up to 18% of the variance in hours.

Table 13 Combined VAR: shock contribution to the forecast error variance of output and hours: 1954q3–2005q4

Horizon	bzp_gov	lrw	rrtaxu	ford_tfp	bzk_ist_news	jpt_mei	ffr
A. Output							
0	5.5	0.1	2.4	15.8	11.8	42.1	0.0
4	1.6	5.6	1.6	15.1	28.8	23.9	2.0
8	1.4	4.8	1.5	13.9	26.9	16.3	6.1
12	3.0	4.8	1.2	12.6	22.1	13.6	8.1
16	4.4	6.9	1.2	11.2	19.6	12.5	7.8
20	4.9	8.5	1.2	10.7	18.6	11.9	7.4
B. Hours							
0	2.3	0.8	0.3	17.6	13.2	20.5	0.0
4	0.5	6.6	0.8	3.7	38.3	22.1	3.2
8	0.9	6.3	0.9	2.4	39.5	14.2	10.9
12	4.1	5.2	0.7	1.8	33.4	11.5	16.8
16	7.3	6.0	0.7	1.7	28.6	10.6	18.3
20	8.9	7.0	0.8	2.0	26.7	10.2	18.1

Notes: These results are from a standard VAR with four lags and a quadratic trend estimated from 1954q3 to 2005q4. The variables are as follows, in this order: Bzp_gov, lrw, rrtaxu, ford_tfp, bzk_ist_news, jpt_mei, log real GDP per capita, log total hours per capita, log commodity prices, log GDP deflator, federal funds rate.
Abbreviations: *bzk*, Ben Zeev and Khan; *bzp*, Ben Zeev and Pappa; *ffr*, federal funds rate; *ford*, Francis, Owyang, Roush, DiCecio; *lrw*, Leeper, Richter, Walker anticipated future tax; *ist*, investment-specific technology; *Jpt*, Justiniano, Primiceri, Tambolotti; *mei*, marginal efficiency of invest; *rrtaxu*, Romer–Romer unanticipated tax; *tfp*, total factor productivity.

In sum, the three fiscal shocks, the three technology shocks, and the federal funds rate shock contribute 63–79% of the variances of output and hours at horizons of 4–20 quarters. While much more work should be done exploring the plausibility of the identifying assumptions, testing the robustness of these shock estimates, and making sure that they do satisfy the properties a shock should satisfy, these results suggest that we are indeed closer to understanding Slutsky's random shocks that drive macroeconomic fluctuations.

ACKNOWLEDGMENTS

I wish to thank John Cochrane, Marco Del Negro, Graham Elliott, Neville Francis, Marc Giannoni, Robert Hall, Arvind Krishnamurthy, Kurt Lunsford, Karel Mertens, Christina Romer, David Romer, James Stock, John Taylor, Harald Uhlig, Mark Watson, Johannes Wieland, and participants at the Stanford Handbook of Macro conference and NBER Monetary Economics conference for very helpful discussions. I am grateful to the numerous authors who sent their estimated technology shocks and to Shihan Xie for providing her updated FAVAR factors. I would also like to express appreciation to the American Economic Association for requiring that all data and programs for published articles be posted. In addition, I am grateful to researchers who publish in journals without that requirement but still post their data and programs on their websites.

REFERENCES

Alesina, A., Ardagna, S., 1998. Tales of fiscal adjustment. Econ. Policy 13 (27), 487–545.

Alesina, A., Ardagna, S., 2010. Large changes in fiscal policy: taxes versus spending. In: Brown, J.R. (Ed.), Tax Policy and the Economy, vol. 24. National Bureau of Economic Research, Cambridge, MA.

Alesina, A., Perotti, R., 1995. Fiscal expansions and fiscal adjustments in OECD countries. Econ. Policy 10 (21), 205–248.

Alesina, A., Perotti, R., 1997. Fiscal adjustments in OECD countries: composition and macroeconomic effects. IMF Staff. Pap. 44, 210–248.

Alexopoulos, M., 2011. Read all about it!! What happens following a technology shock? Am. Econ. Rev. 101 (4), 1144–1179.

Amir Ahmadi, P., Uhlig, H., 2015. Sign Restrictions in Bayesian FAVARs with an Application to Monetary Policy Shocks. NBER Working Paper 21738.

Angrist, J.D., Jordà, O., Kuersteiner, G., 2013. Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited. NBER Working Paper 19355.

Arias, J.E., Caldara, D., Rubio-Ramirez, J.F., 2015a. The Systematic Component of Monetary Policy in SVARS: An Agnostic Identification Procedure. January 2015 Working Paper.

Arias, J.E., Rubio-Ramirez, J.F., Waggoner, D.F., 2015b. Inference Based on SVARs Identified with Sign and Zero Restrictions: Theory and Applications. November 2015 Working Paper.

Auerbach, A., Gorodnichenko, Y., 2012. Measuring the output responses to fiscal policy. Am. Econ. J. Econ. Pol. 4 (2), 1–27.

Auerbach, A., Gorodnichenko, Y., 2013. Fiscal multipliers in recession and expansion. In: Alesina, A., Giavazzi, F. (Eds.), Fiscal Policy After the Financial Crisis. University of Chicago Press, Chicago, IL.

Bagliano, F.C., Favero, C.A., 1999. Information from financial markets and VAR measures of monetary policy. Eur. Econ. Rev. 43 (4–6), 825–837.

Barakchian, S.M., Crowe, C., 2013. Monetary policy matters: evidence from new shocks. J. Monet. Econ. 60 (8), 950–966.

Barro, R.J., 1977. Unanticipated money growth and unemployment in the United States. Am. Econ. Rev. 67 (2), 101–115.

Barro, R.J., 1978. Unanticipated money, output, and the price level in the United States. J. Polit. Econ. 86 (4), 549–580.

Barro, R.J., 1981. Output effects of government purchases. J. Polit. Econ. 89 (6), 1086–1121.

Barro, R.J., Redlick, C.J., 2011. Macroeconomic effects from government purchases and taxes. Q. J. Econ. 126 (1), 51–102.

Barsky, R.B., Sims, E.R., 2011. News shocks and business cycles. J. Monet. Econ. 58 (3), 273–289.

Barsky, R.B., Sims, E., 2012. Information, animal spirits, and the meaning of innovations in consumer confidence. Am. Econ. Rev. 102 (4), 1343–1377.

Barsky, R.B., Basu, S., Lee, K., 2014. Whither News Shocks? NBER Working Paper 20666.

Barth III, M.J., Ramey, V.A., 2002. The cost channel of monetary transmission. In: Bernanke, B.S., Rogoff, K. (Eds.), NBER Macroeconomics Annual 2001, vol. 16. MIT Press, Cambridge, MA, pp. 199–256.

Basu, S., Kimball, M.S., 1997. Cyclical productivity with unobserved input variation. No. w5915, National Bureau of Economic Research.

Basu, S., Fernald, J.G., Kimball, M.S., 2006. Are technology improvements contractionary? Am. Econ. Rev. 96, 1418–1448.

Baumeister, C., Hamilton, J.D., 2015. Sign restrictions, structural vector autoregressions, and useful prior information. Econometrica 83, 1963–1999.

Beaudry, P., Lucke, B., 2010. Letting different views about business cycles compete. In: Acemoglu, D., Rogoff, K., Woodford, M. (Eds.), NBER Macroeconomics Annual 2009, vol. 24. University of Chicago Press, Chicago, IL, pp. 413–455.

Beaudry, P., Portier, F., 2006. Stock prices, news, and economic fluctuations. Am. Econ. Rev. 96 (4), 1293–1307.

Beaudry, P., Portier, F., 2014. News driven business cycles: insights and challenges. J. Econ. Lit. 52 (4), 993–1074.

Beaudry, P., Fève, P., Guay, A., 2015. When is Nonfundamentalness in VARs a Real Problem? An Application to News Shocks. No. W21466, National Bureau of Economic Research.

Ben Zeev, N., Khan, H., 2015. Investment-specific news shocks and U.S. business cycles. J. Money Credit Bank. 47, 1443–1464.

Ben Zeev, N., Pappa, E., forthcoming. Chronicle of a War Foretold: The Macroeconomic Effects of Anticipated Defense Spending Shocks. Econ. J. http://doi.org/10.1111/ecoj.12349.

Bernanke, B.S., 1986. Alternative explanations of the money-income correlation. Carn.-Roch. Conf. Ser. Public Policy 25, 49–99.

Bernanke, B.S., Blinder, A.S., 1992. The Federal funds rate and the channels of monetary transmission. Am. Econ. Rev. 82 (4), 901–921.

Bernanke, B.S., Mihov, I., 1998. Measuring monetary policy. Q. J. Econ. 113 (3), 869–902.

Bernanke, B.S., Boivin, J., Eliasz, P., 2005. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. Q. J. Econ. 120 (1), 387–422.

Bernstein, J., Romer, C.D., 2009. The job impact of the American Recovery and Reinvestment Plan. Office of the Vice President-Elect. Working paper.

Blanchard, O., Perotti, R., 2002. An empirical characterization of the dynamic effects of changes in government spending and taxes on output. Q. J. Econ. 117, 1329–1368.

Blanchard, O., Quah, D., 1989. The dynamic effects of aggregate demand and supply disturbances. Am. Econ. Rev. 79 (4), 655–673.

Blanchard, O., Watson, M.W., 1986. Are all business cycles alike? In: Gordon, R.J. (Ed.), The American Business Cycle: Continuity and Change. NBER. The University of Chicago Press, Chicago, IL.

Blanchard, O.J., L'Huillier, J.P., Lorenzoni, G., 2013. News, noise, and fluctuations: an empirical exploration. Am. Econ. Rev. 103 (7), 3045–3070.

Boivin, J., 2006. Has U.S. monetary policy changed? Evidence from drifting coefficients and real-time data. J. Money Credit Bank. 38 (5), 1149–1173.

Boivin, J., Giannoni, M.P., 2006. Has monetary policy become more effective? Rev. Econ. Stat. 88 (3), 445–462.

Boivin, J., Kiley, M.T., Mishkin, F.S., 2010. How has the monetary transmission mechanism evolved over time? In: Friedman, B.M., Woodford, M. (Eds.), Handbook of Monetary Economics. Elsevier.

Boschen, J.F., Mills, L.O., 1995. The relation between narrative and money market indicators of monetary policy. Econ. Inq. 33 (1), 24–44.

Burnside, C., Eichenbaum, M., Rebelo, S., 1995. Capital utilization and returns to scale. In: Bernanke, B.S., Rotemberg, J.J. (Eds.), NBER Macroeconomics Annual 1995, vol. 10. MIT Press, Cambridge, MA, pp. 67–124.

Burnside, C., Eichenbaum, M., Fisher, J., 2004. Fiscal shocks and their consequences. J. Econ. Theory 115, 89–117.

Caldara, D., Kamps, C., 2012. The Analytics of SVARs: A Unified Framework to Measure Fiscal Multipliers. Finance and Economics Discussion Series Divisions of Research & Statistics and Monetary Affairs Federal Reserve Board, Washington, DC.

Campbell, J.R., Evans, C.L., Fisher, J.D.M., Justiniano, A., 2012. Macroeconomic effects of federal reserve forward guidance. Brook. Pap. Econ. Act. (Spring), 1–80.

Canova, F., De Nicolo, G., 2002. Monetary disturbances matter for business fluctuations in the G-7. J. Monet. Econ. 49 (6), 1131–1159.

Canova, F., Pina, J.P., 2005. What VAR tell us about DSGE models? In: Diebolt, C., Kyrtsou, C. (Eds.), New Trends in Macroeconomics. Springer, Berlin Heidelberg, pp. 89–123.

Canova, F., Sala, L., 2009. Back to square one: identification issues in DSGE models. J. Monet. Econ. 56, 431–449.

Chang, P.-L., Sakata, S., 2007. Estimation of impulse response functions using long autoregression. Econ. J. 10 (2), 453–469.

Chari, V.V., Kehoe, P.J., McGrattan, E.R., 2008. Are structural VARs with long-run restrictions useful in developing business cycle theory? J. Monet. Econ. 55 (8), 1337–1352.

Christiano, L.J., Eichenbaum, M., 1992. Liquidity effects and the monetary transmission mechanism. Am. Econ. Rev. 82 (2), 346.

Christiano, L.J., Eichenbaum, M., Evans, C.L., 1999. What have we learned and to what end? In: Woodford, M., Taylor, J.D. (Eds.), Handbook of Macroeconomics. Elsevier, Amsterdam.

Christiano, L.J., Eichenbaum, M., Vigfusson, R., 2003. What happens after a technology shock? In: NBER Working Paper Series 9819. National Bureau of Economic Research, Cambridge, MA.

Christiano, L.J., Eichenbaum, M., Evans, C.L., 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. J. Polit. Econ. 113 (1), 1–45.

Clemens, J., Miran, S., 2012. Fiscal policy multipliers on subnational government spending. Am. Econ. J.: Econ. Pol. 4 (2), 46–68.

Cloyne, J., 2013. Discretionary tax changes and the macroeconomy: new narrative evidence from the United Kingdom. Am. Econ. Rev. 103 (4), 1507–1528.

Cochrane, J., 1994. Shocks. Carn.-Roch. Conf. Ser. Public Policy 41, 295–364.

Cochrane, J., Comments on ‘A New Measure of Monetary Shocks: Derivation and Implications’ By Christina Romer and David Romer. July 17, 2004, presented at NBER EFG Meeting. http://faculty.chicagobooth.edu/john.cochrane/research/papers/talk_notes_new_measure_2.pdf.

Cochrane, J., Piazzesi, M., 2002. The fed and interest rates—a high-frequency identification. Am. Econ. Rev. 92 (2), 90–95.

Cogan, J.F., Cwik, T., Taylor, J.B., Wieland, V., 2010. New Keynesian versus old Keynesian government spending multipliers. J. Econ. Dyn. Control. 34 (3), 281–295.

Coibion, O., 2012. Are the effects of monetary policy shocks big or small? Am. Econ. J. Macroecon. 4 (2), 1–32.

Coibion, O., Gorodnichenko, Y., 2011. Monetary policy, trend inflation, and the great moderation: an alternative interpretation. Am. Econ. Rev. 101 (1), 341–370.

Cover, J.P., 1992. Asymmetric effects of positive and negative money-supply shocks. Q. J. Econ. 107 (4), 1261–1282.

D’Amico, S., King, T.B., 2015. What Does Anticipated Monetary Policy Do? Federal Reserve Bank of Chicago Working Paper 2015–10, November 2015.

Davis, S.J., Haltiwanger, J., 2001. Sectoral job creation and destruction responses to oil price changes. J. Monet. Econ. 48 (3), 465–512.

Edelberg, W., Eichenbaum, M.S., Fisher, J.D.M., 1999. Understanding the effects of a shock to government purchases. Rev. Econ. Dynamics. 2 (1), 166–206.

Eichenbaum, M.S., 1992. Comment on interpreting the macroeconomic time series facts: the effects of monetary policy. Eur. Econ. Rev. 36 (5), 1001–1011.

Elliott, G., 1998. On the robustness of cointegration methods when regressors almost have unit roots. Econometrica 66 (1), 149–158.

Erceg, C.J., Guerrieri, L., Gust, C., 2005. Can long-run restrictions identify technology shocks? J. Eur. Econ. Assoc. 3 (6), 1237–1278.

Evans, M.K., 1969. Reconstruction and estimation of the balanced budget multiplier. R. Econ. Stats. 51 (1), 14–25.

Evans, C.L., 1992. Productivity shocks and real business cycles. J. Monet. Econ. 29 (2), 191–208.

Farhi, E., Werning, I., 2012. Fiscal Multipliers: Liquidity Traps and Currency Unions. NBER Working Paper No. 18381, September 2012.

Fatás, A., Mihov, I., 2001. The effects of fiscal policy on consumption and employment: theory and evidence. CEPR Discussion Paper No. 2760.

Faust, J., 1998. The robustness of identified VAR conclusions about money. Carn.-Roch. Conf. Ser. Public Policy 49, 207–244.

Faust, J., Leeper, E.M., 1997. When do long-run identifying restrictions give reliable results? J. Bus. Econ. Stat. 15 (3), 345–353.

Faust, J., Swanson, E.T., Wright, J.H., 2004. Identifying VARS based on high frequency futures data. J. Monet. Econ. 51 (6), 1107–1113.

Favero, C., Giavazzi, F., 2012. Measuring tax multipliers: the narrative method in fiscal VARs. Am. Econ. J. Econ. Pol. 4 (2), 69–94.

Fernald, J.G., 1999. Roads to prosperity? Assessing the link between public capital and productivity. Am. Econ. Rev. 89 (3), 619–638.

Fernald, J.G., 2007. Trend breaks, long-run restrictions, and contractionary technology improvements. J. Monet. Econ. 54 (8), 2467–2485.

Fernald, J.G., 2014. A Quarterly, Utilization-Adjusted Series on Total Factor Productivity. Federal Reserve Bank of San Francisco Working Paper 2012–19, April 2014.

Fernandez-Villaverde, J., Rubio-Ramirez, J., Sargent, T.J., Watson, M.W., 2007. A, B, C's (and D's) of understanding VARs. Am. Econ. Rev. 97 (3), 1021–1026.

Fisher, J.D.M., 2006. The dynamic effects of neutral and investment-specific technology shocks. J. Polit. Econ. 114 (3), 413–451.

Fisher, J.D.M., 2010. Comment on 'Letting Different Views of the Business Cycle Compete'. In: Acemoglu, D., Rogoff, K., Woodford, M. (Eds.), NBER Macroeconomics Annual 2009, vol. 24, University of Chicago Press, Chicago, IL, pp. 457–474.

Fisher, J.D.M., Peters, R., 2010. Using stock returns to identify government spending shocks. Econ. J. 120, 414–436.

Forni, M., Gambetti, L., Sala, L., 2014. No news in business cycles. Econ. J. 124, 1168–1191.

Francis, N., Ramey, V.A., 2006. The source of historical fluctuations: an analysis using long-run restrictions. In: Clarida, R., Frankel, J., Giavazzi, F., West, K. (Eds.), NBER International Seminar on Macroeconomics 2004. The MIT Press, Cambridge, MA, pp. 17–49.

Francis, N., Ramey, V.A., 2005. Is the technology-driven real business cycle hypothesis dead? Shocks and aggregate fluctuations revisited. J. Monet. Econ. 52 (8), 1379–1399.

Francis, N., Ramey, V.A., 2009. Measures of per capita hours and their implications for the technology-hours debate. J. Money Credit Bank. 41 (6), 1071–1097.

Francis, N., Owyang, M.T., Rousch, J.E., DiCecio, R., 2014. A flexible finite-horizon alternative to long-run restrictions with an application to technology shocks. Rev. Econ. Stat. 96 (4), 638–647.

Friedman, M., Schwartz, A., 1963. A Monetary History of the United States: 1867–1960. National Bureau of Economic Research, Princeton University Press, Princeton, NJ.

Frisch, R., 1933. Propagation problems and impulse problems in dynamic economics. In: Economic Essays in Honor of Gustav Cassel. Allen & Unwin, London, pp. 171–205.

Galí, J., 1999. Technology, employment, and the business cycle: do technology shocks explain aggregate fluctuations. Am. Econ. Rev. 89, 249–271.

Galí, J., David López-Salido, J., Vallés, J., 2007. Understanding the effects of government spending on consumption. J. Eur. Econ. Assoc. 5 (1), 227–270.

Gechert, S., 2015. What Fiscal Policy Is Most Effective? A Meta Regression Analysis. Oxford Economic Papers. 67 (3), 553–580.

Gertler, M., Karadi, P., 2015. Monetary policy surprises, credit costs, and economic activity. Am. Econ. J. Macroecon. 7 (1), 44–76.

Giavazzi, F., Pagano, M., 1990. Can severe fiscal consolidations be expansionary? Tales of two small European countries. In: Blanchard, O., Fischer, S. (Eds.), NBER Macroeconomics Annual, Vol. 5. National Bureau of Economic Research, Cambridge, Massachusetts.

Giavazzi, F., Pagano, M., 1996. Non-Keynesian effects of fiscal policy changes: international evidence and the Swedish experience. Swedish Econ. Pol. Rev. 3 (1), 67–103.

Gilchrist, S., Zakrajšek, E., 2012. Credit spreads and business cycle fluctuations. Am. Econ. Rev. 102 (4), 1692–1720.

Gordon, R.J., Krenn, R., 2010. The End of the Great Depression: VAR Insight on the Roles of Monetary and Fiscal Policy. NBER Working Paper 16380, September.

Gospodinov, N., María Herrera, A., Pesavento, E., 2013. Unit roots, cointegration, and pretesting in VAR models. In: Fomby, T.B., Kilian, L., Murphy, A. (Eds.), VAR Models in Macroeconomics—New Developments and Applications: Essays in Honor of Christopher A. Sims. Emerald Group Publishing Limited, pp. 81–115.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37 (3), 424–438.

Greenwood, J., Jovanovic, B., 1999. The information-technology revolution and the stock market. Am. Econ. Rev. 89 (2), 116–122.

Greenwood, J., Hercowitz, Z., Huffman, G.W., 1988. Investment, capacity utilization, and the real business cycle. Am. Econ. Rev. 78 (3), 402–417.

Greenwood, J., Hercowitz, Z., Krusell, P., 2000. The role of investment-specific technological change in the business cycle. Eur. Econ. Rev. 44 (1), 91–115.

Gürkaynak, R.S., Sack, B., Swanson, E., 2005. The sensitivity of long-term interest rates to economic news: evidence and implications for macroeconomic models. Am. Econ. Rev. 95 (1), 425–436.

Hall, R.E., 1980. Labor supply and aggregate fluctuations. In: Carnegie-Rochester Conference Series on Public Policy, vol. 12, North-Holland.

Hall, R.E., 1986. The role of consumption in economic fluctuations. In: Gordon, R.J. (Ed.), The American Business Cycle: Continuity and Change. NBER, University of Chicago Press, Chicago, IL, pp. 237–266.

Hall, R.E., 1988. The relation between price and marginal cost in U.S. industry. J. Polit. Econ. 96 (5), 921–947.

Hall, R.E., 1990. Invariance properties of Solow's productivity residual. In: Diamond, P. (Ed.), Growth/Productivity/Unemployment: Essays to Celebrate Bob Solow's Birthday. MIT Press, Cambridge, MA, pp. 71–112.

Hall, R.E., 2009. By how much does GDP rise if the government buys more output? Brook. Pap. Econ. Act. 2, 183–231.

Hamilton, J.D., 1983. Oil and the macroeconomy since World War II. J. Polit. Econ. 91 (2), 228–248.

Hamilton, J.D., 1985. Historical causes of postwar oil shocks and recessions. Energy J. 6 (1), 97–116.

Hamilton, J.D., 2003. What is an oil shock? J. Econ. 113 (2), 363–398.

Hamilton, J.D., 2010. Macroeconomics and ARCH. In: Bollerslev, T., Russell, J., Watson, M. (Eds.), Volatility and Time Series Econometrics: Essays in Honor of Robert Engle. Oxford University Press, Oxford, pp. 79–96.

Hansen, L.P., Sargent, T.J., 1991. Two difficulties in interpreting vector autoregressions. In: Hansen, L.P., Sargent, T.J. (Eds.), Rational Expectations Econometrics. Westview Press, Boulder, CO, pp. 77–119.

Hanson, S.G., Stein, J.C., 2015. Monetary policy and long-term real rates. J. Financ. Econ. 115 (3), 429–448.

Hausman, J.K., 2016. Fiscal policy and economic recovery: the case of the 1936 Veterans' bonus. Am. Econ. Rev. 106 (4), 1100–1143.

Hobijn, B., Jovanovic, B., 2001. The information-technology revolution and the stock market: evidence. Am. Econ. Rev. 91 (5), 1203–1220.

Hoover, K.D., Perez, S.J., 1994. Post hoc ergo propter once more an evaluation of 'does monetary policy matter?' In the spirit of James Tobin. J. Monet. Econ. 34 (1), 47–74.

Jordà, Ò., 2005. Estimation and inference of impulse responses by local projections. Am. Econ. Rev. 95 (1), 161–182.

Justiniano, A., Primiceri, G.E., Tambalotti, A., 2010. Investment shocks and business cycles. J. Monet. Econ. 57 (2), 132–145.

Justiniano, A., Primiceri, G.E., Tambalotti, A., 2011. Investment shocks and the relative price of investment. Rev. Econ. Dyn. 14 (1), 102–121.

Keynes, J.M., 1936. The General Theory of Employment, Interest and Money. Macmillan, London.

Khan, H., Tsoukalas, J., 2012. The quantitative importance of news shocks in estimated DSGE models. J. Money Credit Bank. 44 (8), 1535–1561.

Kilian, L., 2009. Not all oil price shocks are alike: disentangling demand and supply shocks in the crude oil market. Am. Econ. Rev. 99 (3), 1053–1069.

Kilian, L., Vigfusson, R.J., 2011. Are the responses of the US economy asymmetric in energy price increases and decreases? Quant. Econ. 2 (3), 419–453.

King, R., Plosser, C., Stock, J., Watson, M.W., 1991. Stochastic trends and economic fluctuations. Am. Econ. Rev. 81 (4), 819–840.

Kliem, M., Kriwoluzky, A., 2013. Reconciling narrative monetary policy disturbances with structural VAR model shocks? Econ. Lett. 121 (2), 247–251.

Komunjer, I., Ng, S., 2011. Dynamic identification of DSGE models. Econometrica 79 (6), 1995–2032.

Koop, G., Hashem Pesaran, M., Potter, S.M., 1996. Impulse response analysis in nonlinear multivariate models. J. Econ. 74 (1), 119–147.

Krishnamurthy, A., Vissing-Jorgensen, A., 2011. The effects of quantitative easing on interest rates. Brook. Pap. Econ. Acti. Fall, 215–287.

Kurmann, A., Otrok, C., 2013. News shocks and the slope of the term structure of interest rates. Am. Econ. Rev. 103 (6), 2612–2632.

Kurmann, A., Mertens, E., 2014. Stock prices, news, and economic fluctuations: comment. Am. Econ. Rev. 104 (4), 1439–1445.

Kuttner, K.N., 2001. Monetary policy surprises and interest rates: evidence from the Fed funds futures market. J. Monet. Econ. 47 (3), 523–544.

Kydland, F.E., Prescott, E.C., 1982. Time to build and aggregate fluctuations. Econometrica 50 (6), 1345–1370.

Lakdawala, A., 2015. Changes in Federal Reserve Preferences. Michigan State University Working Paper, April 2015.

Leduc, Sylvain, Wilson, Daniel, 2013. Roads to prosperity or bridges to nowhere? Theory and evidence on the impact of public infrastructure investment. In: Acemoglu, D., Parker, J., Woodford, M. (Eds.), NBER Macroecon. Annu. 2012. Vol. 27, pp. 89–142.

Leeper, E.M., 1997. Narrative and VAR approaches to monetary policy: common identification problems. J. Monet. Econ. 40, 641–657.

Leeper, E.M., Richter, A., Walker, T.B., 2012. Quantitative effects of fiscal foresight. Am. Econ. J. Econ. Pol. 4 (2), 1–27.

Leeper, E.M., Walker, T.B., Susan Yang, S.-C., 2013. Fiscal foresight and information flows. Econometrica 81 (3), 1115–1145. Also, unpublished supplement at: https://www.econometricsociety.org/sites/default/files/8337_extensions_0.pdf.

Leigh, D., Devries, P., Freedman, C., Guajardo, J., Laxton, D., Pescatori, A., 2010. Will it hurt? Macroeconomic effects of fiscal consolidation. World Economic Outlook, IMF, October 2010 (Chapter 3).

Litterman, R.B., Weiss, L., 1985. Money, real interest rates, and output: a reinterpretation of post-war data. Econometrica 53 (1), 129–156.

Lundsford, K.G., 2015. Identifying Structural VARs with a Proxy Variable and a Test for a Weak Proxy. Federal Reserve Bank of Cleveland Working Paper, December 15–28, 2015.

Marcellino, M., Stock, J.H., Watson, M.W., 2006. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. J. Econ. 135 (1), 499–526.

McCallum, B.T., 1983. A reconsideration of Sims' evidence concerning monetarism. Econ. Lett. 13 (2–3), 167–171.

McGrattan, E.R., 1994. The macroeconomic effects of distortionary taxation. J. Monet. Econ. 33 (3), 573–601.

Mertens, K., Ravn, M.O., 2011a. Technology-hours redux: tax changes and the measurement of technology shocks. In: NBER International Seminar on Macroeconomics 2010.

Mertens, K., Ravn, M.O., 2011b. Understanding the aggregate effects of anticipated and unanticipated tax policy shocks. Rev. Econ. Dyn. 14 (1), 27–54.

Mertens, K., Ravn, M.O., 2012. Empirical evidence on the aggregate effects of anticipated and unanticipated US tax policy shocks. Am. Econ. J. Econ. Pol. 4 (2), 145–181.

Mertens, K., Ravn, M.O., 2013. The dynamic effects of personal and corporate income tax changes in the United States. Am. Econ. Rev. 103 (4), 1212–1247.

Mertens, K., Ravn, M.O., 2014. A reconciliation of SVAR and narrative estimates of tax multipliers. J. Monet. Econ. 68, S1–S19.

Miyamoto, W., Nguyen, T.L., 2015. News Shocks and Business Cycles: Evidence from Forecast Data. Santa Clara University working paper.

Montiel Olea, J.L., Pflueger, J., 2013. A robust test for weak instruments. J. Bus. Econ. Stat. 31 (3), 358–369.

Montiel Olea, J.L., Stock, J.H., Watson, M.W., 2015. Uniform Inference in SVARs with External Instruments. December 2015 manuscript.

Mountford, A., Uhlig, H., 2009. What are the effects of fiscal policy shocks? J. Appl. Econ. 24, 960–992.

Nakamura, E., Steinsson, J., 2014. Fiscal stimulus in a monetary union: evidence from US regions. Am. Econ. Rev. 104 (3), 753–792.

Nakamura, E., Steinsson, J., 2015. High Frequency Identification of Monetary Non-Neutrality. October 2015 Working Paper.

Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55 (3), 703–708.

Oh, H., Reis, R., 2012. Targeted transfers and the fiscal response to the great recession. J. Monet. Econ. 59, S50–S64.

Olivei, G., Tenreyro, S., 2007. The timing of monetary policy shocks. Am. Econ. Rev. 97, 636–663.

Owyang, M., Ramey, G., 2004. Regime switching and monetary policy measurement. J. Monet. Econ. 51 (8), 1577–1597.

Owyang, M.T., Ramey, V.A., Zubairy, S., 2013. Are government spending multipliers greater during periods of slack? Evidence from twentieth-century historical data. Am. Econ. Rev. 103 (3), 129–134.

Pappa, E., 2009. The effects of fiscal shocks on employment and the real wage. Int. Econ. Rev. 50 (1), 217–244.

Perotti, R., 2005. Estimating the effects of fiscal policy in OECD countries. CEPR Discussion Paper 4842, January.

Perotti, R., 2011. Expectations and Fiscal Policy: An Empirical Investigation. Bocconi Working Paper.

Perron, P., 1989. The great crash, the oil price shock, and the unit root hypothesis. Econometrica 57 (6), 1361–1401.

Pflueger, C.E., Wang, S., 2015. A robust test for weak instruments in Stata. Stata J. 15 (1), 216–225.

Piazzesi, M., Swanson, E.T., 2008. Futures prices as risk-adjusted forecasts of monetary policy. J. Monet. Econ. 55 (4), 677–691.

Pigou, A.C., 1927. Industrial Fluctuations. Macmillan, London.

Poterba, J.M., 1986. Explaining the yield spread between taxable and tax-exempt bonds: the role of expected tax policy. In: Rosen, H.S. (Ed.), Studies in State and Local Public Finance. University of Chicago Press, pp. 5–52.

Prescott, E.C., 1986. Theory ahead of business-cycle measurement. Carn.-Roch. Conf. Ser. Public Policy 25, 11–44.

Primiceri, G., 2005. Time varying structural vector autoregressions and monetary policy. Rev. Econ. Stud. 72 (3), 821–852.

Ramey, V.A., 2009. Identifying Government Spending Shocks: It's All in the Timing. NBER Working Paper No. 15464, October 2009.

Ramey, V.A., 2011a. Identifying government spending shocks: it's all in the timing. Q. J. Econ. 126, 1–50.

Ramey, V.A., 2011b. Can government purchases stimulate the economy? J. Econ. Lit. 49 (3), 673–685.

Ramey, V.A., 2013. Government spending and private activity. In: Alesina, A., Giavazzi, F. (Eds.), Fiscal Policy After the Financial Crisis. University of Chicago Press, Chicago, IL.

Ramey, V.A., Shapiro, M., 1998. Costly capital reallocation and the effects of government spending. Carn.-Roch. Conf. Ser. Public Policy 48, 145–194.

Ramey, V.A., Vine, D.J., 2011. Oil, automobiles, and the US economy: how much have things really changed? In: Acemoglu, D., Woodford, M. (Eds.), NBER Macroeconomics Annual 2010, Vol. 25. University of Chicago Press, pp. 333–367.

Ramey, V.A., Zubairy, S., 2014. Government Spending Multipliers in Good Times and in Bad: Evidence from 20th Century Historical Data. November 2014 Working Paper.

Romer, C.D., Romer, D.H., 1989. Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz. NBER Macroeconomic Annual 1989.

Romer, C.D., Romer, D.H., 1997. Identification and the narrative approach: a reply to Leeper. J. Monet. Econ. 40, 659–665.

Romer, C.D., Romer, D.H., 2000. Federal reserve information and the behavior of interest rates. Am. Econ. Rev. 90 (3), 429–457.

Romer, C.D., Romer, D.H., 2004. A new measure of monetary policy shocks: derivation and implications. Am. Econ. Rev. 94 (4), 1055–1084.

Romer, C.D., Romer, D.H., 2010. The macroeconomic effects of tax changes: estimates based on a new measure of fiscal shocks. Am. Econ. Rev. 100, 763–801.

Romer, C.D., Romer, D.H., 2016. Transfer payments and the macroeconomy: the effects of social security benefit increases, 1952-1991. Berkeley working paper, March.

Rotemberg, J., Woodford, M., 1992. Oligopolistic pricing and the effects of aggregate demand on economic activity. J. Polit. Econ. 100, 1153–1297.

Rudebusch, G., 1998. Do measures of monetary policy in a VAR make sense? In: Symposium on Forecasting and Empirical Methods in Macroeconomics and Finance. International Economic Review, vol. 39(4), pp. 907–931.

Schmitt-Grohe, S., Uribe, M., 2012. What's news in business cycles? Econometrica 80 (6), 2733–2764.

Shapiro, M.D., 1993. Cyclical productivity and the workweek of capital. Am. Econ. Rev. 83 (2), 229–233.

Shapiro, M.D., 1994. Federal reserve policy: cause and effect. In: Gregory Mankiw, N. (Ed.), Monetary Policy. National Bureau of Economic Research, The University of Chicago Press, Chicago, IL.

Shapiro, M.D., Watson, M.W., 1988. The sources of business cycle fluctuations. In: Fischer, S. (Ed.), NBER Macroeconomics Annual, vol. 3. MIT Press, Cambridge, MA.

Sims, C.A., 1972. Money, income, and causality. Am. Econ. Rev. 62 (4), 540–552.

Sims, C.A., 1980a. Macroeconomics and reality. Econometrica 48, 1–48.

Sims, C.A., 1980b. Comparison of interwar and postwar business cycles: monetarism reconsidered. Am. Econ. Rev. 70 (2), 250–257.

Sims, C.A., 1992. Interpreting the macroeconomic time series facts: the effects of monetary policy. Eur. Econ. Rev. 36 (5), 975–1000.

Sims, C.A., 1998. Discussion of Glenn Rudebusch, "Do measures of monetary policy in a VAR make sense?" In: Symposium on Forecasting and Empirical Methods in Macroeconomics and Finance (Nov. 1998). International Economic Review, vol. 39, pp. 907–931.

Sims, C.A., Zha, T., 2006a. Were there Regime Switches in U.S. monetary policy? Am. Econ. Rev. 96 (1), 54–81.

Sims, C.A., Zha, T., 2006b. Does monetary policy generate recessions? Macroecon. Dyn. 10 (2), 231–272.

Sims, C.A., Stock, J.H., Watson, M.W., 1990. Inference in linear time series models with some unit roots. Econometrica 58 (1), 113–144.

Slutsky, E., 1937. The summation of random causes as the source of cyclic processes. Econometrica 5 (2), 105–146.

Smets, F., Wouters, R., 2003. An estimated dynamic stochastic general equilibrium model of the Euro area. J. Eur. Econ. Assoc. 1 (5), 1123–1175.

Smets, F., Wouters, R., 2007. Shocks and frictions in U.S. business cycles: a Bayesian DSGE approach. Am. Econ. Rev. 97 (3), 586–606.

Solow, R.M., 1957. Technical change and the aggregate production function. Rev. Econ. Stat. 39 (3), 312–320.

Stock, J.H., Watson, M.W., 2002. Forecasting using principal components from a large number of predictors. J. Am. Stat. Assoc. 97 (460), 1167–1179.

Stock, J.H., Watson, M.W., 2008. NBER Summer Institute Minicourse 2008: What's New in Econometrics—Time Series, Lecture 7: Structural VARs. National Institute for Economic Research, Cambridge, MA. www.nber.org/minicourse_2008.html.

Stock, J.H., Watson, M.W., 2012. Disentangling the channels of the 2007–09 recession. Brook. Pap. Econ. Act. 2012 (Spring), 81–135.

Stock, J.H., Watson, M.W., 2016. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In: Taylor, J.B., Uhlig, H. (Eds.), Handbook of Macroeconomics, vol. 2A. Elsevier, Amsterdam, Netherlands, pp. 415–525.

Stock, J.H., Wright, J.H., Yogo, M., 2002. A survey of weak instruments and weak identification in generalized method of moments. J. Bus. Econ. Stat. 20 (4), 518–529.

Strongin, S., 1995. The identification of monetary policy disturbances explaining the liquidity puzzle. J. Monet. Econ. 35 (3), 463–497.

Tenreyro, S., Thwaites, G., forthcoming. Pushing on a String: US Monetary Policy is Less Powerful in Recessions. Am. Econ. J.: Macro.

Thoma, M.A., 1994. Subsample instability and asymmetries in money-income causality. J. Econ. 64 (1–2), 279–306.

Uhlig, H., 1997. What are the effects of monetary policy on output? Results from an agnostic identification procedure. Tilburg University Manuscript.

Uhlig, H., 2003. What drives GNP? Unpublished manuscript, Euro Area Business Cycle Network.

Uhlig, H., 2004. Do technology shocks lead to a fall in total hours worked? J. Eur. Econ. Assoc. 2 (2–3), 361–371.

Uhlig, H., 2005. What are the effects of monetary policy on output? Results from an agnostic identification procedure. J. Monet. Econ. 52 (2), 381–419.

Uhlig, H., 2010. Some fiscal calculus. Am. Econ. Rev. 100 (2), 30–34.

Velde, F.R., 2009. Chronicle of a deflation unforetold. J. Polit. Econ. 117 (4), 591–634.

Weise, C.L., 1999. The asymmetric effects of monetary policy: a nonlinear vector autoregression approach. J. Money, Credit, Bank. 31 (1), 85–108.

Wieland, J., Yang, M.-J., 2015. Financial Dampening. September 2015 Working Paper.

Wu, J.C., Xia, F.D., 2016. Measuring the macroeconomic impact of monetary policy at the zero lower bound. J. Money, Credit, Bank. 48 (2–3), 253–291.

Yang, S.-C.S., 2005. Quantifying tax effects under policy foresight. J. Monet. Econ. 52 (8), 1557–1568.

CHAPTER 3

Macroeconomic Regimes and Regime Shifts

J.D. Hamilton
University of California, San Diego, La Jolla, CA, United States

Contents

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.004

Abstract

Many economic time series exhibit dramatic breaks associated with events such as economic recessions, financial panics, and currency crises. Such changes in regime may arise from tipping points or other nonlinear dynamics and are core to some of the most important questions in macroeconomics. This chapter surveys the literature for studying regime changes and summarizes available methods. Section 1 introduces some of the basic tools for analyzing such phenomena, using for illustration the move of an economy into and out of recession. Section 2 focuses on empirical methods, providing a detailed overview of econometric analysis of time series that are subject to changes in regime. Section 3 discusses theoretical treatment of macroeconomic models with changes in regime and reviews applications in a number of areas of macroeconomics. Some brief concluding recommendations for applied researchers are offered in Section 4.

Keywords

Changes in regime, Markov switching, EM algorithm, Nonlinear macroeconomic dynamics, Vector autoregressions

JEL Codes

C32, E32, E37

1. INTRODUCTION: ECONOMIC RECESSIONS AS CHANGES IN REGIME

Fig. 1 plots the US unemployment rate since World War II. Shaded regions highlight a feature of the data that is very familiar to macroeconomists—periodically the US economy enters an episode in which the unemployment rate rises quite rapidly. These shaded regions correspond to periods that the Dating Committee of the National Bureau of Economic Research chose to designate as economic recessions. But what exactly does such a designation signify?

One view is that the statement that the economy has entered a recession does not have any intrinsic objective meaning. According to this view, the economy is always subject to unanticipated shocks, some favorable, others unfavorable. A recession is then held to be nothing more than a string of unusually bad shocks, with the bifurcation of the observed sample into periods of "recession" and "expansion," an essentially arbitrary way of summarizing the data.

Such a view is implicit in many theoretical models used in economics today insofar as it is a necessary implication of the linearity we often assume in order to make our models more tractable. But the convenience of linear models is not a good enough reason to assume that no fundamental changes in economic dynamics occur when the economy goes into a recession. For example, we understand reasonably well that in an expansion, GDP will rise more quickly at some times than others, depending on the pace of new technological innovations. But what exactly would we mean by a negative technology shock? The assumption that such events are just like technological improvements but

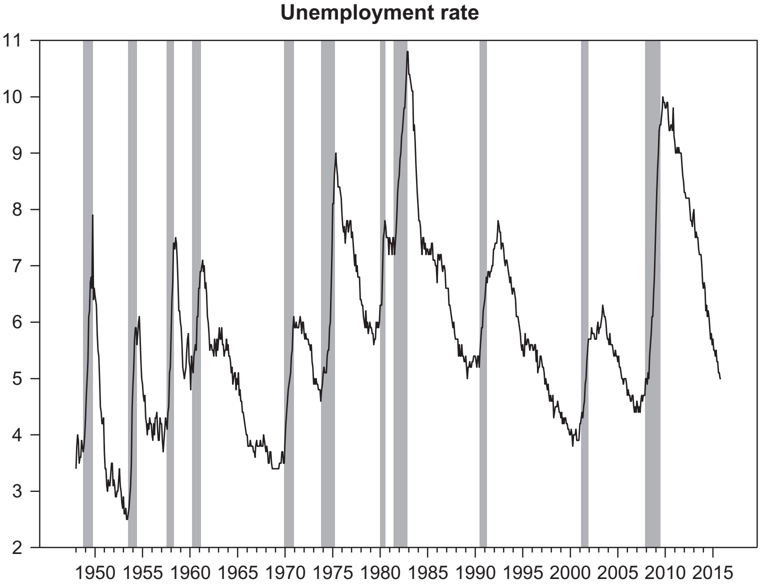


Fig. 1 US civilian unemployment rate, seasonally adjusted, 1948:M1–2015:M11. *Shaded regions* correspond to NBER recession dates.

with a negative sign does not seem like the place we should start if we are trying to understand what really happens during an economic downturn.

An alternative view is that on occasion some forces that are very different from the usual technological growth take over to determine employment and output, resulting for example when a simultaneous drop in product demand across different sectors and a rapid increase in unemployed workers introduce new feedbacks of their own. The idea that there might be a tipping point at which different economic dynamics begin to take over will be a recurrent theme in this chapter.

Let us begin with a very simple model with which we can explore some of the issues. We could represent the possibility that there are two distinct phases for the economy using the random variable s_t. When $s_t = 1$, the economy is in expansion in period t and when $s_t = 2$, the economy is in recession. Suppose that an observed variable y_t such as GDP growth has an average value of $m_1 > 0$ when $s_t = 1$ and average value $m_2 < 0$ when $s_t = 2$, as in

$$y_t = m_{s_t} + \varepsilon_t \tag{1}$$

where $\varepsilon_t \sim$ i.i.d. $N(0,\sigma^2)$. Suppose that the transition between regimes is governed by a Markov chain that is independent of ε_t,

$$\text{Prob}(s_t = j | s_{t-1} = i, s_{t-2} = k, \ldots, y_{t-1}, y_{t-2}, \ldots) = p_{ij} \qquad i, j = 1, 2. \tag{2}$$

Note that if both s_t and ε_t were observed directly, (1)–(2) in fact could still be described as a linear process. We can verify directly from (2) that[a]

$$E(m_{s_t}|m_{s_{t-1}}, m_{s_{t-2}}, \ldots) = a + \phi m_{s_{t-1}} \tag{3}$$

where $a = p_{21}m_1 + p_{12}m_2$ and $\phi = p_{11} - p_{21}$. In other words, m_{s_t} follows an AR(1) process,

$$m_{s_t} = a + \phi m_{s_{t-1}} + v_t. \tag{4}$$

The innovation v_t can take on only one of four possible values (depending on the realization of s_t and s_{t-1}) but by virtue of (3), v_t can be characterized as a martingale difference sequence.

Suppose however that we do not observe s_t and ε_t directly, but only have observations of GDP up through date $t-1$ (denoted $\Omega_{t-1} = \{y_{t-1}, y_{t-2}, \ldots\}$) and want to forecast the value of y_t. Notice from Eq. (1) that y_t is the sum of an AR(1) process (namely (4)) and a white noise process ε_t. Recall (eg, Hamilton, 1994, p. 108) that the result could be described as an ARMA(1,1) process. Thus the linear projection of GDP on its own lagged values is given by

$$\hat{E}(y_t|\Omega_{t-1}) = a + \phi y_{t-1} + \theta[y_{t-1} - \hat{E}(y_{t-1}|\Omega_{t-2})] \tag{5}$$

where θ is a known function of ϕ, σ^2, and the variance of v_t (Hamilton, 1994, eq. [4.7.12]). Note that we are using the notation $\hat{E}(y_t|\Omega_{t-1})$ to denote a linear projection (the forecast that produces the smallest mean squared error among the class of all linear functions of Ω_{t-1}) to distinguish it from the conditional expectation $E(y_t|\Omega_{t-1})$ (the forecast that produces the smallest mean squared error among the class of all functions of Ω_{t-1}).

Because of the discrete nature of s_t, the linear projection (5) would not yield the optimal forecast of GDP. We can demonstrate this using the law of iterated expectations (White, 1984, p. 54):

$$\begin{aligned} E(y_t|\Omega_{t-1}) &= \sum_{i=1}^{2} E(y_t|s_{t-1} = i, \Omega_{t-1})\text{Prob}(s_{t-1} = i|\Omega_{t-1}) \\ &= \sum_{i=1}^{2} (a + \phi m_i)\text{Prob}(s_{t-1} = i|\Omega_{t-1}). \end{aligned} \tag{6}$$

Because a probability is necessarily between 0 and 1, the optimal inference $\text{Prob}(s_{t-1} = i|\Omega_{t-1})$ is necessarily a nonlinear function of Ω_{t-1}. If data through $t-1$ have persuaded us that the economy was in expansion at that point, the optimal forecast is going to be close

[a] That is,

$$\begin{aligned} E(m_{s_{t+1}}|m_{s_t} = m_1) &= p_{11}m_1 + p_{12}m_2 = p_{11}m_1 + a - p_{21}m_1 = a + \phi m_1 \\ E(m_{s_{t+1}}|m_{s_t} = m_2) &= p_{21}m_1 + p_{22}m_2 = a - p_{12}m_2 + p_{22}m_2 = a - (1 - p_{11})m_2 + (1 - p_{21})m_2 = a + \phi m_2. \end{aligned}$$

to $a + \phi m_1$, whereas if we become convinced the economy was in recession, the optimal forecast approaches $a + \phi m_2$. It is in this sense that we could characterize (1) as a nonlinear process in terms of its observable implications for GDP.

Calculation of the nonlinear inference $\text{Prob}(s_{t-1} = i|\Omega_{t-1})$ is quite simple for this process. We could start for $t = 0$ for example with the ergodic probabilities of the Markov chain:

$$\text{Prob}(s_0 = 1|\Omega_0) = \frac{p_{21}}{p_{21} + p_{12}}$$

$$\text{Prob}(s_0 = 2|\Omega_0) = \frac{p_{12}}{p_{21} + p_{12}}.$$

Given a value for $\text{Prob}(s_{t-1} = i|\Omega_{t-1})$, we can arrive at the value for $\text{Prob}(s_t = j|\Omega_t)$ using Bayes's law:

$$\text{Prob}(s_t = j|\Omega_t) = \frac{\text{Prob}(s_t = j|\Omega_{t-1}) f(y_t|s_t = j, \Omega_{t-1})}{f(y_t|\Omega_{t-1})}. \tag{7}$$

Here $f(y_t|s_t = j, \Omega_{t-1})$ is the $N(m_j, \sigma^2)$ density,

$$f(y_t|s_t = j, \Omega_{t-1}) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[\frac{-(y_t - m_j)^2}{2\sigma^2}\right], \tag{8}$$

$\text{Prob}(s_t = j|\Omega_{t-1})$ is the predicted regime given past observations,

$$\text{Prob}(s_t = j|\Omega_{t-1}) = p_{1j}\text{Prob}(s_{t-1} = 1|\Omega_{t-1}) + p_{2j}\text{Prob}(s_{t-1} = 2|\Omega_{t-1}), \tag{9}$$

and $f(y_t|\Omega_{t-1})$ is the predictive density for GDP:

$$f(y_t|\Omega_{t-1}) = \sum_{i=1}^{2} \text{Prob}(s_t = i|\Omega_{t-1}) f(y_t|s_t = i, \Omega_{t-1}). \tag{10}$$

Given a value for $\text{Prob}(s_{t-1} = i|\Omega_{t-1})$, we can thus use (7) to calculate $\text{Prob}(s_t = j|\Omega_t)$, and proceed iteratively in this fashion through the data for $t = 1, 2, \ldots, T$ to calculate the necessary magnitude for forming the optimal nonlinear forecast given in (6).

Note that another by-product of this recursion is calculation in (10) of the predictive density for the observed data. Thus one could estimate the vector of unknown population parameters $\boldsymbol{\lambda} = (m_1, m_2, \sigma, p_{11}, p_{22})'$ by maximizing the log-likelihood function of the observed sample of GDP growth rates,

$$\mathcal{L}(\boldsymbol{\lambda}) = \sum_{t=1}^{T} \log f(y_t|\Omega_{t-1}; \boldsymbol{\lambda}). \tag{11}$$

If the objective is to form an optimal inference about when the economy was in a recession, one can use the same principles to obtain an even better inference as more data accumulate. For example, an inference using data observed through date $t + k$ about the regime at date t is known as a k-period-ahead smoothed inference,

$$\text{Prob}(s_t = i|\boldsymbol{\Omega}_{t+k}),$$

calculation of which will be explained in Eq. (22).

Though this is a trivially simple model, it seems to do a pretty good job at capturing what is being described by the NBER's business cycle chronology. If we select only those quarters for which the NBER declared the US economy to be in expansion, we calculate an average annual growth rate of 4.5%, suggesting a value for the parameter $m_1 = 4.5$. And we observe that if the NBER determined the economy to be in expansion in quarter t, 95% of the time it said the same thing in quarter $t + 1$, consistent with a value of $p_{11} = 0.95$. These values implied by the NBER chronology are summarized in column 3 of Table 1. On the other hand, if we ignore the NBER dates altogether, but simply maximize the log likelihood (11) of the observed GDP data alone, we end up with very similar estimates, as seen in column 4.

Moreover, even given the challenges of data revision, the one-quarter-ahead smoothed probabilities have an excellent out-of-sample record at tracking the NBER dates. Fig. 2 plots historical values for $\text{Prob}(s_t = 2|\boldsymbol{\Omega}_{t+1}, \hat{\boldsymbol{\lambda}}_{t+1})$ where only GDP data as they were actually released as of date $t + 1$ were used to estimate parameters and form the inference plotted for date t. Values before the vertical line are "simulated real-time" inferences from Chauvet and Hamilton (2006), that is, values calculated in 2005 using a separate historical real-time data vintage for each date t shown. Values after the vertical line are true real-time out-of-sample inferences as they have been published individually on www.econbrowser.com each quarter since 2005 without revision.

One attractive feature of this approach is that the linearity of the model conditional on s_t makes it almost as tractable as a fully linear model. For example, an optimal

Table 1 Parameter values for describing U.S. recessions

Parameter (1)	Interpretation (2)	Value from NBER classifications (3)	Value from GDP alone (4)
m_1	Average growth in expansion	4.5	4.62
m_2	Average growth in recession	−1.2	−0.48
σ	Standard deviation of growth	3.5	3.34
p_{11}	Prob. expansion continues	0.95	0.92
p_{22}	Prob. recession continues	0.78	0.74

Parameter estimates based on characteristics of expansions and recessions as classified by NBER (column 3), and values that maximize the observed sample log likelihood of postwar GDP growth rates (column 4), 1947:Q2–2004:Q2.
Source: Chauvet, M., Hamilton, J.D., 2006. Dating business cycle turning points. In: Costas Milas, P.R., van Dijk, D. (Eds.), Nonlinear Analysis of Business Cycles. Elsevier, Amsterdam, pp. 1–54.

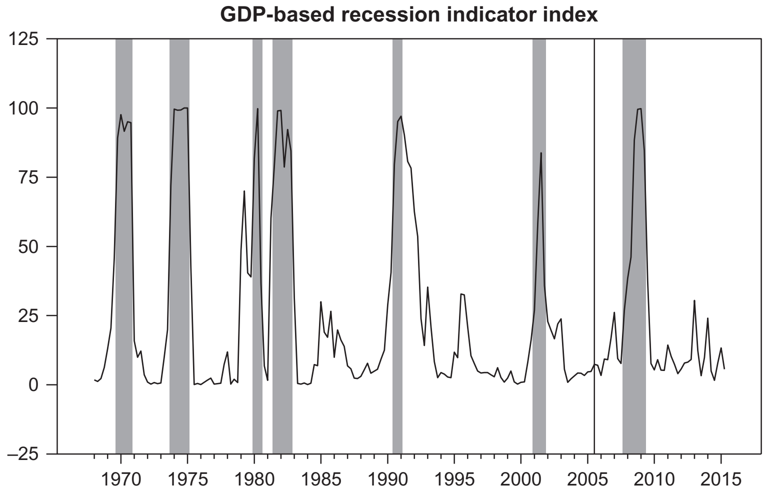


Fig. 2 One-quarter-ahead smoothed probabilities Prob$(s_t = 2|\Omega_{t+1}, \hat{\lambda}_{t+1})$, 1967:Q4–2014:Q2, as inferred using solely GDP data as reported as of date t + 1. *Shaded regions* correspond to NBER recession dates which were not used in any way in constructing the probabilities. Prior to 2005, each point on the graph corresponds to a simulated real-time inference that was constructed from a data set as it would have been available 4 months after the indicated date, as reported in Chauvet and Hamilton (2006). After 2005, points on the graph correspond to actual announcements that were publicly released 4 months after the indicated date. *Source: Updated from Hamilton, J.D., 2011. Calling recessions in real time. Int. J. Forecast. 27, 1006–1026 and www.econbrowser.com.*

k-period-ahead forecast of GDP growth based only on observed growth through date t can be calculated immediately using (4),

$$E(y_{t+k}|\Omega_t) = \mu + \phi^k \sum_{i=1}^{2} (m_i - \mu)\,\text{Prob}(s_t = i|\Omega_t) \tag{12}$$

for $\mu = a/(1-\phi)$. Results like this make this model of changes in regime very convenient to work with.

2. ECONOMETRIC TREATMENT OF CHANGES IN REGIME

This section discusses econometric inference for data that may be subject to changes in regime, while Section 3 examines methods to incorporate changes in regime into theoretical economic models.

2.1 Multivariate or Non-Gaussian Processes

Although the model in Section 1 was quite stylized, the same basic principles can be used to investigate changes in regime in much richer settings. Suppose we have a vector of variables $\mathbf{y}_t$ observed at date t and hypothesize that the density of $\mathbf{y}_t$ conditioned on its past history $\Omega_{t-1} = \{\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots\}$ depends on parameters $\boldsymbol{\theta}$, some or all of which are different depending on the regime s_t:

$$f(\mathbf{y}_t | s_t = i, \Omega_{t-1}) = f(\mathbf{y}_t | \Omega_{t-1}; \boldsymbol{\theta}_i) \quad \text{for } i = 1, \ldots, N. \tag{13}$$

In the example in Section 1, there were $N = 2$ possible regimes with $\boldsymbol{\theta}_1 = (m_1, \sigma)'$, $\boldsymbol{\theta}_2 = (m_2, \sigma)'$, and $f(\mathbf{y}_t | \Omega_{t-1}; \boldsymbol{\theta}_i)$ the $N(m_i, \sigma^2)$ density. But the same basic approach would work for an n-dimensional vector autoregression in which some or all of the parameters change with the regime,

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\Phi}_{s_t,1}\mathbf{y}_{t-1} + \boldsymbol{\Phi}_{s_t,2}\mathbf{y}_{t-2} + \cdots + \boldsymbol{\Phi}_{s_t,r}\mathbf{y}_{t-r} + \mathbf{c}_{s_t} + \boldsymbol{\varepsilon}_t \\ &= \boldsymbol{\Phi}_{s_t}\mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t \end{aligned} \tag{14}$$

$$\boldsymbol{\varepsilon}_t | s_t, \Omega_{t-1} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{s_t}), \tag{15}$$

a class of models discussed in detail in Krolzig (1997). Here $\mathbf{x}_{t-1}$ is an $(nr + 1) \times 1$ vector consisting of a constant term and r lags of $\mathbf{y}$:

$$\mathbf{x}_{t-1} = (\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \ldots, \mathbf{y}'_{t-r}, 1)'.$$

In this case the density of $\mathbf{y}_t$ conditional on its own past values and the regime s_t taking the value i would be

$$f(\mathbf{y}_t | s_t = i, \Omega_{t-1}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-(1/2)(\mathbf{y}_t - \boldsymbol{\Phi}_i\mathbf{x}_{t-1})'\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_t - \boldsymbol{\Phi}_i\mathbf{x}_{t-1})\right]. \tag{16}$$

There is also no reason that a Gaussian density has to be used. For example, Dueker (1997) proposed a model of stock returns in which the innovation comes from a Student t distribution whose degrees of freedom parameter η changes with the regime.

2.2 Multiple Regimes

A convenient representation for a model with $N > 2$ regimes is obtained by collecting the transition probabilities in a matrix $\mathbf{P}$ whose row j, column i element corresponds to p_{ij} (so that columns of $\mathbf{P}$ sum to unity). We likewise summarize the regime at date t by an $(N \times 1)$ vector $\boldsymbol{\xi}_t$ whose ith element is unity when $s_t = i$ and is zero otherwise—in other words, $\boldsymbol{\xi}_t$ corresponds to column s_t of $\mathbf{I}_N$. Notice that $E(\boldsymbol{\xi}_t | s_{t-1} = i)$ has the interpretation

$$E(\boldsymbol{\xi}_t | s_{t-1} = i) = \begin{bmatrix} \text{Prob}(s_t = 1 | s_{t-1} = i) \\ \vdots \\ \text{Prob}(s_t = N | s_{t-1} = i) \end{bmatrix} = \begin{bmatrix} p_{i1} \\ \vdots \\ p_{iN} \end{bmatrix}$$

meaning

$$E(\boldsymbol{\xi}_t | \boldsymbol{\xi}_{t-1}) = \mathbf{P}\boldsymbol{\xi}_{t-1}$$

and

$$\boldsymbol{\xi}_t = \mathbf{P}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t$$

for $\mathbf{v}_t$ a discrete-valued martingale difference sequence whose elements always sum to zero. Thus the Markov chain admits a VAR(1) representation, with k-period-ahead regime probabilities conditional on the observed data $\boldsymbol{\Omega}_t$ given by

$$\begin{bmatrix} \text{Prob}(s_{t+k} = 1|\boldsymbol{\Omega}_t) \\ \vdots \\ \text{Prob}(s_{t+k} = N|\boldsymbol{\Omega}_t) \end{bmatrix} = \mathbf{P}^k \begin{bmatrix} \text{Prob}(s_t = 1|\boldsymbol{\Omega}_t) \\ \vdots \\ \text{Prob}(s_t = N|\boldsymbol{\Omega}_t) \end{bmatrix}. \tag{17}$$

Calculation of the moments and discussion of stationarity conditions for general processes subject to changes in regime can be found in Tjøstheim (1986), Yang (2000), Timmermann (2000), and Francq and Zakoïan (2001).

Although most applications assume a relatively small number of regimes, Sims and Zha (2006) used Bayesian prior information in a model with N as large as 10, while Calvet and Fisher (2004) estimated a model with thousands of regimes by imposing a functional restriction on the ways parameters vary across regimes.

2.3 Processes That Depend on Current and Past Regimes

In the original model proposed by Hamilton (1989) for describing economic recessions, the conditional density of GDP growth y_t was presumed to depend not just on the current regime but also on the r previous regimes:

$$y_t = m_{s_t} + \phi_1(y_{t-1} - m_{s_{t-1}}) + \phi_2(y_{t-2} - m_{s_{t-2}}) + \cdots + \phi_r(y_{t-r} - m_{s_{t-r}}) + \varepsilon_t. \tag{18}$$

While at first glance this might not appear to be a special case of the general formulation given in (13), this in fact is just a matter of representing (18) using the right notation. Taking $r = 1$ for illustration, define

$$s_t^* = \begin{cases} 1 & \text{when } s_t = 1 \text{ and } s_{t-1} = 1 \\ 2 & \text{when } s_t = 2 \text{ and } s_{t-1} = 1 \\ 3 & \text{when } s_t = 1 \text{ and } s_{t-1} = 2 \\ 4 & \text{when } s_t = 2 \text{ and } s_{t-1} = 2 \end{cases}.$$

Then s_t^* itself follows a four-state Markov chain with transition matrix

$$\mathbf{P}^* = \begin{bmatrix} p_{11} & 0 & p_{11} & 0 \\ p_{12} & 0 & p_{12} & 0 \\ 0 & p_{21} & 0 & p_{21} \\ 0 & p_{22} & 0 & p_{22} \end{bmatrix}$$

and the model (18) can indeed be viewed as a special case of (13), with for example

$$f(y_t|s_t^* = 2, \Omega_{t-1}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{[y_t - m_2 - \phi_1(y_{t-1} - m_1)]^2}{2\sigma^2}\right\}.$$

2.4 Inference About Regimes and Evaluating the Likelihood for the General Case

For any of the examples above we could collect the set of possible densities conditional on one of N different possible regimes in an $(N \times 1)$ vector $\boldsymbol{\eta}_t$ whose ith element is $f(\mathbf{y}_t|s_t = i, \Omega_{t-1}; \boldsymbol{\lambda})$ for $\Omega_{t-1} = \{\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_1\}$ and $\boldsymbol{\lambda}$ a vector consisting of all the unknown population parameters. For example, for the Markov-switching vector autoregression the ith element of $\boldsymbol{\eta}_t$ is given by (16) and $\boldsymbol{\lambda}$ is a vector collecting the unknown elements of $\{\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_N, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N, \mathbf{P}\}$ for $\mathbf{P}$ the $(N \times N)$ matrix whose row j column i element is $\text{Prob}(s_{t+1} = j|s_t = i)$ (so columns of $\mathbf{P}$ sum to unity). We likewise can define the $(N \times 1)$ vector $\hat{\boldsymbol{\xi}}_{t|t}$ whose ith element is the probability $\text{Prob}(s_t = i|\Omega_t; \boldsymbol{\lambda})$. One goal is to take the inference $\hat{\boldsymbol{\xi}}_{t-1|t-1}$ and update it to calculate $\hat{\boldsymbol{\xi}}_{t|t}$ using the observation on $\mathbf{y}_t$.Hamilton (1994, p. 692) showed that this can be accomplished by calculating

$$\hat{\boldsymbol{\xi}}_{t|t-1} = \mathbf{P}\hat{\boldsymbol{\xi}}_{t-1|t-1}$$

$$\hat{\boldsymbol{\xi}}_{t|t} = \frac{(\hat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t)}{\mathbf{1}'(\hat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t)} \tag{19}$$

where $\mathbf{1}$ denotes an $(N \times 1)$ vector of ones and $\odot$ denotes element-by-element vector multiplication.

If the Markov chain is known to be ergodic, we could begin the recursion for $t = 1$ by setting $\hat{\boldsymbol{\xi}}_{1|0}$ to the vector of unconditional probabilities, which as in Hamilton (1994, p. 684) can be found from the $(N + 1)$th column of the matrix $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ for

$$\underset{(N+1)\times N}{\mathbf{A}} = \begin{bmatrix} \mathbf{I}_N - \mathbf{P} \\ \mathbf{1}' \end{bmatrix}. \tag{20}$$

Alternative options are to treat the initial probabilities as separate parameters,

$$\begin{bmatrix} \text{Prob}(s_1 = 1|\Omega_0) \\ \vdots \\ \text{Prob}(s_1 = N|\Omega_0) \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \vdots \\ \rho_N \end{bmatrix}$$

where ρ_i could reflect prior beliefs (eg, $\rho_1 = 1$ if the analyst knows the sample begins in regime 1), complete ignorance ($\rho_i = 1/N$ for $i = 1, \ldots, N$), or $\boldsymbol{\rho}$ could be a separate vector of parameters also to be chosen by maximum likelihood. Any of the last three options is particularly attractive if the EM algorithm described in Section 2.5 or Gibbs sampler in

Section 2.8 are used, or if one wants to allow the possibility of a permanent regime shift which would mean that the Markov chain is not ergodic.

In a generalization of (10) and (11), the log-likelihood function for the observed data is naturally calculated as a by-product of the above recursion:

$$\mathcal{L}(\boldsymbol{\lambda}) = \sum_{t=1}^{T} \log f(\mathbf{y}_t | \boldsymbol{\Omega}_{t-1}; \boldsymbol{\lambda}) = \sum_{t=1}^{T} \log[\mathbf{1}'(\hat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t)]. \tag{21}$$

From (17) the k-period-ahead forecast of the regime, $\text{Prob}(s_{t+k} = j | \boldsymbol{\Omega}_t; \boldsymbol{\lambda})$ is found from the jth element of $\mathbf{P}^k \hat{\boldsymbol{\xi}}_{t|t}$.

It is also often of interest to calculate an inference about the regime at date t conditional on the full set of all observations through the end of the sample T, known as the "smoothed probability." The smoothed $\text{Prob}(s_t = i | \boldsymbol{\Omega}_T; \boldsymbol{\lambda})$ is obtained from the ith element of $\hat{\boldsymbol{\xi}}_{t|T}$ which can be calculated as in Hamilton (1994, p. 694) by iterating backward for $t = T-1, T-2, \ldots, 1$ on

$$\hat{\boldsymbol{\xi}}_{t|T} = \hat{\boldsymbol{\xi}}_{t|t} \odot \{\mathbf{P}'[\hat{\boldsymbol{\xi}}_{t+1|T} (\div) \hat{\boldsymbol{\xi}}_{t+1|t}]\} \tag{22}$$

where $(\div)$ denotes element-by-element division.

2.5 EM Algorithm

The unknown parameters $\boldsymbol{\lambda}$ could be estimated by maximizing the likelihood function (21) using numerical search methods. Alternatively, Hamilton (1990) noted that the EM algorithm is often a convenient method for finding the maximum of the likelihood function. This algorithm is simplest if we treat the initial probabilities $\hat{\boldsymbol{\xi}}_{1|0}$ as a vector of free parameters $\boldsymbol{\rho}$ rather than using ergodic probabilities from (20). This section describes how the EM algorithm would be implemented for the case of an unrestricted Markov-switching VAR (14), in which case $\boldsymbol{\lambda}$ includes $\boldsymbol{\rho}$ along with the elements of $\{\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_N, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N, \mathbf{P}\}$. The EM algorithm is an iterative procedure for generating a sequence of estimates $\{\hat{\boldsymbol{\lambda}}^{(\ell)}\}$ where the algorithm guarantees that the log likelihood (21) evaluated at $\hat{\boldsymbol{\lambda}}^{(\ell+1)}$ is greater than or equal to that at $\hat{\boldsymbol{\lambda}}^{(\ell)}$. Iterating until convergence leads to a local maximum of the likelihood function.

To calculate the value of $\hat{\boldsymbol{\lambda}}^{(\ell+1)}$ we first use $\hat{\boldsymbol{\lambda}}^{(\ell)}$ in Eq. (22) to evaluate the smoothed probabilities $\text{Prob}(s_t = i | \boldsymbol{\Omega}_T; \hat{\boldsymbol{\lambda}}^{(\ell)})$ and also smoothed joint probabilities Prob $(s_t = i, s_{t+1} = j | \boldsymbol{\Omega}_T; \hat{\boldsymbol{\lambda}}^{(\ell)})$. The latter are obtained from the row i column j element of the $(N \times N)$ matrix[b]

[b] The row i column j element of this matrix corresponds to

$$\text{Prob}(s_t = i | \boldsymbol{\Omega}_t) \frac{\text{Prob}(s_{t+1} = j | \boldsymbol{\Omega}_T)}{\text{Prob}(s_{t+1} = j | \boldsymbol{\Omega}_t)} p_{ij}$$

which from equation [22.A.21] in Hamilton (1994) equals $\text{Prob}(s_t = i, s_{t+1} = j | \boldsymbol{\Omega}_T)$.

$$\{\hat{\boldsymbol{\xi}}_{t|t}(\hat{\boldsymbol{\lambda}}^{(\ell)})[\hat{\boldsymbol{\xi}}_{t+1|T}(\hat{\boldsymbol{\lambda}}^{(\ell)})(\div)(\hat{\mathbf{P}}^{(\ell)}\hat{\boldsymbol{\xi}}_{t|t}(\hat{\boldsymbol{\lambda}}^{(\ell)}))]'\}\odot\hat{\mathbf{P}}^{(\ell)'}. \tag{23}$$

We then use these smoothed probabilities to generate a new estimate $\hat{\boldsymbol{\rho}}^{(\ell+1)}$, whose ith element is obtained from $\text{Prob}(s_1=i|\Omega_T;\hat{\boldsymbol{\lambda}}^{(\ell)})$, along with a new estimate $\hat{\mathbf{P}}^{(\ell+1)}$ whose row j column i element is given by

$$\hat{p}_{ij}^{(\ell+1)}=\frac{\sum_{t=1}^{T-1}\text{Prob}(s_t=i,s_{t+1}=j|\Omega_T;\hat{\boldsymbol{\lambda}}^{(\ell)})}{\sum_{t=1}^{T-1}\text{Prob}(s_t=i|\Omega_T;\hat{\boldsymbol{\lambda}}^{(\ell)})}.$$

Updated estimates of the VAR parameters for $i = 1,\ldots,N$ are given by

$$\hat{\boldsymbol{\Phi}}_i^{(\ell+1)}=\left(\sum_{t=1}^{T}\mathbf{y}_t\mathbf{x}_{t-1}'\text{Prob}(s_t=i|\Omega_T;\hat{\boldsymbol{\lambda}}^{(\ell)})\right)\left(\sum_{t=1}^{T}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'\text{Prob}(s_t=i|\Omega_T;\hat{\boldsymbol{\lambda}}^{(\ell)})\right)^{-1} \tag{24}$$

$$\hat{\boldsymbol{\Sigma}}_i^{(\ell+1)}=\frac{\sum_{t=1}^{T}(\mathbf{y}_t-\hat{\boldsymbol{\Phi}}_i^{(\ell+1)}\mathbf{x}_{t-1})(\mathbf{y}_t-\hat{\boldsymbol{\Phi}}_i^{(\ell+1)}\mathbf{x}_{t-1})'\text{Prob}(s_t=i|\Omega_T;\hat{\boldsymbol{\lambda}}^{(\ell)})}{\sum_{t=1}^{T}\text{Prob}(s_t=i|\Omega_T;\hat{\boldsymbol{\lambda}}^{(\ell)})}.$$

We thus simply iterate between calculating smoothed probabilities and OLS regressions of $\mathbf{y}_t$ on its lags weighted by those smoothed probabilities. The algorithm will converge to a point that is at least a local maximum of the log likelihood (21) with respect to $\boldsymbol{\lambda}$ subject to the constraints that $\boldsymbol{\rho}'\mathbf{1}=1, \mathbf{1}'\mathbf{P}=\mathbf{1}'$, all elements of $\boldsymbol{\rho}$ and $\mathbf{P}$ are nonnegative, and $\boldsymbol{\Sigma}_j$ is positive semidefinite for $j = 1,\ldots,N$.

2.6 EM Algorithm for Restricted Models

Often we might want to use a more parsimonious representation to which the EM algorithm is easily adapted. For example, suppose that we assume that there are no changes in regime for the equations describing the first n_1 variables in the system:

$$\mathbf{y}_{1t}=\mathbf{A}\mathbf{x}_{t-1}+\boldsymbol{\varepsilon}_{1t} \tag{25}$$

$$\mathbf{y}_{2t}=\mathbf{B}_{s_t}\mathbf{x}_{t-1}+\boldsymbol{\varepsilon}_{2t} \tag{26}$$

$$E\left\{\begin{bmatrix}\boldsymbol{\varepsilon}_{1t}\\ \boldsymbol{\varepsilon}_{2t}\end{bmatrix}\begin{bmatrix}\boldsymbol{\varepsilon}_{1t}' & \boldsymbol{\varepsilon}_{2t}'\end{bmatrix}\middle| s_t\right\}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12,s_t}\\ \boldsymbol{\Sigma}_{21,s_t} & \boldsymbol{\Sigma}_{22,s_t}\end{bmatrix}.$$

As in Hamilton (1994, p. 310) it is convenient to reparameterize the system by premultiplying (25) by $\boldsymbol{\Sigma}_{21,s_t}\boldsymbol{\Sigma}_{11}^{-1}$ and subtracting the result from (26) to obtain

$$\mathbf{y}_{2t}=\mathbf{C}_{s_t}\mathbf{y}_{1t}+\mathbf{D}_{s_t}\mathbf{x}_{t-1}+\mathbf{v}_{2t} \tag{27}$$

where $\mathbf{C}_{s_t}=\boldsymbol{\Sigma}_{21,s_t}\boldsymbol{\Sigma}_{11}^{-1}$, $\mathbf{D}_{st}=\mathbf{B}_{s_t}-\boldsymbol{\Sigma}_{21,s_t}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{A}$, $\mathbf{v}_{2t}=\boldsymbol{\varepsilon}_{2t}-\boldsymbol{\Sigma}_{21,s_t}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\varepsilon}_{1t}$, and $E(\mathbf{v}_{2t}\mathbf{v}_{2t}'|s_t)=\mathbf{H}_{s_t}=\boldsymbol{\Sigma}_{22,s_t}-\boldsymbol{\Sigma}_{21,s_t}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12,s_t}$. Then the likelihood associated with the system (25) and (27) factors into a regime-switching component and a regime-independent component parameterized by $\mathbf{A}, \boldsymbol{\Sigma}_{11}$. In the absence of restrictions on $\mathbf{B}_{s_t}$, $\boldsymbol{\Sigma}_{21,s_t}$, and $\boldsymbol{\Sigma}_{22,s_t}$, the values for $\mathbf{A}$ and $\boldsymbol{\Sigma}_{11}$ do not restrict the likelihood for the regime-switching block, meaning full-information maximum likelihood for the complete system can be implemented by maximizing the likelihood separately for the two blocks. For the regime-independent block, the MLE is obtained by simple OLS:

$$\hat{\mathbf{A}} = \left(\sum_{t=1}^{T} \mathbf{y}_{1t}\mathbf{x}'_{t-1}\right)\left(\sum_{t=1}^{T} \mathbf{x}_{t-1}\mathbf{x}'_{t-1}\right)^{-1}$$

$$\hat{\Sigma}_{11} = T^{-1}\sum_{t=1}^{T}(\mathbf{y}_{1t} - \hat{\mathbf{A}}\boldsymbol{x}_{t-1})(\mathbf{y}_{1t} - \hat{\mathbf{A}}\boldsymbol{x}_{t-1})'.$$

The MLE for the regime-switching block can be found using the EM algorithm,

$$\hat{\mathbf{G}}_i^{(\ell+1)} = \left(\sum_{t=1}^{T} \mathbf{y}_{2t}\mathbf{z}'_t p_{it}^{(\ell)}\right)\left(\sum_{t=1}^{T} \mathbf{z}_t\mathbf{z}'_t p_{it}^{(\ell)}\right)^{-1}$$

$$\hat{\mathbf{H}}_i^{(\ell+1)} = \frac{\left(\sum_{t=1}^{T}(\mathbf{y}_{2t} - \hat{\mathbf{G}}_i^{(\ell+1)}\mathbf{z}_t)(\mathbf{y}_{2t} - \hat{\mathbf{G}}_i^{(\ell+1)}\mathbf{z}_t)' p_{it}^{(\ell)}\right)}{\left(\sum_{t=1}^{T} p_{it}^{(\ell)}\right)}$$

$$p_{it}^{(\ell)} = \text{Prob}(s_t = i|\boldsymbol{\Omega}_T;\hat{\boldsymbol{\lambda}}^{(\ell)})$$

with $\mathbf{z}_t = (\mathbf{y}'_{1t}, \mathbf{x}'_{t-1})'$ and $\mathbf{G}_j = \left[\mathbf{C}_j \quad \mathbf{D}_j\right]$. The MLE for the original parameterization is then found simply by reversing the transformation that led to (27), for example, $\hat{\boldsymbol{\Sigma}}_{21,j} = \hat{\mathbf{C}}_j\hat{\boldsymbol{\Sigma}}_{11}$ and $\hat{\mathbf{B}}_j = \hat{\mathbf{D}}_j + \hat{\boldsymbol{\Sigma}}_{21,j}\hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\mathbf{A}}$.

Alternatively, suppose we want to restrict the switching coefficients to apply only to a subset $\mathbf{x}_{2,t-1}$ of the original regressors, as for example in a VAR in which only the intercept (the last element of $\mathbf{x}_{t-1}$) is changing with regime,

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_{1,t-1} + \mathbf{B}_{s_t}\mathbf{x}_{2,t-1} + \boldsymbol{\varepsilon}_t \tag{28}$$

with $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_t) = \boldsymbol{\Sigma}$. In this case the EM equations take the form[c]

$$\left[\hat{\mathbf{A}}^{(\ell+1)} \quad \hat{\mathbf{B}}_1^{(\ell+1)} \quad \hat{\mathbf{B}}_2^{(\ell+1)} \quad \cdots \quad \hat{\mathbf{B}}_N^{(\ell+1)}\right] = \mathbf{S}_{yx}(\hat{\boldsymbol{\lambda}}^{(\ell)})\mathbf{S}_{xx}^{-1}(\hat{\boldsymbol{\lambda}}^{(\ell)}) \tag{29}$$

$$\mathbf{S}_{yx}(\hat{\boldsymbol{\lambda}}^{(\ell)}) = \sum_{t=1}^{T}\mathbf{y}_t\left[\mathbf{x}'_{1,t-1} \quad \mathbf{x}'_{2,t-1}p_{1t}^{(\ell)} \quad \mathbf{x}'_{2,t-1}p_{2t}^{(\ell)} \quad \cdots \quad \mathbf{x}'_{2,t-1}p_{Nt}^{(\ell)}\right]$$

[c] See the Appendix for more details.

$$\mathbf{S}_{xx}(\hat{\boldsymbol{\lambda}}^{(\ell)}) = \sum_{t=1}^{T} \begin{bmatrix} \mathbf{x}_{1,t-1}\mathbf{x}'_{1,t-1} & \mathbf{x}_{1,t-1}\mathbf{x}'_{2,t-1}p_{1t}^{(\ell)} & \mathbf{x}_{1,t-1}\mathbf{x}'_{2,t-1}p_{2t}^{(\ell)} & \cdots & \mathbf{x}_{1,t-1}\mathbf{x}'_{2,t-1}p_{Nt}^{(\ell)} \\ \mathbf{x}_{2,t-1}\mathbf{x}'_{1,t-1}p_{1t}^{(\ell)} & \mathbf{x}_{2,t-1}\mathbf{x}'_{2,t-1}p_{1t}^{(\ell)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{x}_{2,t-1}\mathbf{x}'_{1,t-1}p_{2t}^{(\ell)} & \mathbf{0} & \mathbf{x}_{2,t-1}\mathbf{x}'_{2,t-1}p_{2t}^{(\ell)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}_{2,t-1}\mathbf{x}'_{1,t-1}p_{Nt}^{(\ell)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_{2,t-1}\mathbf{x}'_{2,t-1}p_{Nt}^{(\ell)} \end{bmatrix}$$

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}^{(\ell+1)} = {} & T^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\mathbf{y}_t - \hat{\mathbf{A}}^{(\ell+1)}\mathbf{x}_{1,t-1} - \hat{\mathbf{B}}_i^{(\ell+1)}\mathbf{x}_{2,t-1}\right] \\ & \times\left[\mathbf{y}_t - \hat{\mathbf{A}}^{(\ell+1)}\mathbf{x}_{1,t-1} - \hat{\mathbf{B}}_i^{(\ell+1)}\mathbf{x}_{2,t-1}\right]' p_{it}^{(\ell)}. \end{aligned} \tag{30}$$

2.7 Structural Vector Autoregressions and Impulse-Response Functions

A Gaussian structural vector autoregression takes the form

$$\mathbf{A}_{s_t}\mathbf{y}_t = \mathbf{B}_{s_t}\mathbf{x}_{t-1} + \mathbf{u}_t$$

where $\mathbf{x}_{t-1} = (\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \ldots, \mathbf{y}'_{t-r}, 1)'$ and $\mathbf{u}_t|s_t, \boldsymbol{\Omega}_{t-1} \sim N(\mathbf{0}, \mathbf{D}_{s_t})$. Here the elements of $\mathbf{u}_t$ are interpreted as different structural shocks which are identified by imposing certain restrictions on $\mathbf{A}_i$, $\mathbf{B}_i$, and $\mathbf{D}_i$. For example, the common Cholesky identification assumes that the structural equations are recursive, with $\mathbf{D}_i$ diagonal and $\mathbf{A}_i$ lower triangular with ones along the diagonal. For an identified structure we could estimate parameters by setting the ith element of $\boldsymbol{\eta}_t$ in (19) to

$$\eta_{it} = \frac{1}{(2\pi)^{n/2}}\frac{\sqrt{|\mathbf{A}_i|^2}}{\sqrt{|\mathbf{D}_i|}}\exp\left[-(1/2)(\mathbf{A}_i\mathbf{y}_t - \mathbf{B}_i\mathbf{x}_{t-1})'\mathbf{D}_i^{-1}(\mathbf{A}_i\mathbf{y}_t - \mathbf{B}_i\mathbf{x}_{t-1})\right]$$

and then choosing $\{\mathbf{A}_1,\ldots,\mathbf{A}_N,\mathbf{B}_1,\ldots,\mathbf{B}_N,\mathbf{D}_1,\ldots,\mathbf{D}_N,\mathbf{P}\}$ to maximize the likelihood (21).

A faster algorithm is likely to be obtained by first finding the MLEs $\{\hat{\boldsymbol{\Phi}}_1,\ldots,\hat{\boldsymbol{\Phi}}_N,\hat{\boldsymbol{\Sigma}}_1,\ldots,\hat{\boldsymbol{\Sigma}}_N,\hat{\mathbf{P}},\hat{\boldsymbol{\rho}}\}$ for the reduced form (14)–(15) using the EM algorithm in Section 2.5. If the model is just identified, we can just translate these into the implied structural parameters $\{\hat{\mathbf{A}}_1,\ldots,\hat{\mathbf{A}}_N,\hat{\mathbf{B}}_1,\ldots,\hat{\mathbf{B}}_N,\hat{\mathbf{D}}_1,\ldots,\hat{\mathbf{D}}_N,\mathbf{P},\hat{\rho}\}$, while for an overidentified model we could find the values for the structural parameters that are closest to the reduced form using minimum chi-square estimation (eg, Hamilton and Wu, 2012). For example, for the Cholesky formulation we would just find the Cholesky factorization $\hat{\mathbf{P}}_i\hat{\mathbf{P}}_i' = \hat{\Sigma}_i$ for each i. The row j column j element of $\hat{\mathbf{D}}_i$ is then the square of the row j column j element of $\hat{\mathbf{P}}_i$. Then $\hat{\mathbf{A}}_i = \hat{\mathbf{D}}_i^{1/2}\hat{\mathbf{P}}_i^{-1}$ and $\hat{\mathbf{B}}_i = \hat{\mathbf{A}}_i\hat{\boldsymbol{\Phi}}_i$.

Users of structural vector autoregressions are often interested in structural impulse-response functions, which in this case are functions of the regime at date t:

$$\mathbf{H}_{mj}=\frac{\partial E(\mathbf{y}_{t+m}|s_t=j,\boldsymbol{\Omega}_t)}{\partial \mathbf{u}_t'}=\frac{\partial E(\mathbf{y}_{t+m}|s_t=j,\boldsymbol{\Omega}_t)}{\partial \boldsymbol{\varepsilon}_t'}\frac{\partial \boldsymbol{\varepsilon}_t}{\partial \mathbf{u}_t'}=\boldsymbol{\Psi}_{mj}\mathbf{A}_j^{-1}.$$

The nonorthogonalized or reduced-form IRF, $\boldsymbol{\Psi}_{mj}$, can be found as follows. Suppose we first condition not just on the regime j at date t but also on a particular regime j_1 for date $t+1, j_2$ for $t+2$, and j_m for $t+m$, and consider the value of

$$\tilde{\boldsymbol{\Psi}}_{m,j,j_1,\ldots,j_m}=\frac{\partial E(\mathbf{y}_{t+m}|s_t=j,s_{t+1}=j_1,\ldots,s_{t+m}=j_m,\boldsymbol{\Omega}_t)}{\partial \boldsymbol{\varepsilon}_t'}.$$

Karamé (2010) noted that this $(n \times n)$ matrix can be calculated from the recursion

$$\tilde{\boldsymbol{\Psi}}_{m,j,j_1,\ldots,j_m}=\boldsymbol{\Phi}_{1,j_m}\tilde{\boldsymbol{\Psi}}_{m-1,j,j_1,\ldots,j_{m-1}}+\boldsymbol{\Phi}_{2,j_m}\tilde{\boldsymbol{\Psi}}_{m-2,j,j_1,\ldots,j_{m-2}}+\cdots+\boldsymbol{\Phi}_{r,j_m}\tilde{\boldsymbol{\Psi}}_{m-r,j,j_1,\ldots,j_{m-r}}$$

for $m=1,2,\ldots$ where $\tilde{\boldsymbol{\Psi}}_{0j}=\mathbf{I}_n$ and $\mathbf{0}=\tilde{\boldsymbol{\Psi}}_{-1,.}=\tilde{\boldsymbol{\Psi}}_{-2,.}=\cdots$. The object of interest is found by integrating out the conditioning variables,

$$\begin{aligned}\boldsymbol{\Psi}_{mj}&=\sum_{j_1=1}^{N}\cdots\sum_{j_m=1}^{N}\tilde{\boldsymbol{\Psi}}_{m,j,j_1,\ldots,j_m}\text{Prob}(s_{t+1}=j_1,\ldots,s_{t+m}=j_m|s_t=j)\\&=\sum_{j_1=1}^{N}\cdots\sum_{j_m=1}^{N}\tilde{\boldsymbol{\Psi}}_{m,j,j_1,\ldots,j_m}p_{j,j_1}p_{j_1,j_2}\cdots p_{j_{m-1},j_m}.\end{aligned}$$

These magnitudes can either be calculated analytically for modest m and N or by simulation.

Such regime-specific impulse-response functions are of interest for questions such as whether monetary policy (Lo and Piger, 2005) or fiscal policy (Auerbach and Gorodnichenko, 2012) has different effects on the economy during an expansion or recession.

2.8 Bayesian Inference and the Gibbs Sampler

Bayesian methods offer another popular approach for econometric inference. The Bayesian begins with prior beliefs about the unknown parameters $\boldsymbol{\lambda}$ which are represented using a probability density $f(\boldsymbol{\lambda})$ that associates a higher probability with values of $\boldsymbol{\lambda}$ that are judged to be more plausible. The goal of inference is to revise these beliefs in the form of a posterior density $f(\boldsymbol{\lambda}|\boldsymbol{\Omega}_T)$ based on the observed data $\boldsymbol{\Omega}_T=\{\mathbf{y}_1,\ldots,\mathbf{y}_T\}$. Often the prior distribution $f(\boldsymbol{\lambda})$ is assumed to be taken from a particular parametric family known as a natural conjugate distribution. These have the property that the prior and posterior are from the same family, as would be the case for example if the prior beliefs were based on an earlier sample of data. Natural conjugates are helpful because they allow many of the results to be obtained using known analytic solutions.

Again we will illustrate some of the main ideas using a Markov-switching vector autoregression:

$$\mathbf{y}_t=\boldsymbol{\Phi}_{s_t}\mathbf{x}_{t-1}+\boldsymbol{\varepsilon}_t$$

$$\boldsymbol{\varepsilon}_t | s_t = i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$$
$$\text{Prob}(s_1 = i) = \rho_i$$
$$\text{Prob}(s_t = j | s_{t-1} = i) = p_{ij}.$$

2.8.1 Prior Distributions

The Dirichlet distribution is the natural conjugate for the parameters that determine the Markov transition probabilities. Suppose $\mathbf{z} = (z_1, \ldots, z_N)'$ is an $(N \times 1)$ vector of nonnegative random variables that sum to unity. The Dirichlet density with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$ is given by

$$f(\mathbf{z}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_N)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_N)} z_1^{\alpha_1 - 1} \cdots z_N^{\alpha_N - 1}$$

for $\Gamma(.)$ the gamma function, with the constant ensuring that the density integrates to unity over the set of vectors $\mathbf{z}$ satisfying the specified conditions. The beta distribution is a special case when $N = 2$, usually expressed as a function of the scalar $z_1 \in (0,1)$,

$$f(z_1) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z_1^{\alpha_1 - 1}(1 - z_1)^{\alpha_2 - 1}.$$

We then represent prior beliefs over the $(N \times 1)$ vector of initial probabilities as $(\rho_1, \ldots, \rho_N) \sim D(\alpha_1, \ldots, \alpha_N)$ and those for transition probabilities as $(p_{i1}, \ldots, p_{iN}) \sim D(\alpha_{i1}, \ldots, \alpha_{iN})$ for $i = 1, \ldots, N$. The natural conjugate for $\boldsymbol{\Sigma}_j$, the innovation variance matrix for regime j, is provided by the Wishart distribution. Let $\mathbf{z}_i$ be independent $(n \times 1)$ $N(\mathbf{0}, \boldsymbol{\Lambda}^{-1})$ vectors and consider the matrix $\mathbf{W} = \mathbf{z}_1\mathbf{z}_1' + \cdots + \mathbf{z}_\eta \mathbf{z}_\eta'$ for $\eta > n - 1$. This matrix is said to have an n-dimensional Wishart distribution with η degrees of freedom and scale matrix $\boldsymbol{\Lambda}^{-1}$, whose density is

$$f(\mathbf{W}) = c|\boldsymbol{\Lambda}|^{\eta/2}|\mathbf{W}|^{(\eta - n - 1)/2} \exp[-(1/2)\text{tr}(\mathbf{W}\boldsymbol{\Lambda})]$$

where tr(.) denotes the trace (sum of diagonal elements). For a univariate regression ($n = 1$) this becomes Λ^{-1} times a $\chi^2(\eta)$ variable, or equivalently a gamma distribution with mean η/Λ and variance $2\eta/\Lambda^2$. The constant c is chosen so that the density integrates to unity over the set of all positive definite symmetric matrices $\mathbf{W}$ (eg, DeGroot, 1970, p. 57):

$$c = \left[2^{\eta n/2} \pi^{n(n-1)/4} \prod_{j=1}^{n} \Gamma\left(\frac{\eta + 1 - j}{2}\right) \right]^{-1}.$$

The natural conjugate prior for $\boldsymbol{\Sigma}_j^{-1}$, the inverse of the innovation variance matrix in regime j, takes the form of a Wishart distribution with η_j degrees of freedom and scale $\boldsymbol{\Lambda}_j^{-1}$:

$$f(\mathbf{\Sigma}_j^{-1}) = c|\mathbf{\Lambda}_j|^{\eta_j/2}|\mathbf{\Sigma}_j|^{-(\eta_j-n-1)/2}\exp\left[-(1/2)\mathrm{tr}\left(\mathbf{\Sigma}_j^{-1}\mathbf{\Lambda}_j\right)\right].$$

Prior information about the regression coefficients $\boldsymbol{\varphi}_j = \mathrm{vec}(\mathbf{\Phi}_j')$ for regime j can be represented with an $N(\mathbf{m}_j,\mathbf{M}_j)$ distribution. The formulas are much simpler in the case of no useful prior information about these coefficients (which can be viewed as the limit of the inference as $\mathbf{M}_j^{-1} \to \mathbf{0}$), and this limiting case will be used for the results presented here.

2.8.2 Likelihood Function and Conditional Posterior Distributions

Collect the parameters that characterize Markov probabilities in a set $p = \{\rho_j, p_{1j}, \ldots, p_{Nj}\}_{j=1}^{N}$, those for variances in a set $\sigma = \{\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_N\}$, and VAR coefficients $\varphi = \{\mathbf{\Phi}_1,\ldots,\mathbf{\Phi}_N\}$. If we were to condition on all of these parameters along with a particular numerical value for the realization of the regime for every date $\mathcal{S} = \{s_1, \ldots, s_T\}$ the likelihood function of the observed data $\mathbf{\Omega}_T = \{\mathbf{y}_1,\ldots,\mathbf{y}_T\}$ would be

$$f(\Omega_T|p,\sigma,\varphi,\mathcal{S}) = \prod_{t=1}^{T}\frac{1}{(2\pi)^{n/2}}|\mathbf{\Sigma}_{s_t}|^{-1/2}\exp\left[-(1/2)(\mathbf{y}_t - \mathbf{\Phi}_{s_t}\mathbf{x}_{t-1})'\mathbf{\Sigma}_{s_t}^{-1}(\mathbf{y}_t - \mathbf{\Phi}_{s_t}\mathbf{x}_{t-1})\right]$$

$$= \prod_{t=1}^{T}\frac{1}{(2\pi)^{n/2}}\sum_{j=1}^{N}\delta_{jt}|\mathbf{\Sigma}_j|^{-1/2}\exp\left[-(1/2)(\mathbf{y}_t - \mathbf{\Phi}_j\mathbf{x}_{t-1})'\mathbf{\Sigma}_j^{-1}(\mathbf{y}_t - \mathbf{\Phi}_j\mathbf{x}_{t-1})\right]$$

where $\delta_{jt} = 1$ if $s_t = j$ and is zero otherwise. With independent priors the joint density of the data, parameters, and regimes is then

$$f(\Omega_T, p, \sigma, \varphi, \mathcal{S}) = f(\Omega_T|p,\sigma,\varphi,\mathcal{S})f(p)f(\sigma)f(\varphi)f(\mathcal{S}|p) \tag{31}$$

$$f(p) \propto \prod_{j=1}^{N}\rho_j^{\alpha_j-1}p_{1j}^{\alpha_{1j}-1}\cdots p_{Nj}^{\alpha_{Nj}-1}$$

$$f(\sigma) \propto \prod_{j=1}^{N}|\mathbf{\Sigma}_j|^{-(\eta_j-n-1)/2}\exp\left[-(1/2)\mathrm{tr}\left(\mathbf{\Sigma}_j^{-1}\mathbf{\Lambda}_j\right)\right]$$

$$f(\varphi) \propto 1$$

$$f(\mathcal{S}|p) = \rho_{s_1}p_{s_1,s_2}p_{s_2,s_3}\cdots p_{s_{T-1},s_T}$$

where p_{s_{t-1},s_t} denotes the parameter p_{ij} when $s_{t-1} = i$ and $s_t = j$.

Let $\Delta(j) = \{t \in 1,\ldots,T : \delta_{jt} = 1\}$ denote the set of dates for which the regime is j. From (31) the posterior distribution of $\mathbf{\Sigma}_j$ conditional on $\mathbf{\Omega}_T, p, \varphi, \mathcal{S}$ is given by

$$
\begin{aligned}
f(\Sigma_j^{-1}|\Omega_T,p,\varphi,\mathcal{S}) \propto{}& |\boldsymbol{\Sigma}_j|^{-(\eta_j-n-1)/2}\exp\left[-(1/2)\mathrm{tr}\left(\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Lambda}_j\right)\right]\times \\
&\prod_{t\in\Delta(j)}|\boldsymbol{\Sigma}_j|^{-1/2}\exp[-(1/2)(\mathbf{y}_t-\boldsymbol{\Phi}_j\mathbf{x}_{t-1})'\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_t-\boldsymbol{\Phi}_j\mathbf{x}_{t-1})] \\
={}& |\boldsymbol{\Sigma}_j|^{-(T_j+\eta_j-n-1)/2}\exp[-(1/2)\mathrm{tr}[\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\Lambda}_j+\mathbf{H}_j)] \qquad (32)
\end{aligned}
$$

for $T_j=\sum_{t=1}^{T}\delta_{jt}$ the number of dates characterized by regime j and $\mathbf{H}_j=\sum_{t=1}^{T}\delta_{jt}(\mathbf{y}_t-\boldsymbol{\Phi}_j\mathbf{x}_{t-1})(\mathbf{y}_t-\boldsymbol{\Phi}_j\mathbf{x}_{t-1})'$ the sum of outer products of the residual vectors for those observations. In other words, $\boldsymbol{\Sigma}_j^{-1}|\Omega_T,p,\varphi,\mathcal{S}$ has a Wishart distribution with $T_j+\eta_j$ degrees of freedom and scale matrix $(\boldsymbol{\Lambda}_j+\mathbf{H}_j)^{-1}$.

Likewise for $\hat{\boldsymbol{\Phi}}_j=\left(\sum_{t=1}^{T}\delta_{jt}\mathbf{y}_t\mathbf{x}_{t-1}'\right)\left(\sum_{t=1}^{T}\delta_{jt}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)^{-1}$, that is for $\hat{\boldsymbol{\Phi}}_j$ the OLS regression coefficients using only observations for regime j, the posterior distribution of $\boldsymbol{\varphi}_j=\mathrm{vec}(\boldsymbol{\Phi}_j')$ conditional on $\Omega_T,p,\sigma,\mathcal{S}$ is

$$
\begin{aligned}
f(\boldsymbol{\varphi}_j|\Omega_T,p,\sigma,\mathcal{S}) &\propto \prod_{t\in\Delta(j)}\exp[-(1/2)(\mathbf{y}_t-\boldsymbol{\Phi}_j\mathbf{x}_{t-1})'\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_t-\boldsymbol{\Phi}_j\mathbf{x}_{t-1})] \\
&= \prod_{t\in\Delta(j)}\exp[-(1/2)(\mathbf{y}_t-\hat{\boldsymbol{\Phi}}_j\mathbf{x}_{t-1}+\hat{\boldsymbol{\Phi}}_j\mathbf{x}_{t-1}-\boldsymbol{\Phi}_j\mathbf{x}_{t-1})'\boldsymbol{\Sigma}_j^{-1}\times \\
&\qquad (\mathbf{y}_t-\hat{\boldsymbol{\Phi}}_j\mathbf{x}_{t-1}+\hat{\boldsymbol{\Phi}}_j\mathbf{x}_{t-1}-\boldsymbol{\Phi}_j\mathbf{x}_{t-1})] \\
&\propto \prod_{t\in\Delta(j)}\exp[-(1/2)\mathbf{x}_{t-1}'(\hat{\boldsymbol{\Phi}}_j-\boldsymbol{\Phi}_j)'\boldsymbol{\Sigma}_j^{-1}(\hat{\boldsymbol{\Phi}}_j-\boldsymbol{\Phi}_j)\mathbf{x}_{t-1}] \\
&= \prod_{t\in\Delta(j)}\exp\{-(1/2)(\hat{\boldsymbol{\phi}}_j-\boldsymbol{\varphi}_j)'[\boldsymbol{\Sigma}_j^{-1}\otimes\mathbf{x}_{t-1}\mathbf{x}_{t-1}'](\hat{\boldsymbol{\phi}}_j-\boldsymbol{\varphi}_j)\} \\
&= \exp\left\{-(1/2)(\boldsymbol{\varphi}_j-\hat{\boldsymbol{\phi}}_j)'\left[\boldsymbol{\Sigma}_j\otimes\left(\sum_{t=1}^{T}\delta_{jt}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)^{-1}\right]^{-1}(\boldsymbol{\varphi}_j-\hat{\boldsymbol{\phi}}_j)\right\}
\end{aligned}
$$

establishing that $\boldsymbol{\varphi}_j|\Omega_T,p,\sigma,\mathcal{S}\sim N\left(\hat{\boldsymbol{\phi}}_j,\boldsymbol{\Sigma}_j\otimes\left(\sum_{t=1}^{T}\delta_{jt}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)^{-1}\right)$ for $\hat{\boldsymbol{\phi}}_j=\mathrm{vec}(\hat{\boldsymbol{\Phi}}_j')$.

The conditional posterior distribution of $\boldsymbol{\rho}$ is found from

$$
f(\boldsymbol{\rho}|\Omega_T,\varphi,\sigma,\mathcal{S})\propto\rho_1^{a_1-1}\cdots\rho_N^{\alpha_N-1}\rho_{s_t}
$$

which will be recognized as $D(\alpha_1+\delta_{11},\ldots,\alpha_N+\delta_{N1})$, in other words, a Dirichlet distribution in which we have increased the parameter associated with the realized regime for observation 1 by unity and kept all other parameters the same. We similarly have $(p_{i1},\ldots,p_{iN})|\Omega_T,\varphi,\sigma,\mathcal{S}\sim D(\alpha_{i1}+T_{i1},\ldots,\alpha_{iN}+T_{iN})$ for $T_{ij}=\sum_{t=2}^{T}\delta_{i,t-1}\delta_{jt}$ the number of times that regime i is followed by j in the given sequence $\mathcal{S}$.

2.8.3 Gibbs Sampler

The idea behind the Gibbs sampler is to take advantage of the above known conditional distributions to generate a sequence of random variables whose unconditional distribution will turn out to be the object we're interested in. Suppose that as the result of a previous iteration ℓ we had generated particular numerical values for $\varphi,\sigma,p,\mathcal{S}$. We could for example begin iteration $\ell=1$ with arbitrary initial guesses for the parameters along with a possible realization of the regime for each date. Given the numbers from iteration ℓ, we could generate $\mathbf{\Sigma}_j^{(\ell+1)}$ from expression (32), namely, $\mathbf{\Sigma}_j^{(\ell+1)}$ is the inverse of a draw from a Wishart distribution with $T_j^{(\ell)}+\eta_j$ degrees of freedom and scale matrix $\left(\mathbf{\Lambda}_j+\mathbf{H}_j^{(\ell)}\right)^{-1}$, where $T_j^{(\ell)}=\sum_{t=1}^{T}\delta_{jt}^{(\ell)}$ is a simple count of the number of elements in $\{s_1^{(\ell)},\ldots,s_T^{(\ell)}\}$ that take the value j and $\mathbf{H}_j^{(\ell)}$ is the sum of the residual outer products $\sum_{t=1}^{T}\delta_{jt}^{(\ell)}(\mathbf{y}_t-\mathbf{\Phi}_j^{(\ell)}\mathbf{x}_{t-1})$ $(\mathbf{y}_t-\mathbf{\Phi}_j^{(\ell)}\mathbf{x}_{t-1})'$ for those $T_j^{(\ell)}$ observations. Doing so for each $j=1,\ldots,N$ gives us the new $\sigma^{(\ell+1)}$. We get a new value for the VAR coefficients by generating $\boldsymbol{\varphi}_j^{(\ell+1)}\sim N\left(\hat{\boldsymbol{\varphi}}_j^{(\ell+1)},\mathbf{\Sigma}_j^{(\ell+1)}\otimes\left(\sum_{t=1}^{T}\delta_{jt}^{(\ell)}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)^{-1}\right)$ where $\hat{\boldsymbol{\varphi}}_j^{(\ell+1)}=\text{vec}\left[\hat{\mathbf{\Phi}}_j^{(\ell+1)'}\right]$ is obtained from OLS regression on these $T_j^{(\ell)}$ observations: $\hat{\mathbf{\Phi}}_j^{(\ell+1)}=\left(\sum_{t=1}^{T}\delta_{jt}^{(\ell)}\mathbf{y}_t\mathbf{x}_{t-1}'\right)$ $\left(\sum_{t=1}^{T}\delta_{jt}^{(\ell)}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'\right)^{-1}$. New initial probabilities $(\rho_1^{(\ell+1)},\ldots,\rho_N^{(\ell+1)})$ are generated from $D(\alpha_1+\delta_{11}^{(\ell)},\ldots,\alpha_N+\delta_{N1}^{(\ell)})$ and new Markov probabilities $(p_{i1}^{(\ell+1)},\ldots,p_{iN}^{(\ell+1)})$ from $D(\alpha_{i1}+T_{i1}^{(\ell)},\ldots.,\alpha_{iN}+T_{iN}^{(\ell)})$ for $T_{ij}^{(\ell)}$ the number of times $s_t^{(\ell)}=i$ is followed by $s_{t+1}^{(\ell)}=j$ within the particular realization $(s_1^{(\ell)},\ldots,s_T^{(\ell)})$.

Finally, we can get a new realization $(s_1^{(\ell+1)},\ldots,s_T^{(\ell+1)})$ as a draw from the conditional posterior $f(\mathcal{S}|\Omega_T,p^{(\ell+1)},\sigma^{(\ell+1)},\varphi^{(\ell+1)})$ by iterating backward on a variant of the smoothing algorithm in Section 2.4. Specifically, given the values $(p^{(\ell+1)},\sigma^{(\ell+1)},\varphi^{(\ell+1)})$ we can iterate on (19) for $t=1,\ldots,T$ to calculate the $(N\times 1)$ vector $\{\hat{\boldsymbol{\xi}}_{t|t}^{(\ell+1)}\}_{t=1}^{T}$ whose jth element is $\text{Prob}(s_t=j|\Omega_t,p^{(\ell+1)},\sigma^{(\ell+1)},\varphi^{(\ell+1)})$. To generate $s_T^{(\ell+1)}$, we first generate a $U(0,1)$ variate. If this is smaller than the calculated $\text{Prob}(s_T=1|\Omega_T,p^{(\ell+1)},\sigma^{(\ell+1)},\varphi^{(\ell+1)})$, we set $s_T^{(\ell+1)}=1$. If the uniform variable turns out to be between $\text{Prob}(s_T=1|\Omega_T,p^{(\ell+1)},\sigma^{(\ell+1)},\varphi^{(\ell+1)})$ and the sum $\text{Prob}(s_T=1|\Omega_T,p^{(\ell+1)},\sigma^{(\ell+1)},\varphi^{(\ell+1)})+\text{Prob}(s_T=2|\Omega_T,p^{(\ell+1)},\sigma^{(\ell+1)},\varphi^{(\ell+1)})$, we set $s_T^{(\ell+1)}=2$, and so on. After we have generated a particular value for $s_T^{(\ell+1)}$ we can use (23) to calculate the probability

$$\begin{aligned}\text{Prob}(s_{T-1}&=i|s_T=s_T^{(\ell+1)},\Omega_T,p^{(\ell+1)},\sigma^{(\ell+1)},\varphi^{(\ell+1)})=\\&\frac{\text{Prob}(s_{T-1}=i,s_T=s_T^{(\ell+1)}|\Omega_T,p^{(\ell+1)},\sigma^{(\ell+1)},\varphi^{(\ell+1)})}{\text{Prob}(s_T=s_T^{(\ell+1)}|\Omega_T,p^{(\ell+1)},\sigma^{(\ell+1)},\varphi^{(\ell+1)})}\end{aligned}$$

with which we generate a draw $s_{T-1}^{(\ell+1)}$. Iterating backward in this manner gives us the full sequence $\mathcal{S}^{(\ell+1)}$.

We now have a complete new set $p^{(\ell+1)}, \boldsymbol{\sigma}^{(\ell+1)}, \boldsymbol{\varphi}^{(\ell+1)}, \mathcal{S}^{(\ell+1)}$, from which we can then generate values for $\ell + 2, \ell + 3$, and so on. The idea behind the Gibbs sampler (eg, Albert and Chib, 1993) is that the sequence corresponds to a Markov chain whose ergodic distribution under general conditions is the true posterior distribution $f(p, \sigma, \varphi, \mathcal{S}|\Omega_T)$. The proposal is then to discard the first say 10^6 draws and retain the next 10^6 draws as a sample from the posterior distribution.

One can also adapt approaches like those in Section 2.6 to apply the Gibbs sampler to restricted models. For example, if regime switching is confined to a subset of the equations, we can use the parameterization (27) and perform inference on the regime-switching subset independently from the rest of the system.

Although very convenient for many applications, one caution to be aware of in applying the Gibbs sampler is the role of label switching. Strategies for dealing with this are discussed by Celeux et al. (2000), Frühwirth-Schnatter (2001), and Geweke (2007).

2.9 Time-Varying Transition Probabilities

While the calculations above assumed that regimes are characterized by an exogenous Markov chain, this is easily generalized. We could replace (2) with

$$\operatorname{Prob}(s_t = j | s_{t-1} = i, s_{t-2} = k, \ldots, \boldsymbol{\Omega}_{t-1}) = p_{ij}(\mathbf{x}_{t-1}; \boldsymbol{\lambda}) \quad i, j = 1, \ldots, N \tag{33}$$

where $\mathbf{x}_{t-1}$ is a subset of $\boldsymbol{\Omega}_{t-1}$ or other observed variables on which one is willing to condition and $p_{ij}(\mathbf{x}_{t-1}; \boldsymbol{\lambda})$ is a specified parametric function. The generalization of (9) then becomes

$$\operatorname{Prob}(s_t = j | \boldsymbol{\Omega}_{t-1}) = \sum_{i=1}^{N} p_{ij}(\mathbf{x}_{t-1}; \boldsymbol{\lambda}) \operatorname{Prob}(s_{t-1} = i | \boldsymbol{\Omega}_{t-1}),$$

where the sequence $\operatorname{Prob}(s_t = i | \boldsymbol{\Omega}_t)$ can still be calculated iteratively as in (7),

$$\operatorname{Prob}(s_t = j | \boldsymbol{\Omega}_t) = \frac{\operatorname{Prob}(s_t = j | \boldsymbol{\Omega}_{t-1}) f(\mathbf{y}_t | \boldsymbol{\Omega}_{t-1}; \boldsymbol{\theta}_j)}{f(\mathbf{y}_t | \boldsymbol{\Omega}_{t-1})} \tag{34}$$

with the predictive density in the denominator now

$$f(\mathbf{y}_t | \boldsymbol{\Omega}_{t-1}) = \sum_{i=1}^{N} \operatorname{Prob}(s_t = i | \boldsymbol{\Omega}_{t-1}) f(\mathbf{y}_t | \boldsymbol{\Omega}_{t-1}; \boldsymbol{\theta}_i). \tag{35}$$

Diebold et al. (1994) showed how the EM algorithm works in such a setting, while Filardo and Gordon (1998) developed a Gibbs sampler. Other interesting applications with time-varying transition probabilities include Filardo (1994) and Peria (2002).

2.10 Latent-Variable Models with Changes in Regime

A more involved case that cannot be handled using the above devices is when the conditional density of $\mathbf{y}_t$ depends on the full history of regimes $(s_t,s_{t-1},\ldots,s_1)$ through date t. One important case in which this arises is when a process moving in and out of recession phase is proposed as an unobserved latent variable influencing an $(n \times 1)$ vector of observed variables $\mathbf{y}_t$. For example, Chauvet (1998) specified a process for an unobserved scalar business-cycle factor F_t characterized by

$$F_t = \alpha_{s_t} + \phi F_{t-1} + \eta_t$$

which influences the observed $\mathbf{y}_t$ according to

$$\mathbf{y}_t = \boldsymbol{\psi} F_t + \mathbf{q}_t$$

for $\boldsymbol{\psi}$ an $(n \times 1)$ vector of factor loadings and elements of $\mathbf{q}_t$ presumed to follow separate autoregressions. This can be viewed as a state-space model with regime-dependent parameters in which the conditional density (13) turns out to depend on the complete history $(s_t,s_{t-1},\ldots,s_1)$.

One approach for handling such models is an approximation to the log likelihood and optimal inference developed by Kim (1994). Chauvet and Hamilton (2006) and Chauvet and Piger (2008) demonstrated the real-time usefulness of this approach for recognizing US recessions with $\mathbf{y}_t$ a (4×1) vector of monthly indicators of sales, income, employment, and industrial production, while Camacho et al. (2014) have had success using a more detailed model for the Euro area.

The Gibbs sampler offers a particularly convenient approach for this class of models. We simply add the unobserved sequence of factors $\{F_1,\ldots,F_T\}$ as another random block to be sampled from along with p,σ,φ, and $\mathcal{S}$. Conditional on $\{F_1,\ldots,F_T\}$, draws for those other blocks can be performed exactly as in Section 2.8, while draws for $\{F_1,\ldots,F_T\}$ conditional on the regimes and other parameters can be calculated using well-known algorithms associated with the Kalman filter; see Kim and Nelson (1999a) for details.

2.11 Selecting the Number of Regimes

Often one would want to test the null hypothesis that there are N regimes against the alternative of $N+1$, and in particular to test the null hypothesis that there are no changes in regime at all $(H_0 : N = 1)$. A natural idea would be to compare the values achieved for the log likelihood (21) for N and $N+1$. Unfortunately, the likelihood ratio does not have the usual asymptotic χ^2 distribution because under the null hypothesis, some of the parameters of the model become unidentified. For example, if one thought of the null hypothesis in (1) as $m_1 = m_2$, when the null is true the maximum likelihood estimates $\hat{p}_{11}$ and $\hat{p}_{22}$ do not converge to any population values. Hansen (1992) and Garcia (1998) examined the distribution of the likelihood ratio statistic in this setting, though implementing their procedures can be quite involved if the model is at all complicated. Cho and White (2007) and Carter and Steigerwald (2012, 2013) suggested quasi-likelihood ratio tests that ignore the Markov property of s_t. For discussion of some of the subtleties and possible solutions for the case of i.i.d. regime changes, see Hall and Stewart (2005) and Chen and Li (2009).

An alternative is to calculate instead general measures that trade off the fit of the likelihood against the number of parameters estimated. Popular methods such as Schwarz's (1978) Bayesian criterion rely for their asymptotic justification on the same regularity conditions whose failure causes the likelihood ratio statistic to have a nonstandard distribution. But Smith et al. (2006) developed a simple test that can be used to select the number of regimes for a Markov-switching regression,

$$y_t = \mathbf{x}_t' \boldsymbol{\beta}_{s_t} + \sigma_{s_t} \varepsilon_t \tag{36}$$

where $\varepsilon_t \sim N(0,1)$ and s_t follows an N-state Markov chain. The authors proposed to estimate the parameter vector $\boldsymbol{\lambda} = \left(\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_N', \sigma_1, \ldots, \sigma_N, p_{ij, i=1,\ldots,N; j=1,\ldots,N-1}\right)^{-1}$ by maximum likelihood for each possible choice of N and calculate

$$\hat{T}_i = \sum_{t=1}^{T} \text{Prob}(s_t = i | \boldsymbol{\Omega}_T; \hat{\boldsymbol{\lambda}}_{\text{MLE}}) \quad \text{for } i = 1, \ldots, N$$

using the full-sample smoothed probabilities. They suggested choosing the value of N for which

$$\text{MSC} = -2\mathcal{L}(\hat{\boldsymbol{\lambda}}_{\text{MLE}}) + \sum_{i=1}^{N} \frac{\hat{T}_i(\hat{T}_i + Nk)}{\hat{T}_i - Nk - 2}$$

is smallest, where k is the number of elements in the regression vector $\boldsymbol{\beta}$. Other alternatives are to use Bayesian methods to find the value of N that leads to the largest value for the marginal likelihood (Chib, 1998) or the highest Bayes factor (Koop and Potter, 1999).

Another promising test of the null hypothesis of no change in regime was developed by Carrasco et al. (2014). Let $\ell_t = \log f(y_t | \boldsymbol{\Omega}_{t-1}; \boldsymbol{\lambda})$ be the log of the predictive density of the tth observation under the null hypothesis of no switching. For the Markov-switching regression (36), $\boldsymbol{\lambda}$ would correspond to the fixed-regime regression coefficients and variance $(\boldsymbol{\beta}', \sigma^2)'$:

$$\ell_t = -(1/2)\log(2\pi\sigma^2) - \frac{(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2}{2\sigma^2}.$$

Define $\mathbf{h}_t$ to be the derivative of the log density with respect to the parameter vector,

$$\mathbf{h}_t = \left.\frac{\partial \ell_t}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}_0}$$

where $\hat{\boldsymbol{\lambda}}_0$ denotes the MLE under the null hypothesis of no change in regime. For example,

$$\mathbf{h}_t = \begin{bmatrix} \dfrac{(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}})\mathbf{x}_t}{\hat{\sigma}^2} \\ -\dfrac{1}{2\hat{\sigma}^2} + \dfrac{(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}})^2}{2\hat{\sigma}^4} \end{bmatrix}$$

where $\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\sum_{t=1}^{T}\mathbf{x}_t y_t\right)$ and $\hat{\sigma}^2 = T^{-1}\sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}})^2$. To implement the Carrasco et al. (2014) test of the null hypothesis of no change in regime against the alternative that the first element of $\boldsymbol{\beta}$ switches according to a Markov chain, let $\ell_t^{(1)}$ denote the first element of $\mathbf{h}_t$ and calculate

$$\ell_t^{(2)} = \left.\frac{\partial^2 \ell_t}{\partial \lambda_1^2}\right|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}_0}$$

$$\gamma_t(\rho) = \ell_t^{(2)} + \left[\ell_t^{(1)}\right]^2 + 2\sum_{s<t}\rho^{t-s}\ell_t^{(1)}\ell_s^{(1)}$$

where ρ is an unknown parameter characterizing the persistence of the Markov chain. We then regress $(1/2)\gamma_t(\rho)$ on $\mathbf{h}_t$, save the residuals $\hat{\varepsilon}_t(\rho)$, and calculate

$$C(\rho) = \frac{1}{2}\left[\max\left\{0, \frac{\sum_{t=1}^{T}\gamma_t(\rho)}{2\sqrt{\sum_{t=1}^{T}[\hat{\varepsilon}_t(\rho)]^2}}\right\}\right]^2.$$

We then find the value ρ^* that maximizes $C(\rho)$ over some range (eg, $\rho \in [0.2,0.8]$) and bootstrap to see if $C(\rho^*)$ is statistically significant. This is done by generating data with no changes in regime using the MLE $\boldsymbol{\lambda} = \hat{\lambda}_0$ and calculating $C(\rho^*)$ on each generated sample.

Another option is to conduct generic tests developed by Hamilton (1996) of the hypothesis that an N-regime model accurately describes the data. For example, if the model is correctly specified, the derivative of the log of the predictive density with respect to any element of the parameter vector,

$$\left.\frac{\partial \log p(\mathbf{y}_t|\boldsymbol{\Omega}_{t-1};\boldsymbol{\lambda})}{\partial \lambda_i}\right|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}_{\mathrm{MLE}}},$$

should be impossible to predict from its own lagged values, a hypothesis that can be tested using simple regressions.

2.12 Deterministic Breaks

Another common approach is to treat the changes in regime as deterministic rather than random. If we wanted to test the null hypothesis of constant coefficients against the alternative that a certain subset of the coefficients of a regression switched at fixed known dates $t_1, t_2, \ldots, t_N$, we could do this easily enough using a standard F test (see, for example, Fisher, 1970). If we do not know the dates, we could calculate the value of the F statistic for every set of allowable N partitions, efficient algorithms for which have been described by Bai and Perron (2003) and Doan (2012), with critical values for interpreting the supremum of the F statistics provided by Bai and Perron (1998). Bai and Perron (1998) also described a sequential procedure with which one could first test the null hypothesis of no breaks against the alternative of $N = 1$ break, and then test $N = 1$ against $N = 2$, and so on.

Although simpler to deal with econometrically, deterministic structural breaks have the drawback that they are difficult to incorporate in a sensible way into models based on rational decision makers. Neither the assumption that people knew perfectly that the change was coming years in advance, nor the assumption that they were certain that nothing would ever change (when in the event the change did indeed appear) is very appealing. There is further the practical issue of how users of such econometric models are supposed to form their own future forecasts. Pesaran and Timmermann (2007) suggested estimating models over windows of limited subsamples, watching the data for an indication that it is time to switch to using a new model. Another drawback of interpreting structural breaks as deterministic events is that such approaches make no use of the fact that regimes such as business downturns may be a recurrent event.

2.13 Chib's Multiple Change-Point Model

Chib (1998) offered a way to interpret multiple change-point models that gets around some of the awkward features of deterministic structural breaks. Chib's model assumes that when the process is in regime i, the conditional density of the data is governed by parameter vector $\boldsymbol{\theta}_i$ as in (13). Chib assumed that the process begins at date 1 in regime $s_t = 1$ and parameter vector $\boldsymbol{\theta}_1$, and will stay there the next period with probability p_{11}. With probability $1 - p_{11}$ we get a new value $\boldsymbol{\theta}_2$, drawn perhaps from an $N(\boldsymbol{\theta}_1, \boldsymbol{\Sigma})$ distribution. Conditional on knowing that there were N such breaks, this could be viewed as a special case of an N-state Markov-switching model with transition probability matrix taking the form

$$\mathbf{P} = \begin{bmatrix} p_{11} & 0 & 0 & \cdots & 0 & 0 \\ 1-p_{11} & p_{22} & 0 & \cdots & 0 & 0 \\ 0 & 1-p_{22} & p_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_{N-1,N-1} & 0 \\ 0 & 0 & 0 & \cdots & 1-p_{N-1,N-1} & 1 \end{bmatrix}.$$

The total number of regime changes N could then be selected using one of the methods discussed above.

Again it is not clear how to form out-of-sample forecasts with this specification. Pesaran et al. (2006) proposed embedding Chib's model within a hierarchical prior with which one could forecast future changes in regime based on the size and duration of past breaks.

2.14 Smooth Transition Models

Another econometric approach to changes in regime is the smooth transition regression model (Teräsvirta, 2004):

$$y_t = \frac{\exp[-\gamma(z_{t-1} - c)]}{1 + \exp[-\gamma(z_{t-1} - c)]}\mathbf{x}'_{t-1}\boldsymbol{\beta}_1 + \frac{1}{1 + \exp[-\gamma(z_{t-1} - c)]}\mathbf{x}'_{t-1}\boldsymbol{\beta}_2 + u_t. \tag{37}$$

Here the scalar z_{t-1} could be one of the elements of $\mathbf{x}_{t-1}$ or some known function of $\mathbf{x}_{t-1}$. For $\gamma > 0$, as $z_{t-1} \to -\infty$, the regression coefficients go to $\boldsymbol{\beta}_1$, while when $z_{t-1} \to \infty$, the regression coefficients approach $\boldsymbol{\beta}_2$. The parameter γ governs how quickly the coefficients transition as z_{t-1} crosses the threshold c.

If $\mathbf{x}_{t-1} = (y_{t-1}, y_{t-2}, \ldots, y_{t-r})'$, this is Teräsvirta's (1994) smooth-transition autoregression, for which typically $z_{t-1} = y_{t-d}$ for some lag d. More generally, given a data-generating process for $\mathbf{x}_t$, (37) is a fully specified time-series process for which forecasts at any horizon can be calculated by simulation. One important challenge is how to choose the lag d or more generally the switching variable z_{t-1}. Although in some settings the forecast might be similar to that coming from (6), the weights $\text{Prob}(s_{t-1} = i|\boldsymbol{\Omega}_{t-1})$ in the latter would be a function of the entire history $\{y_{t-1}, y_{t-2}, \ldots, y_1\}$ rather than any single value.

3. ECONOMIC THEORY AND CHANGES IN REGIME

The previous section discussed econometric issues associated with analyzing series subject to changes in regime. This section reviews how these features can appear in theoretical models of the economy.

3.1 Closed-Form Solution of DSGEs and Asset-Pricing Implications

In some settings it is possible to find exact analytical solutions for a full dynamic stochastic general equilibrium model subject to changes in regime. A standard first-order condition in many macro models holds that

$$U'(C_t) = \beta E_t[U'(C_{t+1})(1 + r_{j,t+1})] \tag{38}$$

where C_t denotes consumption of a representative consumer, β a time-discount rate, and $r_{j,t+1}$ the real return on asset j between t and $t + 1$.Lucas (1978) proposed a particularly simple setting in which aggregate output comes solely from nonreproducible assets (sometimes thought of as fruit coming from trees) for which equilibrium turns out to require that C_t equals the aggregate real dividend D_t paid on equities (or the annual crop of fruit). If the utility function exhibits constant relative risk aversion ($U(C) = (1+\gamma)^{-1} C^{(1+\gamma)}$), the aggregate equilibrium real stock price must satisfy

$$P_t = D_t^{-\gamma} \sum_{k=1}^{\infty} \beta^k E_t D_{t+k}^{(1+\gamma)}.$$

Since the dividend process $\{D_{t+k}\}$ is exogenous in this model, one could simply assume that the change in the log of D_t is characterized by a process such as (1). Cecchetti et al. (1990) used calculations related to those in (12) to find the closed-form solution for the general-equilibrium stock price,

$$P_t = \rho_{s_t} D_t,$$

where the values of ρ_1 and ρ_2 are given in equations (11) and (12) in their paper.

Lucas's assumption of an exogenous consumption and dividend process is obviously quite restrictive. Nevertheless, asset-pricing relations such as (38) have to hold regardless of how we close the rest of the model. We can always use (38) or other basic asset-pricing conditions along with an assumed process for returns to find the implications of changes in regime for financial variables in more general settings. There is a very large literature investigating these issues, covering topics such as portfolio allocation (Ang and Bekaert, 2002a; Guidolin and Timmermann, 2008), financial implications of rare-event risk (Barro, 2006; Evans, 1996), option pricing (Elliott et al., 2005), and the term structure of interest rates (Ang and Bekaert, 2002b; Bansal and Zhou, 2002; Bekaert et al., 2001). For a survey of this literature, see Ang and Timmermann (2012).

3.2 Approximating the Solution to DSGEs Using Perturbation Methods

First-order conditions for a much broader class of dynamic stochastic general equilibrium models with Markov regime-switching take the form

$$E_t \mathbf{a}(\mathbf{y}_{t+1}, \mathbf{y}_t, \mathbf{x}_t, \mathbf{x}_{t-1}, \varepsilon_{t+1}, \varepsilon_t, \boldsymbol{\theta}_{s_{t+1}}, \boldsymbol{\theta}_{s_t}) = \mathbf{0}. \tag{39}$$

Here $\mathbf{a}(.)$ is an $[(n_y + n_x) \times 1]$ vector-valued function, $\mathbf{y}_t$ an $(n_y \times 1)$ vector of control variables (also sometimes referred to as endogenous jump variables), $\mathbf{x}_t$ an $(n_x \times 1)$ vector of predetermined endogenous or exogenous variables, ε_t an $(n_\varepsilon \times 1)$ vector of innovations to those elements of $\mathbf{x}_t$ that are exogenous to the model, and s_t follows an N-state Markov chain. The example considered in the previous subsection is a special case of such a system with $n_y = n_x = 1, y_t = P_t/D_t$, $x_t = \ln(D_t/D_{t-1})$, $\theta_{s_t} = m_{s_t}$, and[d]

[d] Notice (38) can be written

$$D_t^{\gamma} = \beta E_t \left[D_{t+1}^{\gamma} \frac{P_{t+1} + D_{t+1}}{P_t} \right]$$
$$1 = \beta E_t \left[\left(\frac{D_{t+1}}{D_t} \right)^{\gamma} \left(\frac{(P_{t+1}/D_{t+1}) + 1}{P_t/D_t} \right) \left(\frac{D_{t+1}}{D_t} \right) \right].$$

$$\mathbf{a}(y_{t+1}, y_t, x_t, x_{t-1}, \varepsilon_{t+1}, \varepsilon_t, m_{s_{t+1}}, m_{s_t})$$
$$= \begin{bmatrix} \beta \exp[(1+\gamma)(m_{s_{t+1}} + \varepsilon_{t+1})][(y_{t+1}+1)/y_t] - 1 \\ x_t - m_{s_t} - \varepsilon_t \end{bmatrix}.$$

For that example we were able to find closed-form solutions of the form

$$\mathbf{y}_t = \boldsymbol{\rho}_{s_t}(\mathbf{x}_{t-1}, \boldsymbol{\varepsilon}_t)$$
$$\mathbf{x}_t = \mathbf{h}_{s_t}(\mathbf{x}_{t-1}, \boldsymbol{\varepsilon}_t),$$

namely $y_t = \rho_{s_t}$ and $x_t = m_{s_t} + \varepsilon_t$.

For more complicated models, solutions cannot be found analytically but can be approximated using the partition perturbation method developed by Foerster et al. (forthcoming). Their method generalizes the now-standard perturbation methods of Schmitt-Grohe and Uribe (2004) for finding linear and higher-order approximations to the solutions to DSGEs with no regime switching. Foerster et al.'s idea is to approximate the solutions $\boldsymbol{\rho}_j(.)$ and $\mathbf{h}_j(.)$ in a neighborhood around the deterministic steady-state values satisfying $\mathbf{a}(\mathbf{y}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{x}^*, \mathbf{0}, \mathbf{0}, \boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = \mathbf{0}$ where $\boldsymbol{\theta}^*$ is the unconditional expectation of $\boldsymbol{\theta}_{s_t}$ calculated from the ergodic probabilities of the Markov chain,

$$\boldsymbol{\theta}^* = \sum_{j=1}^{N} \boldsymbol{\theta}_j \text{Prob}(s_t = j).$$

For the Lucas tree example from the previous subsection, $m^* = (m_1 p_{21} + m_2 p_{12})/(p_{12} + p_{21})$. We then think of a sequence of economies indexed by a continuous scalar χ such that their behavior as $\chi \to 0$ approaches the steady state, while the value at $\chi = 1$ is exactly that implied by (39):

$$\mathbf{y}_t = \boldsymbol{\rho}_{s_t}(\mathbf{x}_{t-1}, \boldsymbol{\varepsilon}_t, \chi) \tag{40}$$
$$\mathbf{x}_t = \mathbf{h}_{s_t}(\mathbf{x}_{t-1}, \boldsymbol{\varepsilon}_t, \chi). \tag{41}$$

As $\chi \to 0$, the randomness coming from $\boldsymbol{\varepsilon}_t$ is suppressed, and it turns out to be necessary to do the same thing for any elements of $\boldsymbol{\theta}$ that influence the steady state in order to have some fixed point around which to calculate the approximation. For elements in $\boldsymbol{\theta}_{s_t}$ that may change with regime but do not matter for the steady state, Foerster et al. (forthcoming) showed that it is not necessary to shrink by χ in order to approximate the dynamic solution. The authors thus specified

$$\boldsymbol{\theta}(s_t, \chi) = \begin{bmatrix} \boldsymbol{\theta}^{A*} + \chi(\boldsymbol{\theta}^A_{s_t} - \boldsymbol{\theta}^{A*}) \\ \boldsymbol{\theta}^B_{s_t} \end{bmatrix}$$

where $\boldsymbol{\theta}^A_{s_t}$ denotes the subset of elements of $\boldsymbol{\theta}_{s_t}$ that influence the steady state. The economy characterized by a particular value of χ thus needs to satisfy

$$\mathbf{0} = \int \sum_{j=1}^{N} p_{s_t,j} \mathbf{a}[\boldsymbol{\rho}_j(\mathbf{x}_t, \chi \boldsymbol{\varepsilon}_{t+1}, \chi), \mathbf{y}_t, \mathbf{x}_t, \mathbf{x}_{t-1}, \chi \varepsilon_{t+1}, \varepsilon_t, \boldsymbol{\theta}(j,\chi), \boldsymbol{\theta}(s_t,\chi)] dF(\boldsymbol{\varepsilon}_{t+1}) \tag{42}$$

where $F(\boldsymbol{\varepsilon}_{t+1})$ denotes the cumulative distribution function for ε_{t+1}. Note (42) is satisfied by construction when evaluated at $\mathbf{y}_t = \mathbf{y}^*$, $\mathbf{x}_t = \mathbf{x}_{t-1} = \mathbf{x}^*$, $\boldsymbol{\varepsilon}_t = \mathbf{0}$, and $\chi = 0$.

We next substitute (40) and (41) into (42) to arrive at a system of $N(n_y + n_x)$ equations of the form

$$\mathbf{Q}_{s_t}(\mathbf{x}_{t-1}, \boldsymbol{\varepsilon}_t, \chi) = \mathbf{0} \quad s_t = 1, \ldots, N$$

which have to hold for all $\mathbf{x}_{t-1}$, $\boldsymbol{\varepsilon}_t$, and χ. Taking derivatives with respect to $\mathbf{x}_{t-1}$ and evaluating at $\mathbf{x}_{t-1} = \mathbf{x}^*$, $\boldsymbol{\varepsilon}_t = \mathbf{0}$, and $\chi = 0$ (that is, using a first-order Taylor approximation around the steady state) yields a system of $N(n_y + n_x)n_x$ quadratic polynomial equations in the $N(n_y + n_x)n_x$ unknowns corresponding to elements of the matrices

$$\underset{(n_y \times n_x)}{\mathbf{R}_j^x} = \left.\frac{\partial \boldsymbol{\rho}_j(\mathbf{x}_{t-1}, \boldsymbol{\varepsilon}_t, \chi)}{\partial \mathbf{x}'_{t-1}}\right|_{\mathbf{x}_{t-1} = \mathbf{x}^*, \boldsymbol{\varepsilon}_t = \mathbf{0}, \chi = 0} \quad j = 1, \ldots, N$$

$$\underset{(n_x \times n_x)}{\mathbf{H}_j^x} = \left.\frac{\partial \mathbf{h}_j(\mathbf{x}_{t-1}, \boldsymbol{\varepsilon}_t, \chi)}{\partial \mathbf{x}'_{t-1}}\right|_{\mathbf{x}_{t-1} = \mathbf{x}^*, \boldsymbol{\varepsilon}_t = \mathbf{0}, \chi = 0} \quad j = 1, \ldots, N.$$

The authors proposed an algorithm for finding the solution to this system of equations, that is, values for the above sets of matrices. Given these, other terms in the first-order Taylor approximation to (42) produce a system of $N(n_y + n_x)n_\varepsilon$ equations that are linear in known parameters and the unknown elements of

$$\underset{(n_y \times n_\varepsilon)}{\mathbf{R}_j^\varepsilon} = \left.\frac{\partial \boldsymbol{\rho}_j(\mathbf{x}_{t-1}, \boldsymbol{\varepsilon}_t, \chi)}{\partial \varepsilon'_t}\right|_{\mathbf{x}_{t-1} = \mathbf{x}^*, \boldsymbol{\varepsilon}_t = \mathbf{0}, \chi = 0} \quad j = 1, \ldots, N$$

$$\underset{(n_x \times n_\varepsilon)}{\mathbf{H}_j^\varepsilon} = \left.\frac{\partial \mathbf{h}_j(\mathbf{x}_{t-1}, \boldsymbol{\varepsilon}_t, \chi)}{\partial \varepsilon'_t}\right|_{\mathbf{x}_{t-1} = \mathbf{x}^*, \boldsymbol{\varepsilon}_t = \mathbf{0}, \chi = 0} \quad j = 1, \ldots, N,$$

from which $\mathbf{R}_j^\varepsilon$ and $\mathbf{H}_j^\varepsilon$ are readily calculated. Another system of $N(n_y + n_x)$ linear equations yields

$$\underset{(n_y \times 1)}{\mathbf{R}_j^\chi} = \left.\frac{\partial \boldsymbol{\rho}_j(\mathbf{x}_{t-1}, \boldsymbol{\varepsilon}_t, \chi)}{\partial \chi}\right|_{\mathbf{x}_{t-1} = \mathbf{x}^*, \boldsymbol{\varepsilon}_t = \mathbf{0}, \chi = 0} \quad j = 1, \ldots, N$$

$$\underset{(n_x \times 1)}{\mathbf{H}_j^\chi} = \left.\frac{\partial \mathbf{h}_j(\mathbf{x}_{t-1}, \boldsymbol{\varepsilon}_t, \chi)}{\partial \chi}\right|_{\mathbf{x}_{t-1} = \mathbf{x}^*, \boldsymbol{\varepsilon}_t = \mathbf{0}, \chi = 0} \quad j = 1, \ldots, N.$$

The approximation to the solution to the regime-switching DSGE is then

$$\begin{aligned}\mathbf{y}_t &= \mathbf{y}^* + \mathbf{R}^x_{s_t}(\mathbf{x}_{t-1} - \mathbf{x}^*) + \mathbf{R}^{\varepsilon}_{s_t}\boldsymbol{\varepsilon}_t + \mathbf{R}^{\chi}_{s_t} \\ \mathbf{x}_t &= \mathbf{x}^* + \mathbf{H}^x_{s_t}(\mathbf{x}_{t-1} - \mathbf{x}^*) + \mathbf{H}^{\varepsilon}_{s_t}\boldsymbol{\varepsilon}_t + \mathbf{H}^{\chi}_{s_t}.\end{aligned}$$

One could then go a step further if desired, taking a second-order Taylor approximation to (42). Once the first step (the linear approximation) has been completed, the second step (quadratic approximation) is actually easier to calculate numerically than the first step was, because all the second-step equations turn out to be linear in the remaining unknown magnitudes.

Lind (2014) developed an extension of this approach that could be used to form approximations to any model characterized by dramatic nonlinearities, even if regime-switching in the form of (39) is not part of the maintained structure. For example, the economic relations may change significantly when interest rates are at the zero lower bound. Lind's idea is to approximate the behavior of a nonlinear model over a set of discrete regions using relations that are linear (or possibly higher-order polynomials) over individual regions, from which one can then use many of the tools discussed above for economic and econometric analysis.

3.3 Linear Rational Expectations Models with Changes in Regime

Economic researchers often use a linear special case of (39) which in the absence of regime shifts takes the form

$$\mathbf{A}E(\mathbf{y}_{t+1}|\boldsymbol{\Omega}_t) = \mathbf{d} + \mathbf{B}\mathbf{y}_t + \mathbf{C}\mathbf{x}_t \tag{43}$$

$$\mathbf{x}_t = \mathbf{c} + \boldsymbol{\Phi}\mathbf{x}_{t-1} + \mathbf{v}_t$$

for $\mathbf{y}_t$ an $(n_y \times 1)$ vector of endogenous variables, $\boldsymbol{\Omega}_t = \{\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_1\}$,$\mathbf{x}_t$ an $(n_x \times 1)$ vector of exogenous variables, and $\mathbf{v}_t$ a martingale difference sequence. Such a system might have been obtained as an approximation to the first-order conditions for a nonlinear DSGE using the standard perturbation algorithm, or often is instead simply postulated as the primitive conditions of the model of interest. If $\mathbf{A}^{-1}$ exists and the number of eigenvalues of $\mathbf{A}^{-1}\mathbf{B}$ whose modulus is less than or equal to unity is equal to the number of predetermined endogenous variables, then a unique stable solution can be found of the form[e]

$$\begin{aligned}\mathbf{k}_{t+1} &= \mathbf{h}_{k0} + \mathbf{H}_{kk}\mathbf{k}_t + \mathbf{H}_{kx}\mathbf{x}_t \\ \mathbf{d}_t &= \mathbf{h}_{d0} + \mathbf{H}_{dk}\mathbf{k}_t + \mathbf{H}_{dx}\mathbf{x}_t\end{aligned}$$

[e] Klein (2000) generalized to the case when $\mathbf{A}$ may not be invertible.

where $\mathbf{k}_t$ denotes the elements of $\mathbf{y}_t$ that correspond to predetermined variables, while $\mathbf{d}_t$ collects the control or jump variables. Algorithms for finding the values of the parameters $\mathbf{h}_{i0}$ and $\mathbf{H}_{ij}$ have been developed by Blanchard and Kahn (1980), Klein (2000), and Sims (2001).

We could also generalize (43) to allow for changes in regime,

$$\mathbf{A}_{s_t} E(\mathbf{y}_{t+1} | \Omega_t, s_t, s_{t-1}, \ldots, s_1) = \mathbf{d}_{s_t} + \mathbf{B}_{s_t} \mathbf{y}_t + \mathbf{C}_{s_t} \mathbf{x}_t \tag{44}$$

where s_t follows an exogenous N-state Markov chain and $\mathbf{A}_j$ denotes an $(n_y \times n_y)$ matrix of parameters when the regime for date t is given by $s_t = j$. To solve such a model, Davig and Leeper (2007) suggested exploiting the feature that conditional on $\mathcal{S} = \{s_t\}_{t=1}^{\infty}$ the model is linear. Let $\mathbf{y}_{jt}$ correspond to the value of $\mathbf{y}_t$ when $s_t = j$ and collect the set of such vectors for all the possible regimes in a larger vector $\mathbf{Y}_t$:

$$\underset{(Nn_y \times 1)}{\mathbf{Y}_t} = \begin{bmatrix} \underset{(n_y \times 1)}{\mathbf{y}_{1t}} \\ \vdots \\ \underset{(n_y \times 1)}{\mathbf{y}_{Nt}} \end{bmatrix}.$$

If we restrict our consideration to solutions that satisfy the minimal-state Markov property, then

$$E(\mathbf{y}_{t+1} | \mathcal{S}, \Omega_t) = E(\mathbf{y}_{t+1} | s_{t+1}, s_t, \Omega_t)$$

and

$$E(\mathbf{y}_{t+1} | s_t = i, \Omega_t) = \sum_{j=1}^{N} E(\mathbf{y}_{t+1} | s_{t+1} = j, s_t = i, \Omega_t) p_{ij}.$$

Hence when $s_t = i$,

$$\mathbf{A}_{s_t} E(\mathbf{y}_{t+1} | s_t, \Omega_t) = (\mathbf{p}_i' \otimes \mathbf{A}_i) E(\mathbf{Y}_{t+1} | \mathbf{Y}_t) \tag{45}$$

where

$$\mathbf{p}_i = \begin{bmatrix} p_{i1} \\ \vdots \\ p_{iN} \end{bmatrix}$$

denotes column i of the Markov transition probabilities, with elements of $\mathbf{p}_i$ summing to unity. Consider then the stacked structural system,

$$\mathbf{A} E(\mathbf{Y}_{t+1} | \mathbf{Y}_t) = \mathbf{d} + \mathbf{B}\mathbf{Y}_t + \mathbf{C}\mathbf{x}_t \tag{46}$$

$$\underset{(Nn_y \times Nn_y)}{\mathbf{A}} = \begin{bmatrix} \underset{(1\times N)}{\mathbf{p}_1'} \otimes \underset{(n_y\times n_y)}{\mathbf{A}_1} \\ \vdots \\ \underset{(1\times N)}{\mathbf{p}_N'} \otimes \underset{(n_y\times n_y)}{\mathbf{A}_N} \end{bmatrix} \quad \underset{(Nn_y\times 1)}{\mathbf{d}} = \begin{bmatrix} \underset{(n_y\times 1)}{\mathbf{d}_1} \\ \vdots \\ \underset{(n_y\times 1)}{\mathbf{d}_N} \end{bmatrix} \tag{47}$$

$$\underset{(Nn_y\times Nn_y)}{\mathbf{B}} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}_N \end{bmatrix} \quad \underset{(Nn_y\times n_x)}{\mathbf{C}} = \begin{bmatrix} \underset{(n_y\times n_x)}{\mathbf{C}_1} \\ \vdots \\ \underset{(n_y\times n_x)}{\mathbf{C}_N} \end{bmatrix}.$$

This is a simple regime-independent system for which a solution can be found using the traditional method. For example, with no predetermined variables, if all eigenvalues of $\mathbf{A}^{-1}\mathbf{B}$ are outside the unit circle, then we can find a unique stable solution of the form

$$\underset{(Nn_y\times 1)}{\mathbf{Y}_t} = \underset{(Nn_y\times 1)}{\mathbf{h}} + \underset{(Nn_y\times n_x)}{\mathbf{H}} \underset{(n_x\times 1)}{\mathbf{x}_t} \tag{48}$$

which implies that

$$\underset{(n_y\times 1)}{\mathbf{y}_t} = \underset{(n_y\times 1)}{\mathbf{h}_{s_t}} + \underset{(n_y\times n_x)}{\mathbf{H}_{s_t}} \underset{(n_x\times 1)}{\mathbf{x}_t} \tag{49}$$

for $\mathbf{h}_i$ and $\mathbf{H}_i$ the *i*th blocks of $\mathbf{h}$ and $\mathbf{H}$, respectively. If (48) is a solution to (46), then (49) is a solution to (44).[f]

[f] If (49) holds, then

$$E(\mathbf{y}_{t+1}|\Omega_t, s_t = i) = \sum_{j=1}^{N} p_{ij}[\mathbf{h}_j + \mathbf{H}_j(\mathbf{c} + \mathbf{\Phi}\mathbf{x}_t)].$$

Thus for (44) to hold it must be the case that for each $i = 1,\ldots,N$,

$$\mathbf{A}_i \sum_{j=1}^{N} p_{ij}\mathbf{H}_j\mathbf{\Phi} = \mathbf{B}_i\mathbf{H}_i + \mathbf{C}_i$$
$$\mathbf{A}_i \sum_{j=1}^{N} p_{ij}(\mathbf{h}_j + \mathbf{H}_j\mathbf{c}) = \mathbf{d}_i + \mathbf{B}_i\mathbf{h}_i.$$

But if (48) is a solution to (46), then

$$\mathbf{A}[\mathbf{h} + \mathbf{H}(\mathbf{c} + \mathbf{\Phi}\mathbf{x}_t)] = \mathbf{d} + \mathbf{B}(\mathbf{h} + \mathbf{H}\mathbf{x}_t) + \mathbf{C}\mathbf{x}_t$$

block *i* of which requires from (47) that

$$(\mathbf{p}_i' \otimes \mathbf{A}_i)\mathbf{H}\mathbf{\Phi} = \mathbf{B}_i\mathbf{H}_i + \mathbf{C}_i$$
$$(\mathbf{p}_i' \otimes \mathbf{A}_i)(\mathbf{h} + \mathbf{H}\mathbf{c}) = \mathbf{d}_i + \mathbf{B}_i\mathbf{h}_i$$

as were claimed to hold.

However, Farmer et al. (2010) demonstrated that while (48) yields one stable solution to (44), it need not be the only stable solution. For further discussion, see Farmer et al. (2009).

3.4 Multiple Equilibria

Other economists have argued that models in which there are multiple possible solutions—for example, system (43) with no predetermined variables and an eigenvalue of $\mathbf{A}^{-1}\mathbf{B}$ inside the unit circle—are precisely those we should be most interested in, given the perception that sometimes consumers or firms seem to become highly pessimistic for no discernible reason, bringing the economy into a self-fulfilling downturn; see Benhabib and Farmer (1999) for a survey of this literature. One factor that could produce multiple equilibria is coordination externalities. The rewards to me of participating in a market may be greatest when I expect large numbers of others to do the same (Cooper, 1994; Cooper and John, 1988). Multiple equilibria could also arise when expectations themselves are a factor determining the equilibrium (Kurz and Motolese, 2001). Kirman (1993) and Chamley (1999) discussed mechanisms by which the economy might tend to oscillate periodically between the possible regimes in multiple-equilibria settings.

A widely studied example is financial market bubbles. In the special case of risk-neutral investors (that is, when $U'(C)$ is some constant independent of consumption C), Eq. (38) relating the price of the stock P_t to its future dividend D_{t+1} becomes

$$P_t = \beta E_t(P_{t+1} + D_{t+1}). \tag{50}$$

One solution to (50) is the market-fundamentals solution given by

$$P_t^* = \sum_{j=1}^{\infty} \beta^j E_t(D_{t+j}).$$

But $P_t = P_t^* + B_t$ also satisfies(50) for B_t any bubble process satisfying $B_t = \beta E_t B_{t+1}$. Hall et al. (1999) proposed an empirical test of whether an observed financial price is occasionally subject to such a bubble regime. This test has been applied in dozens of different empirical studies. However, Hamilton (1985), Driffill and Sola (1998), and Gürkaynak (2008) noted the inherent difficulties in distinguishing financial bubbles from unobserved fundamentals.

3.5 Tipping Points and Financial Crises

In other models, there may be a unique equilibrium but under the right historical conditions, a small change in fundamentals can produce a huge change in observed outcomes. Such dynamics might be well described as locally linear processes that periodically experience changes in regime. Investment dynamics constitute one possible transmission mechanism. The right sequence of events can end up triggering a big investment decline that in turn contributes to a dramatic drop in output and an effective change in regime. Acemoglu and Scott (1997) presented a model where this happens as a result of intertemporal increasing returns, for example, if an investment that leads to a significant new discovery makes additional investments more profitable for a short time. Moore and Schaller (2002), Guo et al. (2005), and Veldkamp (2005) examined different settings in which investment dynamics contribute to tipping points, often through a process of learning about current opportunities. Startz (1998) demonstrated how an accumulation of small shocks could under certain circumstances trigger a dramatic shift between alternative production technologies. Learning by market participants introduces another possible source of tipping-point or regime-shift dynamics (Hong et al., 2007; Branch and Evans, 2010). Gârleanu et al. (2015) demonstrated how tipping points could emerge from the interaction of limited market integration, leveraging, and contagion.

Brunnermeier and Sannikov (2014) developed an intriguing description of tipping points in the context of financial crises. They posited two types of agents, designated "experts" and "households." Experts can invest capital more productively than households, but they are constrained to borrow using only risk-free debt. In normal times, 100% of the economy's equity ends up being held by experts. But as negative shocks cause their net worth to decline, they can end up selling off capital to less productive households, lowering both output and investment. This results in a bimodal stationary distribution in which the economy spends most of its time around the steady state in which experts hold all the capital. But a sequence of negative shocks can lead the economy to become stuck in an inefficient equilibrium from which it can take a long time to recover.

A large number of researchers have used regime-switching models to study financial crises empirically. These include Hamilton's (2005) description of banking crises in the 19th century, Asea and Blomberg's (1998) study of lending cycles in the late 20th century, and an investigation of more recent financial stress by Hubrich and Tetlow (2015).

3.6 Currency Crises and Sovereign Debt Crises

A sudden loss of confidence in a country can lead to a flight from the currency which in turn produces a shock to credit and spending that greatly exacerbates the country's problems. A sudden wave of pessimism could be self-fulfilling, giving rise to multiple equilibria that could exhibit Markov switching (Jeanne and Masson, 2000), or could be characterized by tipping point dynamics where under the right circumstances a small change in fundamentals pushes a country into crisis. Empirical investigations of currency crises using regime-switching models include Peria (2002) and Cerra and Saxena (2005).

Similar dynamics can characterize yields on sovereign debt. If investors lose confidence in a country's ability to service its debt, they will demand a higher interest rate as compensation. The higher interest costs could produce a tipping point that indeed forces a country into default or to make drastic fiscal adjustments (Greenlaw et al., 2013). Analyses of changes in regime in this context include Davig et al. (2011) and Bi (2012).

3.7 Changes in Policy as the Source of Changes in Regime

Another source of a change in regime is a discrete shift in policy itself. One commonly studied possibility is that control of monetary policy may periodically shift between hawks and doves, the latter being characterized by either a higher inflation target or more willingness to tolerate deviations of inflation from target. Analyses using this approach include Owyang and Ramey (2004), Schorfheide (2005), Liu et al. (2011), Bianchi (2013), and Baele et al. (2015).

An alternative possibility is that changes in fiscal regime can be a destabilizing factor. Ruge-Murcia (1995) showed how a lack of credibility of the fiscal stabilization in 1984 contributed to the changes in inflation Israel experienced, while Ruge-Murcia (1999) documented the close connection between changes in fiscal regimes and inflation regimes for Brazil.

4. CONCLUSIONS AND RECOMMENDATIONS FOR RESEARCHERS

We have seen that researchers have a rich set of tools and specifications on which to draw for interpreting data and building economic models for environments in which there may be changes in regime. The chapter closes with some practical recommendations for researchers as to which options are most promising.

Although a researcher might be tempted to use the most general specification possible, with all the parameters changing across a large number of regimes and time-varying transition probabilities, in practice this is usually asking more than the data can deliver. For example, for postwar US data we have only 11 recessions, which economic theory says should be difficult or impossible to predict (Hamilton, 2011). Building a richly parameterized description of the transition into and out of recession could easily result in an overfitted and misspecified model. By contrast, using a simple time-invariant Markov chain is likely to give a reasonable and robust approximation to the key features of the data. Similarly, we know from the analytic characterization of the maximum likelihood estimates (eg, Eq. (24)) that inference about parameters that only show up in regime i can only come from observations within that regime. With postwar quarterly data that would mean about 50 observations from which to estimate all the parameters operating during recessions. One or two parameters could be estimated fairly well, but overfitting is again a potential concern in models with many parameters. For this reason researchers may want to limit the focus to a few of the most important parameters that are likely to change, such as the intercept and the variance.

Where more than two regimes are required, there again are benefits to keeping the model parsimonious. For example, a common finding is that the variance of US output growth permanently decreased in 1984 (McConnell and Perez-Quiros, 2000), while the intercept periodically shifts to negative during recessions. This requires four different regimes—the economy could be in expansion or recession and the date could be before or after the Great Moderation. A useful simplification treats the variance regime as independent of the recession regime, requiring estimation of only 4 transition probabilities rather than 12, as in Kim and Nelson (1999b).

Another feature of which researchers should be aware is that there can be multiple local maxima to the likelihood function. It is therefore good practice to begin the EM iterations from a large number of different starting points to make sure we are always ending up with the same answer, and also as a practical test of whether the algorithm has indeed converged to a fixed point. Likewise with Bayesian methods we want to make sure numerical algorithms converge to the same posterior distribution under alternative starting points and chain dynamics, and the procedure should take into account the label-switching problem.

Provided researchers make note of these issues, these approaches offer a flexible way of modeling some of the key nonlinearities in macroeconomic dynamics without sacrificing the simplicity and tractability of linear models.

APPENDIX

Derivation of EM Equations for Restricted VAR

As noted by Hamilton (1990, p. 47), the M or maximization step of the EM algorithm can be implemented by finding the first-order conditions associated with maximizing the likelihood conditional on a particular set of realizations for the regimes $\mathcal{S}=\{s_1,\ldots,s_T\}$ and then weighting these by the smoothed probability of $\mathcal{S}$ and summing over all the possible realizations of $\mathcal{S}$. For a VAR restricted as in (28), the conditional likelihood is

$$\frac{1}{(2\pi)^{nT/2}|\boldsymbol{\Sigma}|^{T/2}}\exp\left[-(1/2)\sum_{t=1}^{T}(\mathbf{y}_t-\mathbf{A}\mathbf{x}_{1,t-1}-\mathbf{B}_{s_t}\mathbf{x}_{2,t-1})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_t-\mathbf{A}\mathbf{x}_{1,t-1}-\mathbf{B}_{s_t}\mathbf{x}_{2,t-1})\right]$$

with first-order conditions

$$\sum_{t=1}^{T}(\mathbf{y}_t-\mathbf{A}\mathbf{x}_{1,t-1}-\mathbf{B}_{s_t}\mathbf{x}_{2,t-1})\mathbf{x}'_{1,t-1}=\mathbf{0} \qquad \text{(A.1)}$$

$$\sum_{t=1}^{T}(\mathbf{y}_t-\mathbf{A}\mathbf{x}_{1,t-1}-\mathbf{B}_{s_t}\mathbf{x}_{2,t-1})\mathbf{x}'_{2,t-1}\delta(s_t=i)=\mathbf{0}\quad \text{for } i=1,\ldots,N \qquad \text{(A.2)}$$

$$\sum_{t=1}^{T}(1/2)\left[\boldsymbol{\Sigma}-(\mathbf{y}_t-\mathbf{A}\mathbf{x}_{1,t-1}-\mathbf{B}_{s_t}\mathbf{x}_{2,t-1})(\mathbf{y}_t-\mathbf{A}\mathbf{x}_{1,t-1}-\mathbf{B}_{s_t}\mathbf{x}_{2,t-1})'\right]=\mathbf{0} \qquad \text{(A.3)}$$

where $\delta(s_t = i)$ denotes unity if $s_t = i$ and zero otherwise. Stacking (A.1)–(A.2) horizontally gives

$$\sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{A}\mathbf{x}_{1,t-1} - \mathbf{B}_{s_t}\mathbf{x}_{2,t-1})\mathbf{z}'_{t-1} = \sum_{t=1}^{T}(\mathbf{y}_t - [\mathbf{A}\ \ \mathbf{B}_1\ \ \mathbf{B}_2\ \ \cdots\ \ \mathbf{B}_N]\mathbf{z}_{t-1})\mathbf{z}'_{t-1} = \mathbf{0} \tag{A.4}$$

for

$$\underset{[1\times(k_1+Nk_2)]}{\mathbf{z}'_{t-1}} = \left[\mathbf{x}'_{1,t-1}\ \ \mathbf{x}'_{2,t-1}\delta(s_t=1)\ \ \mathbf{x}'_{2,t-1}\delta(s_t=2)\ \ \cdots\ \ \mathbf{x}'_{2,t-1}\delta(s_t=N)\right].$$

Multiplying the tth term within the sums in (A.4) and (A.3) by $\text{Prob}(s_t = i|\boldsymbol{\Omega}_T;\hat{\boldsymbol{\lambda}}^{(\ell)})$, summing over $i = 1,\ldots,N$, and rearranging gives Eqs. (29) and (30).

ACKNOWLEDGMENTS

I thank Marine Carrasco, Steve Davis, Liang Hu, Òscar Jordà, Douglas Steigerwald, John Taylor, Allan Timmermann, and Harald Uhlig for helpful comments on earlier drafts.

REFERENCES

Acemoglu, D., Scott, A., 1997. Asymmetric business cycles: theory and time-series evidence. J. Monet. Econ. 40, 501–533.

Albert, J., Chib, S., 1993. Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. J. Bus. Econ. Stat. 11, 1–15.

Ang, A.A., Bekaert, G., 2002a. International asset allocation with regime shifts. Rev. Financ. Stud. 15, 1137–1187.

Ang, A.A., Bekaert, G., 2002b. Regime switches in interest rates. J. Bus. Econ. Stat. 20, 163–182.

Ang, A.A., Timmermann, A., 2012. Regime changes and financial markets. Ann. Rev. Financ. Econ. 4, 313–337.

Asea, P.K., Blomberg, B., 1998. Lending cycles. J. Econom. 83, 89–128.

Auerbach, A., Gorodnichenko, Y., 2012. Measuring the output responses to fiscal policy. Am. Econ. J. Macroecon. 4, 1–27.

Baele, L., Bekaert, G., Cho, S., Inghelbrecht, K., Moreno, A., 2015. Macroeconomic regimes. J. Monet. Econ. 70, 51–71.

Bai, J., Perron, P., 1998. Testing for and estimation of multiple structural changes. Econometrica 66, 47–78.

Bai, J., Perron, P., 2003. Computation and analysis of multiple structural change models. J. Appl. Econ. 18, 1–22.

Bansal, R., Zhou, H., 2002. Term structure of interest rates with regime shifts. J. Financ. 57, 1997–2042.

Barro, R.J., 2006. Rare disasters and asset markets in the twentieth century. Q. J. Econ. 121, 823–866.

Bekaert, G., Hodrick, R.J., Marshall, D., 2001. Peso problem explanations for term structure anomalies. J. Monet. Econ. 48, 241–270.

Benhabib, J., Farmer, R.E.A., 1999. Indeterminacy and sunspots in macroeconomics. In: Taylor, J., Woodford, M. (Eds.), Handbook of Macroeconomics. In: vol. 1. North Holland, Amsterdam.

Bi, H., 2012. Sovereign default, risk premia, fiscal limits, and fiscal policy. Eur. Econ. Rev. 56, 389–410.

Bianchi, F., 2013. Regime switches, agents' beliefs and post-World War II U.S. macroeconomic dynamics. Rev. Econ. Stud. 80, 463–490.

Blanchard, O.J., Kahn, C.M., 1980. The solution of linear difference models under rational expectations. Econometrica 48, 1305–1317.

Branch, W.A., Evans, G.W., 2010. Asset return dynamics and learning. Rev. Financ. Stud. 23, 1651–1680.

Brunnermeier, M.K., Sannikov, Y., 2014. A macroeconomic model with a financial sector. Am. Econ. Rev. 104, 379–421.

Calvet, L., Fisher, A., 2004. How to forecast long-run volatility: regime-switching and the estimation of multifractal processes. J. Financ. Econ. 2, 49–83.

Camacho, M., Perez-Quiros, G., Poncela, P., 2014. Green shoots and double dips in the Euro are: a real time measure. Int. J. Forecast. 30, 520–535.

Carrasco, M., Hu, L., Ploberger, W., 2014. Optimal test for Markov switching. Econometrica 82, 765–784.

Carter, A.V., Steigerwald, D.G., 2012. Testing for regime switching: a comment. Econometrica 80, 1809–1812.

Carter, A.V., Steigerwald, D.G., 2013. Markov regime-switching tests: asymptotic critical values. J. Econ. Methods 2, 25–34.

Cecchetti, S.G., Lam, P.S., Mark, N.C., 1990. Mean reversion in equilibrium asset prices. Am. Econ. Rev. 80, 398–418.

Celeux, G., Hurn, M., Robert, C.P., 2000. Computational and inferential difficulties with mixture posterior distributions. J. Am. Stat. Assoc. 95, 957–970.

Cerra, V., Saxena, S.C., 2005. Did output recover from the Asian crisis? IMF Staff Pap. 52, 1–23.

Chamley, C., 1999. Coordinating regime switches. Q. J. Econ. 114, 869–905.

Chauvet, M., 1998. An economic characterization of business cycle dynamics with factor structure and regime switches. Int. Econ. Rev. 39, 969–996.

Chauvet, M., Hamilton, J.D., 2006. Dating business cycle turning points. In: Costas Milas, P.R., van Dijk, D. (Eds.), Nonlinear Analysis of Business Cycles. Elsevier, Amsterdam, pp. 1–54.

Chauvet, M., Piger, J., 2008. A comparison of the real-time performance of business cycle dating methods. J. Bus. Econ. Stat. 26, 42–49.

Chen, J., Li, P., 2009. Hypothesis test for normal mixture models: the EM approach. Ann. Stat. 37, 2523–2542.

Chib, S., 1998. Estimation and comparison of multiple change-point models. J. Econom. 86, 221–241.

Cho, J.S., White, H., 2007. Testing for regime switching. Econometrica 75, 1671–1720.

Cooper, R., 1994. Equilibrium selection in imperfectly competitive economics with multiple equilibria. Econ. J. 104, 1106–1122.

Cooper, R., John, A., 1988. Coordinating coordination failures in Keynesian models. Q. J. Econ. 103, 441–463.

Davig, T., Leeper, E.M., 2007. Generalizing the Taylor principle. Am. Econ. Rev. 97, 607–635.

Davig, T., Leeper, E.M., Walker, T.B., 2011. Inflation and the fiscal limit. Eur. Econ. Rev. 55, 31–47.

DeGroot, M.H., 1970. Optimal Statistical Decisions. McGraw-Hill, New York.

Diebold, F.X., Lee, J.H., Weinbach, G.C., 1994. Regime switching with time-varying transition probabilities. In: Hargreaves, C.P. (Ed.), Nonstationary Time Series Analysis and Cointegration. Oxford University Press, Oxford.

Doan, T., 2012. RATS User's Guide, Version 8.2. http://www.estima.com.

Driffill, J., Sola, M., 1998. Intrinsic bubbles and regime-switching. J. Monet. Econ. 42, 357–374.

Dueker, M., 1997. Markov switching in GARCH processes and mean-reverting stock-market volatility. J. Bus. Econ. Stat. 15, 26–34.

Elliott, R.J., Chan, L., Siu, T.K., 2005. Option pricing and Esscher transform under regime switching. Ann. Finance 1, 423–432.

Evans, M.D.D., 1996. Peso problems: their theoretical and empirical implications. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics, vol. 14. Elsevier, Amsterdam.

Farmer, R.E.A., Waggoner, D.F., Zha, T., 2009. Understanding Markov-switching rational expectations models. J. Econ. Theory 144, 1849–1867.

Farmer, R.E.A., Waggoner, D.F., Zha, T., 2010. Generalizing the Taylor principle: comment. Am. Econ. Rev. 100, 608–617.

Filardo, A.J., 1994. Business cycle phases and their transitional dynamics. J. Bus. Econ. Stat. 12, 299–308.

Filardo, A.J., Gordon, S.F., 1998. Business cycle durations. J. Econom. 85, 99–123.

Fisher, F.M., 1970. Tests of equality between sets of coefficients in two linear regressions: an expository note. Econometrica 38, 361–366.

Foerster, A., Rubio-Ramirez, J., Waggoner, D.F., Zha, T., forthcoming. Perturbation methods for Markov-switching DSGE models. Quant. Econom.

Francq, C., Zakoïan, J.M., 2001. Stationarity of multivariate Markov-switching ARMA models. J. Econom. 102, 339–364.

Frühwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. J. Am. Stat. Assoc. 96, 194–209.

Garcia, R., 1998. Asymptotic null distribution of the likelihood ratio test in Markov switching models. Int. Econ. Rev. 39, 763–788.

Gârleanu, N., Panageas, S., Yu, J., 2015. Financial entanglement: a theory of incomplete integration, leverage, crashes, and contagion. Am. Econ. Rev. 105, 1979–2010.

Geweke, J., 2007. Interpretation and inference in mixture models: simple MCMC works. Comput. Stat. Data Anal. 51, 3529–3550.

Greenlaw, D., Hamilton, J.D., Hooper, P., Mishkin, F., 2013. Crunch time: fiscal crises and the role of monetary policy. In: Inproceedings of the U.S. Monetary Policy Forum 2013. Chicago Booth School of Business: Initiative on Global Markets, pp. 3–58.

Guidolin, M., Timmermann, A., 2008. International asset allocation under regime switching, skew, and kurtosis preferences. Rev. Financ. Stud. 21, 889–935.

Guo, X., Miao, J., Morelle, E., 2005. Irreversible investment with regime shifts. J. Econ. Theory 122, 37–59.

Gürkaynak, R.S., 2008. Econometric tests of asset price bubbles: taking stock. J. Econ. Surv. 22, 166–186.

Hall, P., Stewart, M., 2005. Theoretical analysis of power in a two-component normal mixture model. J. Stat. Plan. Inference 134, 158–179.

Hall, S.G., Psaradakis, Z., Sola, M., 1999. Detecting periodically collapsing bubbles: a Markov-switching unit root test. J. Appl. Econ. 14, 43–154.

Hamilton, J.D., 1985. On testing for self-fulfilling speculative price bubbles. Int. Econ. Rev. 27, 545–552.

Hamilton, J.D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57, 357–384.

Hamilton, J.D., 1990. Analysis of time series subject to changes in regime. J. Econom. 45, 39–70.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, NJ.

Hamilton, J.D., 1996. Specification testing in Markov-switching time-series models. J. Econom. 70, 127–157.

Hamilton, J.D., 2005. What's real about the business cycle? Fed. Reserve Bank St. Louis Rev. 87, 435–452.

Hamilton, J.D., 2011. Calling recessions in real time. Int. J. Forecast. 27, 1006–1026.

Hamilton, J.D., Wu, J.C., 2012. Identification and estimation of Gaussian affine term structure models. J. Econom. 168, 315–331.

Hansen, B.E., 1992. The likelihood ratio test under non-standard conditions. J. Appl. Econ. 7, S61–S82. Erratum, 1996, 11, 195–198.

Hong, H., Stein, J.C., Yu, J., 2007. Simple forecasts and paradigm shifts. J. Financ. 62, 1207–1242.

Hubrich, K., Tetlow, R.J., 2015. Financial stress and economic dynamics: the transmission of crises. J. Monet. Econ. 70, 100–115.

Jeanne, O., Masson, P., 2000. Currency crises, sunspots, and Markov-switching regimes. J. Int. Econ. 50, 327–350.

Karamé, F., 2010. Impulse-response functions in Markov-switching structural vector autoregressions: a step further. Econ. Lett. 106, 162–165.

Kim, C.J., 1994. Dynamic linear models with Markov-switching. J. Econom. 60, 1–22.

Kim, C.J., Nelson, C.R., 1999a. State-Space Models with Regime Switching. MIT Press, Cambridge, MA.

Kim, C.J., Nelson, C.R., 1999b. Has the U.S. economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle. Rev. Econ. Stat. 81, 608–616.

Kirman, A., 1993. Ants, rationality, and recruitment. Q. J. Econ. 108, 137–156.

Klein, P., 2000. Using the generalized Schur form to solve a multivariate linear rational expectations model. J. Econ. Dyn. Control 24, 1405–1423.

Koop, G., Potter, S.N., 1999. Bayes factors and nonlinearity: evidence from economic time series. J. Econom. 88, 251–281.

Krolzig, H.M., 1997. Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis. Springer, Berlin.

Kurz, M., Motolese, M., 2001. Endogenous uncertainty and market volatility. Econ. Theory 17, 497–544.

Lind, N., 2014. Regime-switching perturbation for non-linear equilibrium models. Working paper, UCSD.

Liu, Z., Waggoner, D.F., Zha, T., 2011. Sources of macroeconomic fluctuations: a regime-switching DSGE approach. Quant. Econ. 2, 251–301.

Lo, M.C., Piger, J., 2005. Is the response of output to monetary policy asymmetric? Evidence from a regime-switching coefficients model. J. Money Credit Bank. 37, 865–886.

Lucas Jr., R.E., 1978. Asset prices in an exchange economy. Econometrica 66, 1429–1445.

McConnell, M.M., Perez-Quiros, G., 2000. Output fluctuations in the United States: what has changed since the early 1980's? Am. Econ. Rev. 90, 1464–1476.

Moore, B., Schaller, H., 2002. Persistent and transitory shocks, learning, and investment dynamics. J. Money Credit Bank. 34, 650–677.

Owyang, M., Ramey, G., 2004. Regime switching and monetary policy measurement. J. Monet. Econ. 51, 1577–1597.

Peria, M.S.M., 2002. A regime-switching approach to the study of speculative attacks: a focus on EMS crises. In: Hamilton, J.D., Raj, B. (Eds.), Advances in Markov-Switching Models. Physica-Verlag, Heidelberg.

Pesaran, M.H., Pettenuzzo, D., Timmermann, A., 2006. Forecasting time series subject to multiple structural breaks. Rev. Econ. Stud. 73, 1057–1084.

Pesaran, M.H., Timmermann, A., 2007. Selection of estimation window in the presence of breaks. J. Econom. 137, 134–161.

Ruge-Murcia, F.J., 1995. Credibility and changes in policy regime. J. Polit. Econ. 103, 176–208.

Ruge-Murcia, F.J., 1999. Government expenditure and the dynamics of high inflation. J. Dev. Econ. 58, 333–358.

Schmitt-Grohe, S., Uribe, M., 2004. Solving dynamic general equilibrium models using a second-order approximation. J. Econ. Dyn. Control 28, 755–775.

Schorfheide, F., 2005. Learning and monetary policy shifts. Rev. Econ. Dyn. 8, 392–419.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.

Sims, C., 2001. Solving linear rational expectations models. J. Comput. Econ. 20, 1–20.

Sims, C., Zha, T., 2006. Were there regime switches in U.S. monetary policy? Am. Econ. Rev. 96, 54–81.

Smith, A., Naik, P.A., Tsai, C.L., 2006. Markov-switching model selection using Kullback-Leibler divergence. J. Econom. 134, 553–577.

Startz, R., 1998. Growth states and shocks. J. Econ. Growth 3, 203–215.

Teräsvirta, T., 1994. Specification, estimation, and evaluation of smooth transition autoregressive models. J. Am. Stat. Assoc. 89, 208–218.

Teräsvirta, T., 2004. Smooth transition regression modelling. In: Lütkepohl, H., Krätzig, M. (Eds.), Applied Time Series Econometrics. Cambridge University Press, Cambridge, UK, pp. 222–242.

Timmermann, A., 2000. Moments of Markov switching models. J. Econom. 96, 75–111.

Tjøstheim, D., 1986. Some doubly stochastic time series models. J. Time Ser. Anal. 7, 51–72.

Veldkamp, L.L., 2005. Slow boom, sudden crash. J. Econ. Theory 124, 230–257.

White, H., 1984. Asymptotic Theory for Econometricians. Academic Press, Orlando.

Yang, M.X., 2000. Some properties of vector autoregressive processes with Markov-switching coefficients. Econ. Theory 16, 23–43.

CHAPTER 4

The Macroeconomics of Time Allocation

M. Aguiar[*], **E. Hurst**[†]
[*]Princeton University, Princeton, NJ, United States
[†]The University of Chicago Booth School of Business, Chicago, IL, United States

Contents

Abstract

In this chapter we explore the macroeconomics of time allocation. We begin with an overview of the trends in market hours in the United States, both in the aggregate and for key subsamples. After introducing a Beckerian theoretical framework, the chapter then discusses key empirical patterns of time allocation, both in the time series (including business cycle properties) and over the life cycle. We focus on several core nonmarket activities, including home production, child care, and leisure. The chapter concludes with a discussion of why these patterns are important to macroeconomics and spells out directions for future research.

Keywords

Time allocation, Home production, Labor supply, Employment, Leisure, Nonseparable preferences

JEL Classification Codes

J22, E24

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.06.004

1. INTRODUCTION

What drives the time series variation in labor supply? During the last decade, the employment to population ratio of prime-age workers has fallen sharply—particularly for lower skilled workers. As market work falls, how do households allocate their time? Why does labor supply vary so much at business cycle frequencies? Can the ability to produce at home make labor supply more elastic? Can innovations in home production technology explain the rise in female employment and the convergence of male and female labor supply elasticities? Why does consumption vary over the life cycle? As market work falls after middle age, how do household individuals allocate their time? As individuals age, do they allocate more time to home production and shopping reducing their observed market expenditure for a constant consumption basket?

In this chapter, we introduce readers to the importance of time allocation for life cycle, business cycle, and long-run time series movements in labor supply and market consumption. Becker's Presidential Address (1989) provides a nice argument in favor of linking micro time allocation and associated expenditure decisions to key macroeconomic outcomes. The goal of the chapter is to provide an introduction to the literature that examines these issues. In doing so, we highlight differences by both gender and years of accumulated schooling. As we show, the time series and life cycle patterns in time use differ markedly between men and women. Likewise, the time series and life cycle patterns also differ across skill groups. For example, the time women allocate to market work has risen sharply over the last five decades relative to men. Simultaneously, the time women allocate to home production has fallen sharply over the last five decades relative to men. However, the trends in leisure time are nearly identical between men and women. Yet, less-skilled men and women experienced a much larger increase in leisure than higher skilled men and women over the same period.

The chapter begins by exploring patterns in market work over time. We illustrate these patterns over time for different age, sex, and skill groups. These patterns set the stage for the work that follows. In Section 2, we outline a Beckerian model of consumption with multiple goods. The model illustrates the key forces illustrating how changes in the way time is allocated outside of the market sector can explain time series, life cycle, and business cycle movements in both the time allocated to market work and market consumption. This model while simple is quite powerful. Individuals are endowed with a given amount of time and, with said endowment, make choices on how it is allocated across activities given the prices and technologies they face.

In Sections 3–5, we document the time series, business cycle, and life cycle variation in individual time use, respectively. We primarily focus on three uses of time aside from market work. First, we look at home production broadly. These activities include activities like cooking, cleaning, shopping, doing laundry, moving the lawn, and caring for older adults. Second, we look at child care. In doing so, we discuss why the literature treats child care as a distinct activity relative to home production. Lastly, we look at the time individuals spend in leisure activities. This category includes time spent watching television, socializing, going to the movies, playing video games, exercising, and sleeping. On occasion, we discuss the trends in the remaining time-use categories like job search, accumulating human capital, and participating in civic organizations. Throughout all of these sections, we also set these facts in the broader macroeconomics literature. In the final section, we close with a few comments on a future research agenda.

2. TRENDS IN MARKET WORK

In this section we set the stage by reviewing and updating some familiar trends in market labor. In the remainder of the chapter, we discuss how trends in market hours are complemented by trends in other time-intensive activities. The next section provides a theoretical framework which highlights why measuring time allocation across multiple activities may be useful in understanding market hours.

Fig. 1 shows the trends in male hours worked per week allocated to market work (left axis) and employment propensity (right axis) from 1967 through 2014. To compute this

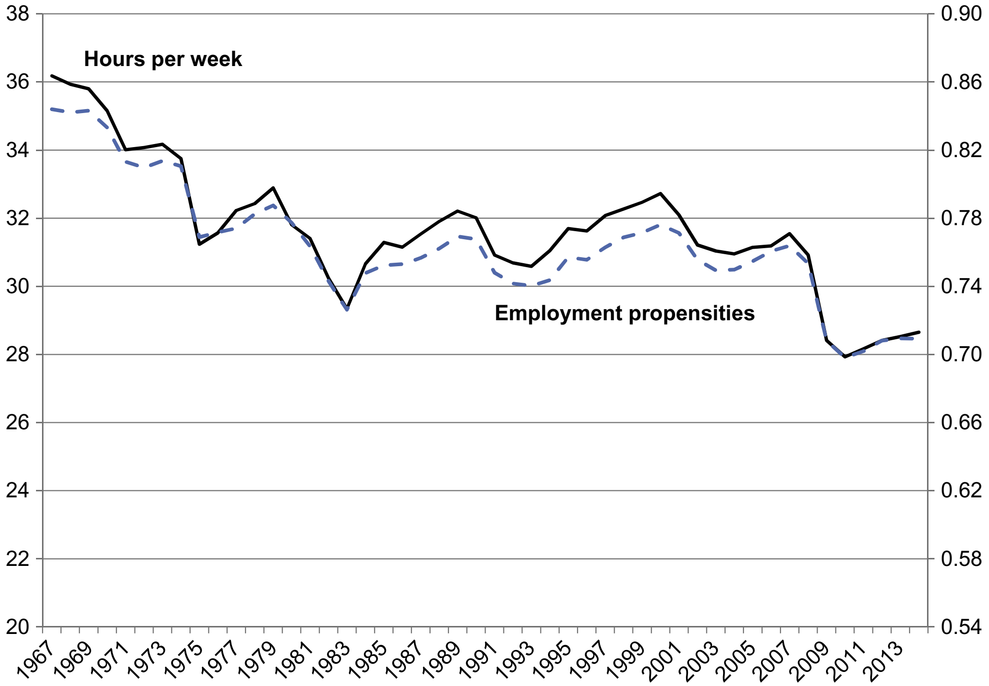


Fig. 1 CPS trends in market hours and employment rates: all men (21–75). Note: Figure shows the trends in market hours per week worked (*solid line—left axis*) and employment propensities (*dashed line—right axis*) between 1967 and 2014. Data come from the March Current Population Survey. The sample includes all men between the ages of 21 and 75 (inclusive) within the survey. Hours worked per week in the market are based on the self-reported response to a question of how many hours the individual worked last week. Employment propensities are based on the amount of people who report being employed in a given week.

figure (and all figures within this section), we use data from the March Current Population Survey (CPS).[a] The only restriction we placed on the data was to restrict the sample to include men between the ages of 21 and 75 (inclusive). Hours per week is measured as the individual's self-reported hours worked on all jobs during the prior week. For those that did not work last week, hours per week is measured as zero. The employment propensity is a dummy variable that takes the value of 1 if the individual reported having a job (regardless of whether or not they worked any hours last week).

As seen from Fig. 1, male hours per week have fallen sharply since the late 1960s. In 1967, the typical male between the ages of 21 and 75 worked roughly 36 h per week. That number fell steadily to the 1980s where, on average, men worked about 31 h per week. During the 2008 recession, male hours fell to only about 28 h per week. That number has not rebounded as of 2014. The movement in hours per week is almost entirely driven by movements on the extensive margin of labor supply. As seen from Fig. 1, employment propensities moved in lock step with the hours movement over this time period. Put another way, hours per week worked conditional on being employed remained roughly constant over this 47-year period. Prior to the 2008 recession, roughly 77% of men in the 21–75 age range were employed. That number fell to 70% during the recession and it has only rebounded to 71% by 2014.

Fig. 2 shows hours per week, conditional on working, for men during the 1967–2014 period. Hours worked per week, conditional on working, have remained roughly constant over the last 50 years. Since 1970, hours worked per week, conditional on working, have bounced around between 40 and 42 h per week. Since the early 2000s, there has been a persistent decline in hours worked per week, conditional on working, from 42 h per week to roughly 40 h per week in 2009. The low hours per week, conditional on working, has remained roughly constant since 2009.

Fig. 3 shows the similar patterns for women. Between the late 1960s and the late 1990s, female time allocated to market work increased sharply. Both hours per week and employment propensities increased continuously during this period. Starting in 2000, however, female hours worked per week and employment propensities fell. The trends in female hours and employment propensities matched their male counterparts. Fig. 4 shows hours per week, conditional on working, for women during the 1967–2014 period. Like men, hours worked per week, conditional on working, have remained roughly constant over the last 50 years. Since 1980, hours worked per week, conditional on working, have remained roughly constant at about 35 h per week. This shows that for women essentially all the change in total hours since 1980 is due to changes in the extensive margin of employment.

Figs. 5 and 6 show the same patterns by educational attainment for men (Fig. 5) and women (Fig. 6). We define higher educated as individuals who completed a bachelor's

[a] We downloaded the data directly from the Integrated Public Use Microdat Series (IPUMS) website: https://www.ipums.org.

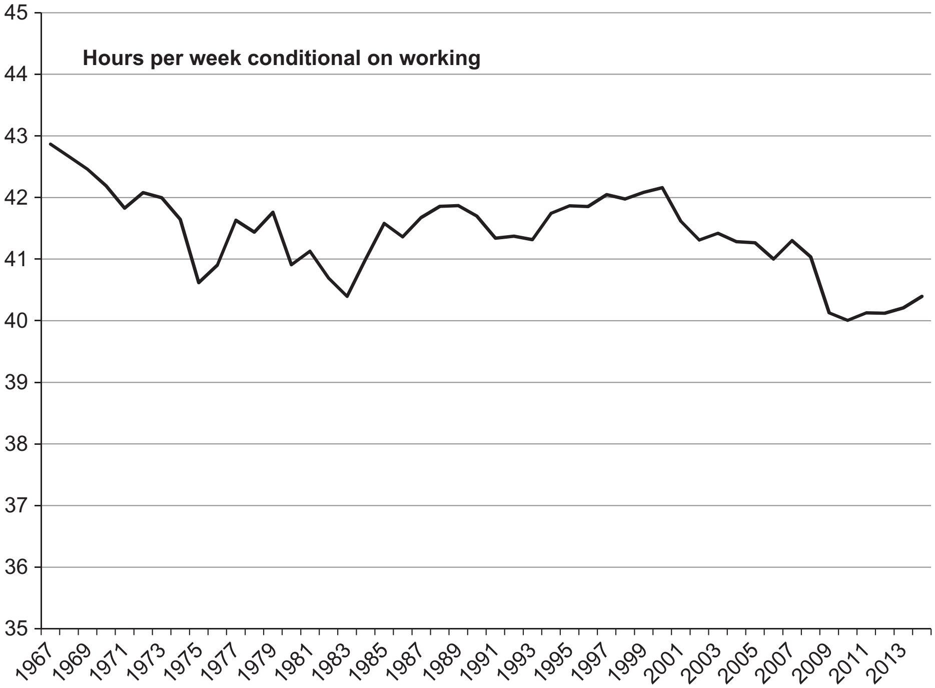


Fig. 2 CPS trends in market hours and employment rates: employed men. Note: Figure shows the trends in market hours per week worked for men, conditional on working. The sample is the same as Fig. 1.

degree or higher. Lower educated individuals include anyone with less than a bachelor's degree. Given that the population has been aging during this time period, Figs. 7 and 8 show the trends in hours work by sex, skill, and age. Fig. 7A shows the patterns for four age groups for higher skilled men. The age groups are 21–40, 41–55, 56–65, and 66–75. Figs. 7B and 8A and B show the analogous age breakdown for lower skilled men, higher skilled women, and lower skilled women, respectively.

The patterns in Figs. 5–8 highlight many of the questions that frame our subsequent analysis. First, hours allocated to market work is falling for men of both skill levels since the late 1960s. Higher educated men experienced a decline in market work hours from about 43 h a week in 1967 to about 34 h a week in 2008. Much of this decline was concentrated prior to 1980 and after 1999. As the population aged during this time, a greater fraction of individuals became retired. In Fig. 7A, we see that hours worked declined for every age group of higher skilled men during the last 47 years. Higher skilled men aged 56–65 saw the largest decline. In 1967, these men worked on average 40 h a week. That number fell to about 30 h a week in 1990 has been relatively constant throughout—even during the 2008 recession. High-skilled men aged 41–55 experienced a steady decline in

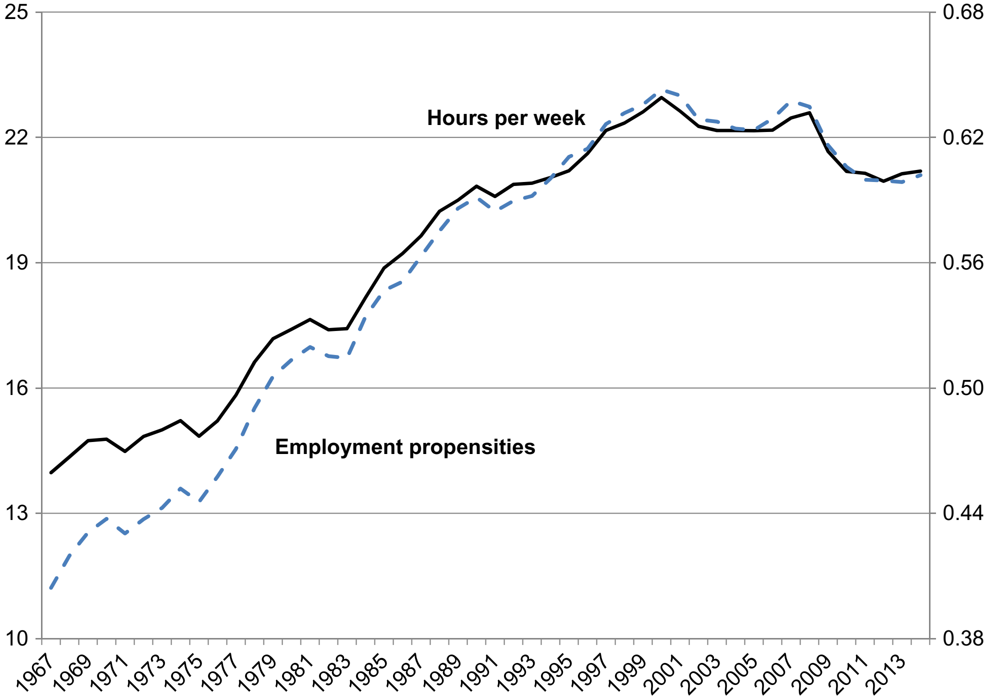


Fig. 3 CPS trends in market hours and employment rates: all women (21–75). Note: Figure shows the trends in market hours per week worked (*solid line—left axis*) and employment propensities (*dashed line—right axis*) between 1967 and 2014. Data come from the March Current Population Survey. The sample includes all women between the ages of 21 and 75 (inclusive) within the survey. Hours worked per week in the market are based on the self-reported response to a question of how many hours the individual worked last week. Employment propensities are based on the amount of people who report being employed in a given week.

hours worked since the late 1960s from 45 h per week in 1967 to 40 h a week in 2014. Like the trend for all high-skilled men regardless of age, much of the decline took place prior to 1980 and after 1999. Younger high-skilled men (those aged 21–40) had relative flat hours through 1999. But, since the late 1990s, younger higher skilled men have reduced their hours from 41 h per week to about 37 h per week in 2014. Conversely, higher skilled men aged 66–75 have increased their hours worked by about 3–4 h.

The qualitative decline in market hours is roughly similar for lower skilled men within each age group. The main quantitative difference, however, is that the declines were much more dramatic for low-skilled men between the ages of 21 and 40 and between the ages of 41 and 55. For this group of relatively young men, there was a marked decline in hours worked relative to their higher educated counterparts. In 1967, younger lower skilled men worked roughly 40 h per week. Yet, by 2014, lower educated men between the ages of 21–40 are only working just over 28 h per week. This 12 h per week decline

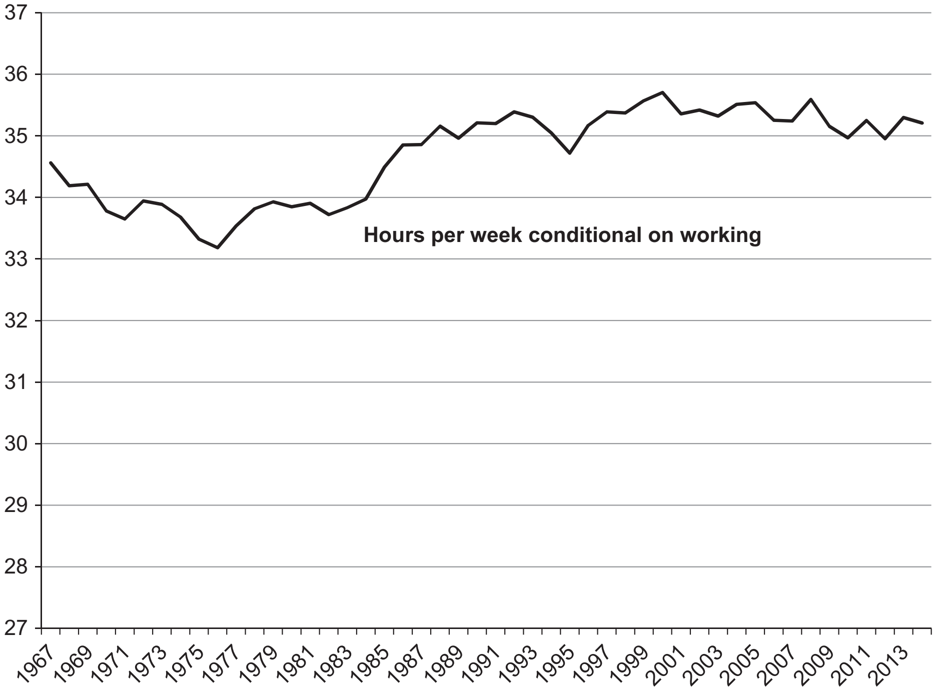


Fig. 4 CPS trends in market hours: employed women. Note: Figure shows the trends in market hours per week worked for women, conditional on working. The sample is the same as Fig. 3.

dwarfs 5-h decline for higher educated men in the same age range. Lower skilled men aged 41–55 decreased their market work hours by 8 h per week on average. This is larger than the 5-h decline experienced by the higher skilled men of the same age. Much of this divergence occurred starting after 1999. Young lower skilled men have dramatically reduced their hours during the last 15 years. As with the patterns in Fig. 1, essentially all of the action is on the extensive margin of employment. There was relatively little movement in hours worked per week conditional on being employed. The increase in inequality in employment propensities between higher and lower prime-aged men is a defining feature of time use since 2000.

Like with men, higher skilled women consistently work more in the market sector than lower skilled women. Like their male counterparts, higher skilled prime-aged women (those 21–40 and those 41–55) reduced their market work hours slightly during the 2000s. This comes as a reversal of trends during the prior decades. From 1967 through 1990, prime-aged higher skilled women increased their market hours by roughly 6–9 h per week. Again, like their male counterparts, prime-aged lower skilled women saw a dramatic reduction in market work hours during the 2000s. For example, younger low-skilled women (those aged 21–40) reduced their market work hours by roughly

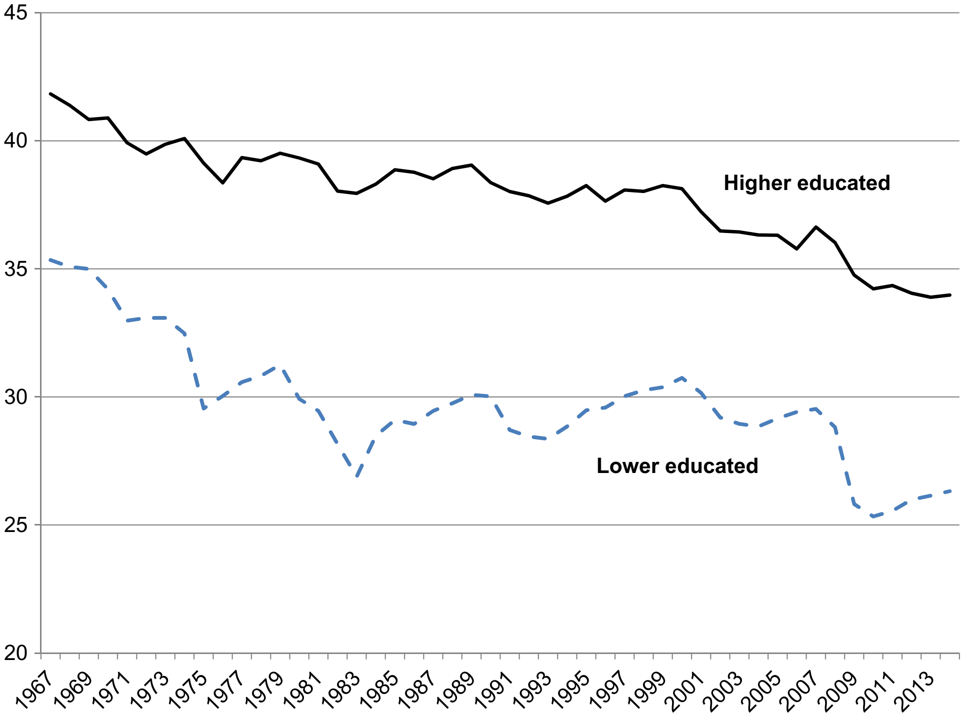


Fig. 5 CPS trends in market hours: men by skill (21–75). Note: Figure shows the trends in market hours per week worked for higher skilled men (*solid line*) and lower skilled men (*dashed line*) between 1967 and 2014. Data come from the March Current Population Survey. The sample includes all men between the ages of 21 and 75 (inclusive) within the survey. Higher educated men are defined as those men with a bachelor's degree or higher. Lower educated men have years of schooling less than 16 years. Hours worked per week in the market are based on the self-reported response to a question of how many hours the individual worked last week.

4 h per week between 1999 and 2014. The combination of these patterns caused inequality market work hours to also increase during the 2000s for lower skilled prime-aged women relative to higher skilled prime-aged women.

Given these large fluctuations in market work hours over time, across genders, across skill groups within gender and across age groups within a gender * skill group, it is interesting to understand how time allocated to activities other than market work have been changing as well. We turn to that analysis now.

3. A THEORY OF TIME USE

The modern theory of time allocation was first laid out in the seminal Becker (1965). The Beckerian approach recognizes the consumption "commodities" produced using both

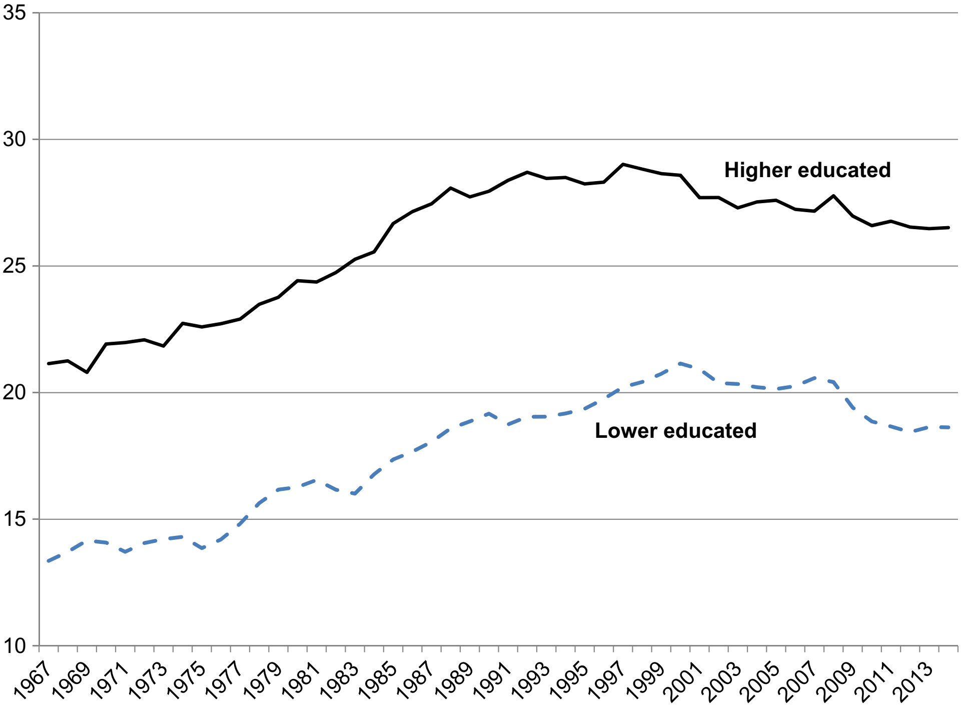


Fig. 6 CPS trends in market hours: women by skill (21–75). Note: Figure shows the trends in market hours per week worked for higher skilled women (*solid line*) and lower skilled women (*dashed line*) between 1967 and 2014. Data come from the March Current Population Survey. The sample includes all women between the ages of 21 and 75 (inclusive) within the survey. Higher educated women are defined as those women with a bachelor's degree or higher. Lower educated women have years of schooling less than 16 years. Hours worked per week in the market are based on the self-reported response to a question of how many hours the individual worked last week.

market goods and one's time. In this section, we highlight a few implications of the Beckerian model that have proved useful in understanding empirical time allocation and associated market expenditures. The version of Becker's model presented below draws on Aguiar and Hurst (2007b) and Aguiar et al. (2012). For expositional reasons, we make a number of simplifying assumptions which can easily be relaxed in order to highlight the key mechanisms.

Consider an agent which enjoys utility over I different consumption commodities, $c_1, \ldots, c_i, \ldots, c_I$. Commodity i is produced using market input x_i and time input h_i according to the technology:

$$c_i = f^i(x_i, h_i).$$

We assume that there is no joint production, so x_i and h_i are used only to produce commodity i.

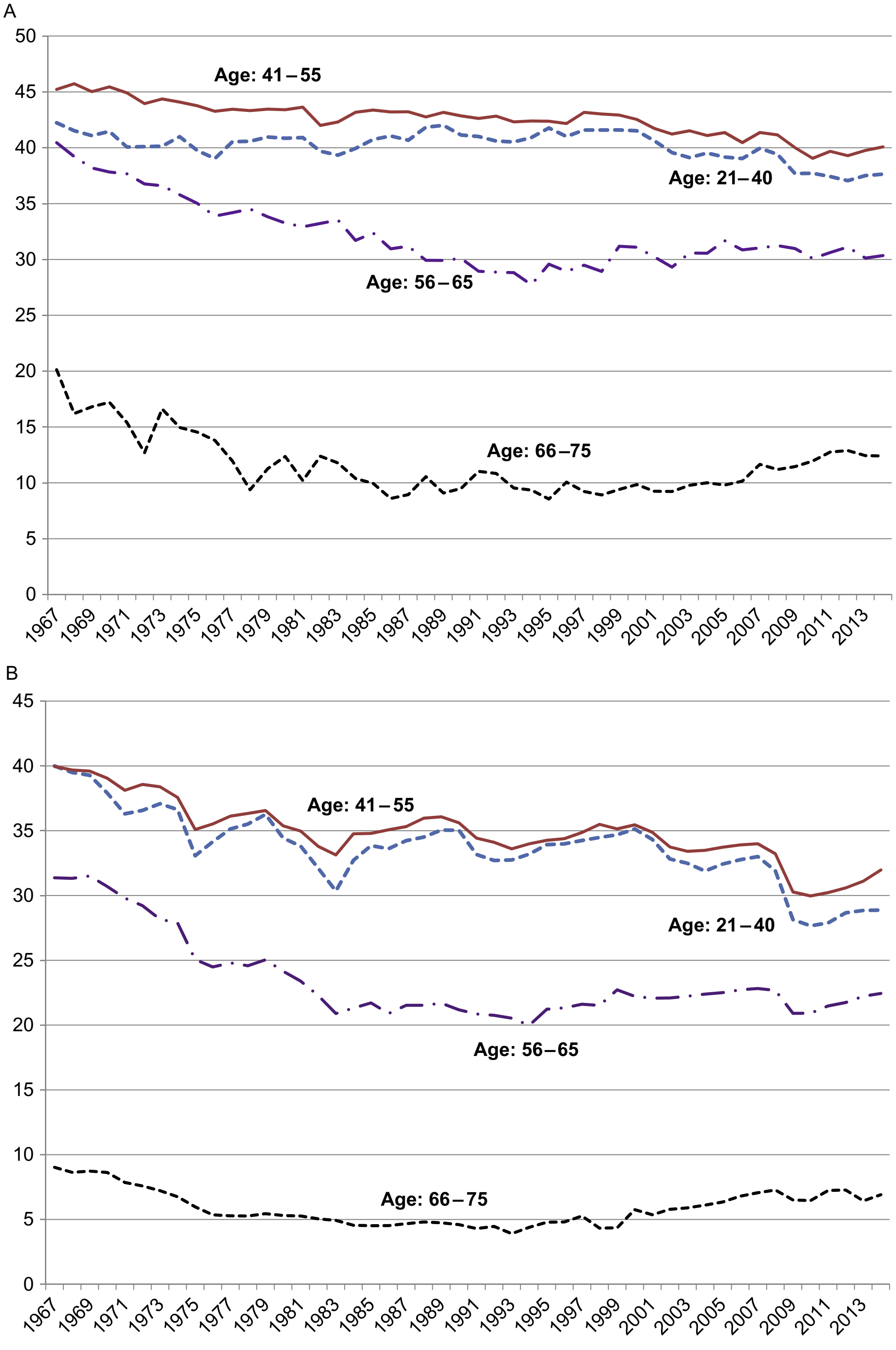


Fig. 7 See legend on opposite page.

To motivate the framework, a commodity could be a meal, which is produced using ingredients (a market good) as well as cooking time. In this example, time and goods are substitutes, as one could purchase the meal partially or completely prepared at a higher goods price but a lower time cost. Another example, in which time and goods are complements, is watching TV. For this commodity, the ability to substitute market expenditures for time inputs is limited; however, the purchase of additional inputs (like a premium channel) raises the value of time spent in the production of the commodity.

The agent lives for T periods and has preferences over sequences of consumption given by:

$$\sum_{t=0}^{T-1} \beta^t u(c_1(t), \ldots, c_I(t)).$$

There is no uncertainty and utility is separable across periods.

We assume that the agent can borrow and lend freely at a an interest rate $R = \beta^{-1}$ and in period t chooses to supply labor $n(t)$ at a market wage $w(t)$. Starting from some initial assets a_0, the budget set is therefore:

$$\sum_{t=0}^{T-1} \beta^t \left(\sum_{i=1}^{I} p_i(t) x_i(t) - w(t) n(t) \right) \leq a_0.$$

We normalize the time endowment to one each period. The time allocation budget constraint is:

$$\sum_i h_i + n \leq 1, h_i, n \geq 0.$$

We shall assume that labor is interior, and so the wage is the opportunity cost of time inputs into home production. We also assume that $h_i \geq 0$ is never binding as well.

If we assume that f^i has constant returns to scale, then the implied price index for a unit of consumption commodity c_i can be expressed by $q^i(p_i, w)$, where q^i solves:

$$q^i(p_i, w) = \min_{x_i, h_i} p_i x_i + w h_i$$

subject to

$$f^i(x_i, h_i) \geq 1.$$

Fig. 7 CPS trends in market hours: men by education and age. Note: Figure shows the trends in market hours per week worked for more educated (A) and less educated (B) men by age between 1967 and 2014. Data come from the March Current Population Survey. The sample includes all men between the ages of 21 and 75 (inclusive) within the survey. More educated men are defined as those men with a bachelor's degree or higher. Less educated men have years of schooling less than 16 years. Hours worked per week in the market are based on the self-reported response to a question of how many hours the individual worked last week.

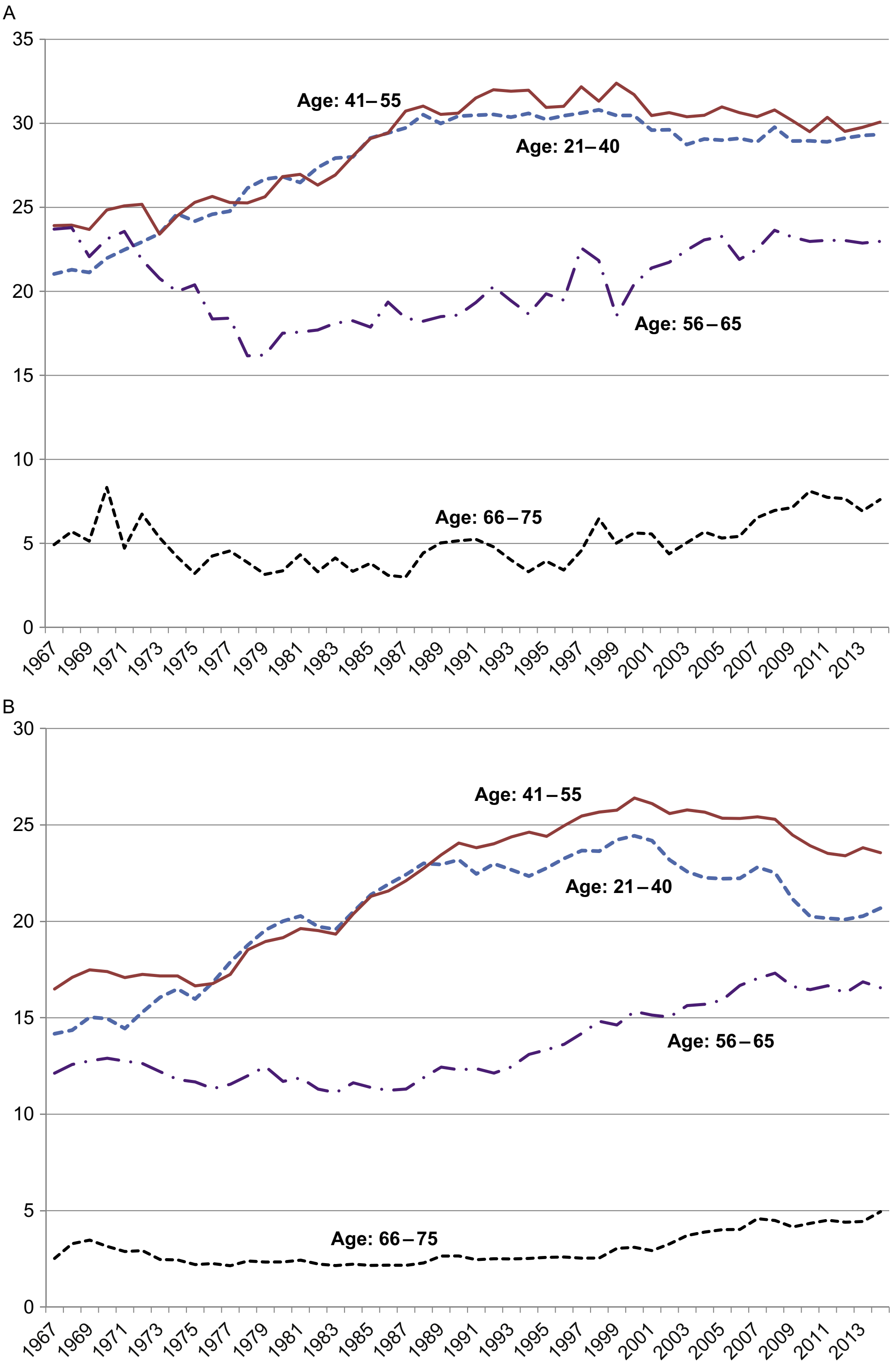


Fig. 8 See legend on opposite page.

Home production implies that the price of a consumption commodity depends on the price of the market input as well as the opportunity cost of time.

It is straightforward from the cost-minimization problem that:

$$\frac{f_h^i}{f_x^i}=\frac{w}{p_i},$$

that is, the marginal rate of technical substitution is set equal to the relative price of inputs. Denote the elasticity of substitution between x_i and h_i associated with the technology f^i by σ_i. As the relative cost of time increases, the agent will reduce the ratio of time to market inputs $\left(\frac{h_i}{x_i}\right)$ in production, the extent of this substitution being governed by σ_i. Again, for notational simplicity, we take σ_i to be constant.

The agent's problem can be rewritten as:

$$\max_{\{c_i(t)\}}\sum_{t=0}^{T-1}\beta^t u(c_1(t),\ldots,c_I(t))$$

subject to

$$\sum_{t=0}^{T-1}\beta^t\left(\sum_i q_i(p_i(t),w(t))x_i(t)-w(t)\right)\le a_0.$$

Letting λ be the multiplier on the budget constraint, the first-order condition is:

$$u_i\ =q^i\lambda.$$

An interesting question is how does time and market inputs vary with the wage holding constant λ. A little algebra leads us to:

$$\left.\frac{d\ \ln x_i}{d\ \ln w}\right|_\lambda\ =s_h^i\left(\sigma_i-\frac{1}{\gamma_i}\right), \tag{1}$$

where:

$$s_h^i=\frac{\partial\ \ln q^i}{\partial\ \ln w}=\frac{wh}{q_i c_i}$$

is the cost share of time input into commodity i and

Fig. 8 CPS trends in market hours: women by education and age. Note: Figure shows the trends in market hours per week worked for more educated (A) and less educated (B) women by age between 1967 and 2014. Data come from the March Current Population Survey. The sample includes all women between the ages of 21 and 75 (inclusive) within the survey. More educated women are defined as those women with a bachelor's degree or higher. Less educated women have years of schooling less than 16 years. Hours worked per week in the market are based on the self-reported response to a question of how many hours the individual worked last week.

$$\frac{1}{\gamma_i} = -\frac{u_i}{u_{ii}c_i}$$

is the intertemporal elasticity of substitution for commodity i.

Eq. (1) states that if the intratemporal elasticity of substitution is greater than the intertemporal elasticity of substitution, an increase in the cost of time (holding λ constant) will lead to an increase in market expenditure, and vice versa if $\sigma_i < \frac{1}{\gamma_i}$. The intuition is the following. An increase in the price of time induces substitution away from h_i and toward x_i for a given level of production. This substitution is governed by σ_i. However, an increase in the price of time raises the cost of consuming today relative to other periods, as $q^i(p_i, w)$ is increasing in both arguments. This induces a shift in consumption away from the high-wage period, and both expenditure and time inputs correspondingly decline. The size of this effect is governed by the intertemporal elasticity of substitution, $1/\gamma_i$. Whether expenditure goes up or down in response to variation in w depends on which effect dominates. Moreover, the effect is scaled by the share of time input into production of the commodity, s_h^i.

Similarly, the agent's first-order conditions imply:

$$\left.\frac{d\ \ln h_i}{d\ \ln w}\right|_{\lambda} = -\sigma_i(1 - s_h^i) - \frac{1}{\gamma_i} s_h^i. \tag{2}$$

This elasticity is unambiguously negative, as both intra- and intertemporal considerations imply reducing time inputs when the wage is high. The total effect is a weighted average of the two elasticities.

Using the time constraint, which implies $\sum_i h_i = 1 - n$, we can express the Frisch elasticity of nonmarket time $1 - n$ as:

$$\left.\frac{d\ \ln n}{d\ \ln w}\right|_{\lambda} = \sum_{i=1}^{I}\left(\frac{h_i}{n}\right)\left(\sigma_i(1 - s_h^i) + \frac{1}{\gamma_i} s_h^i\right), \tag{3}$$

which is a weighted average of the elasticity of each commodity's time input from Eq. (2). Eq. (3) implies that the elasticity of market labor depends on how time is allocated away from the market, and how elastic those activities are with respect to the wage. This insight goes back at least to Mincer (1962), who argued that women have a higher elasticity of market labor as their nonmarket time was concentrated in activities with close market substitutes, which would be high σ_i in our framework. As we shall see, women have been substituting nonmarket time away from home production and toward leisure in recent decades. In the Beckerian framework, this implies a corresponding evolution in the elasticity of labor supply. An interesting question for future research is whether this is reflected in the data.

4. TIME-USE DATA

Before proceeding, it is worth discussing how we measure time away from market work. For our primary data source, we use data from the 2003 to 2013 waves of the American Time-Use Survey (ATUS). The ATUS is conducted by the US Bureau of Labor Statistics (BLS) and individuals in the sample are drawn from the exiting sample of the CPS. On average, individuals are sampled approximately 3 months after completion of their final CPS survey. Given this, we can link each respondent to their labor market conditions when they were in the CPS. The ATUS is a highly detailed and easy-to-use survey, and the link to the CPS makes it straightforward to link time diaries to a long list of covariates.

At the time of the ATUS survey, the BLS updates the respondent's employment and demographic information. Each wave is based on 24-h time diaries where respondents report the activities from the previous day in detailed time intervals. Survey personnel then assign the activities reported by the individual to a specific category in the ATUS's set classification scheme which is comprised of over 400 detailed time-use categories. For more information on the types of activities that are recorded in the ATUS, see Hammermesh et al. (2005). The 2003 wave of the survey includes over 20,000 respondents, while each of the remaining waves includes roughly 13,000 respondents.

We segment the allocation of time into six broad time-use categories. We construct the categories to be mutually exclusive and to sum to the individual's entire time endowment. The six categories we look at are described in detail below and are based on the response for the primary time-use activity. These categories are defined similar to Aguiar et al. (2013).

Market work includes all time spent working in the market sector on main jobs, second jobs, and overtime, including any time spent commuting to or from work and time spent on work-related meals and activities. We separate from total market work the time spent on job search and the time spent on other income-generating activities outside the formal sector. This allows us to study the extent to which households spend time looking for employment or substitute time from the formal to the informal sector.

Job search includes all time spent by the individual searching for a job. As with all time-use categories, we include the time spent commuting associated with job search as part of time spent on job search. Job search includes, among others, activities such as sending out resumes, going on job interviews, researching details about a job, asking about job openings, or looking for jobs in the paper or the Internet.

Child care measures all time spent by the individual caring for, educating, or playing with their children. Guryan et al. (2008) show that the time series and life cycle patterns of time spent on child care differ markedly from the patterns of time spent on home production. In particular, the income elasticity of time spent on child care is large and positive, while the income elasticity of time spent on home production is large and negative. Additionally, some components of child care have a direct leisure component. For example, according to Juster (1985), individuals report spending time playing with their children as among their most enjoyable activities. On the other hand, there is a well-developed market for child care services that parents are willing to pay for to reduce their time spent with their children. Given these dichotomies, we treat child care as a separate category.

Nonmarket work (home production) consists of four subcategories: core home production, activities related to home ownership, obtaining goods and services, and care of other adults. Core home production includes any time spent on meal preparation and cleanup, doing laundry, ironing, dusting, vacuuming, indoor household cleaning, cleaning or repairing vehicles and furniture, and activities related to the management and the organization of the household. Home ownership activities include time spent on household repairs, time spent on exterior cleaning and improvements, time spent on the garden, and lawn care.[b] Time spent obtaining goods and services includes all time spent acquiring any goods or services (excluding medical care, education, and restaurant meals). Examples include grocery shopping, shopping for other household items, comparison shopping, coupon clipping, going to the bank, going to a barber, going to the post office, obtaining government services, and buying goods online. Finally, care of other adults includes any time supervising and caring for other adults, preparing meals and shopping for other adults, helping other adults around the house with cleaning and maintenance, and transporting other adults to doctors offices and grocery stores.

Leisure includes most of the remaining time individuals spend that is not on market work, nonmarket work, job search, or child care. Specifically, we follow Aguiar and Hurst (2007c, 2009) and try to isolate goods for which time and expenditure are complements. The time spent on activities which comprise leisure includes time spent watching television, time spent socializing (relaxing with friends and family, playing games with friends and family, talking on the telephone, attending and hosting social events, etc.), time spent exercising and on sports (playing sports, attending sporting events, exercising, running, etc.), time spent reading (reading books and magazines, reading personal mail and email, etc.), time spent on entertainment and hobbies that do not generate income (going to the movies or theater, listening to music, using the computer for leisure, doing arts and crafts, playing a musical instrument, etc.), time spent with pets, and all other similar activities. We also include in our leisure measure activities that provide direct utility but may also be viewed as intermediate inputs such as time spent sleeping, eating, and

[b] With respect to the long-run trends in time use, there is a debate about whether time spent gardening or spending time with one's pets should be considered as home production or leisure. See, for example, Ramey (2007). Given that the ATUS time-use categories can be disaggregated into finer subcategories, in this paper we include gardening and lawn care in nonmarket work and we include pet care into leisure.

personal care. While we exclude own medical care, we include activities such as grooming, having sex, and eating at home or in restaurants.

Other includes all the remaining time spent on one's education, time spent on civic and religious activities, and time spent on one's own medical and health care. Some of this time can be considered home production as well, as they represent time investments into the stock of health and human capital.[c]

For our main sample, we include all ATUS respondents between the ages of 21 and 75 (inclusive) who had complete time-use record. Specifically, we exclude any respondent who had any time allocation that was not able to be classified by the ATUS staff. In total, we have 107,768 individuals in our base sample. We use the sample weights provided by the ATUS to aggregate responses by age or by year. Throughout our analysis, we also look at subsamples by age, gender, and accumulated schooling.

We also bring in results from Aguiar and Hurst (2007c, 2009) when exploring historical trends in time use. For these historical trends, data are used from the *1965–1966 America's Use of Time* and the *1985 Americans' Use of Time*. The 1965–1966 Americans' Use of Time was conducted by the Survey Research Center at the University of Michigan. The survey sampled one individual per household in 2001 households in which at least one adult person between the ages of 19 and 65 was employed in a nonfarm occupation during the previous year. This survey does not contain sampling weights, so we weight each respondent equally (before adjusting for the day of week of each diary). Of the 2001 individuals, 776 came from Jackson, Michigan. The time-use data were obtained by having respondents keep a complete diary of their activities for a single 24-h period between November 15 and December 15, 1965, or between March 7 and April 29, 1966. When recounting historical trends in Aguiar and Hurst (2007c, 2009), the Jackson, Michigan sample was included. The 1985 Americans' Use of Time survey was conducted by the Survey Research Center at the University of Maryland. The sample of 4939 individuals was nationally representative with respect to adults over the age of 18 living in homes with at least one telephone. The survey sampled its respondents from January 1985 through December 1985. Again, weights were used to ensure that each day of the week was represented equally. The classification scheme for the time-use data used in Aguiar and Hurst (2007c, 2009) was nearly identical to the classification outlined above.[d]

[c] The "other" category also includes any time spent engaging in activities that generate income outside the formal market sector. These include time spent preparing hobbies, crafts, or food for sale through informal channels. Additionally, activities like informal babysitting are included in this category. As shown in Aguiar et al. (2013), this subcategory of time spent on income-generating activities outside the formal market sector is close to zero on average, suggesting that it is not worth analyzing as a separate category.

[d] While nearly identical, there were some differences. In particular, Aguiar and Hurst (2007c, 2009) included lawn care and gardening as a component of "leisure." In the classification using the 2003–3013 ATUS discussed above, lawn and gardening was included as a component of home production.

5. LONG-RUN TRENDS IN TIME USE

5.1 Historical Trends in Time Use

As show above, time spent on market work for men has been falling within the United States since the late 1960s, while time spent on market work for women has been increasing steadily during this time period. Using the detailed time diaries, we can measure the trends in three other time-use categories: nonmarket work, child care, and leisure. For much of the historical trends we document in this section, we draw on the work of Aguiar and Hurst (2007c, 2009). In those papers, Aguiar and Hurst restrict their attention to individuals between the ages of 18 and 65 who are nonretired. The nonretired restriction is necessitated by the restrictions to the 1965 survey which only sampled people who were nonretired. Likewise, the restriction excluding individuals over the age of 65 was necessitated by the 1965 survey not interviewing individuals above the age of 65. While these restrictions are slightly narrower than the restrictions, we impose on the ATUS data in subsequent sections, the restrictions do not alter the main take aways for the time series trends in any meaningful way.

Fig. 9 shows the time series patterns in nonmarket work, child care, and leisure for the full sample, men and women in 1965, 1985, and 2003 as documented by Aguiar and Hurst (2007c). Fig. 9A shows the trends in nonmarket work. Between 1965 and 2003, women dramatically decreased the time they allocated to home production by roughly 10 h per week. Men, conversely, increased their home production between 1965 and 1985 by roughly 3 h per week. Between 1985 and 2003, male home

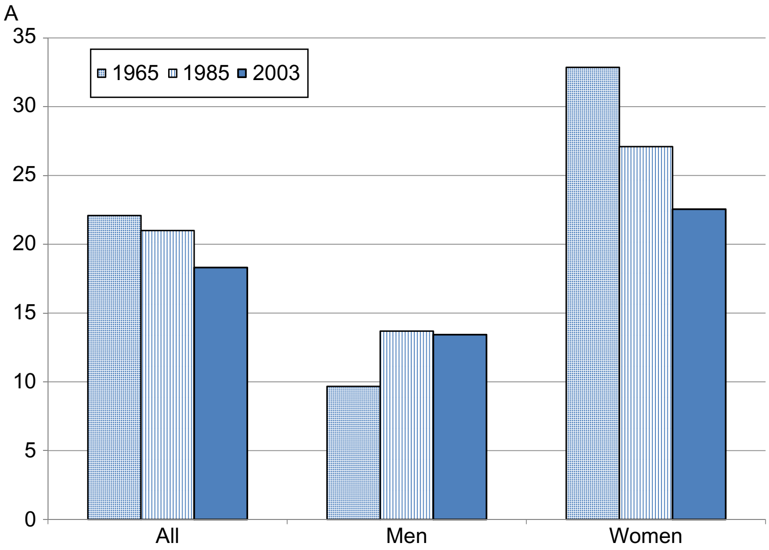


Fig. 9 Trends in time allocation: all men and women. Note: Figure shows the amount of time allocated to nonmarket work (A), child care (B), and leisure (C), in 1965, 1985, and 2003. Results in the figure come from tables II and III of Aguiar and Hurst (2007c). See text for additional details.

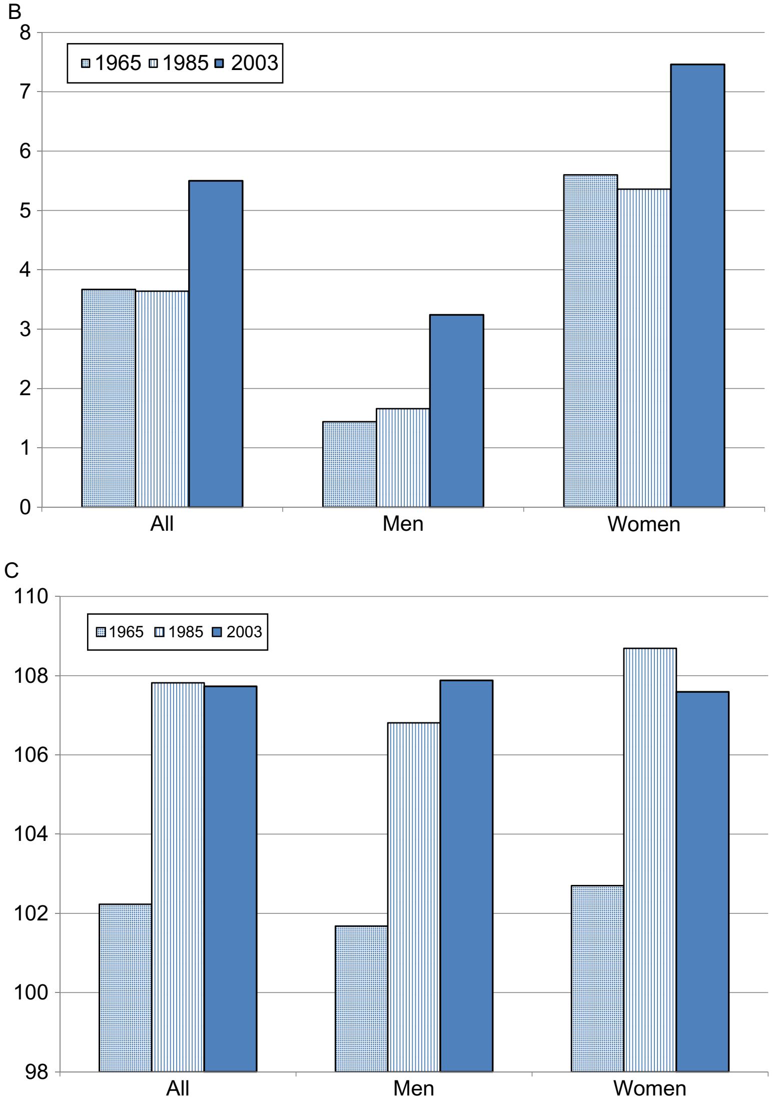


Fig. 9—Cont'd

production hours have been roughly constant. Not only has nonmarket work become less prevalent within the United States during the last 40 years, but also men and women are converging in their nonmarket work levels. Existing work has emphasized that innovations in the nonmarket sector caused women's increase in market work. For example, Greenwood et al. (2005) have shown that innovations in labor-saving devices used in home production allowed women to increase their labor supply in a model where home production is an active margin of substitution.

In Fig. 9B, we see time spent on child care has increased in recent years as well for both men and women. All of the increase took place after 1985. It is hard to tell how much of that increase is real or an artifact of the different survey designs between the 2003 ATUS and the earlier surveys. In particular, the ATUS had as a goal to measure parental time inputs into children. Ramey and Ramey (2010) document that the increase in time spent with children has increased more for high educated parents relative to low educated parents. The increasing gap in time spent with children by education has occurred in all categories of child care time: time spent on basic child care, time spent on educational child care, and time spent on recreational child care. They suggest that the increase in time spent on child care is real and a result of increased competition to get children into elite universities.

In Fig. 9C, the time series trends in leisure are shown. The large declines in market work for men during the 1960s, 1970s, and 1980s led to a large increase in leisure time for males between 1965 and 1985. Likewise, the large declines in home production for women during the 1960s, 1970s, and 1980s led to a large increase in leisure time for females between 1965 and 1975. For both men and women, leisure was roughly constant between 1985 and 2003. Men's leisure increase by roughly 1 h and women's leisure declined by roughly 1 h over the two decades between 1895 and 2003. It is interesting to note, however, that despite very different levels of market work, home production, and child care, men and women's leisure time is nearly identical in each decade. For example, in 2003, both men and women allocated roughly 107 h per week to leisure time activities. The 107 h includes time spent sleeping. Removing sleep from the leisure activities does not change any of the cross-sectional or time series patterns given that sleeping time is roughly constant over the decades and roughly constant between men and women.

Figs. 10 and 11 show the trends in home production and leisure by sex-skill groupings. The take aways from these figures are twofold. First, the trends in home production are nearly identical across educational attainment, conditional on sex. Second, the trends in leisure have diverged sharply between higher skilled and lower skilled individuals. Higher skilled individuals only experienced modest increases in leisure between 1965 and 2003. After experiencing large increases between 1965 and 1985, the leisure gains reversed between 1985 and 2003. Conversely, lower skilled individuals tracked their higher educated counterparts in terms of increased leisure time between 1965 and 1985 but continued to increase their leisure time between 1985 and 2003. The increase in leisure inequality has matched the well-documented increase in income and consumption inequality during the last 30 years documented by many in the literature.[e]

The above facts are drawn from the work of Aguiar and Hurst (2007c, 2009). However, Aguiar and Hurst (2007c, 2009) were not the only papers to harmonize historical

[e] See, for example, Aguiar and Bils (2015).

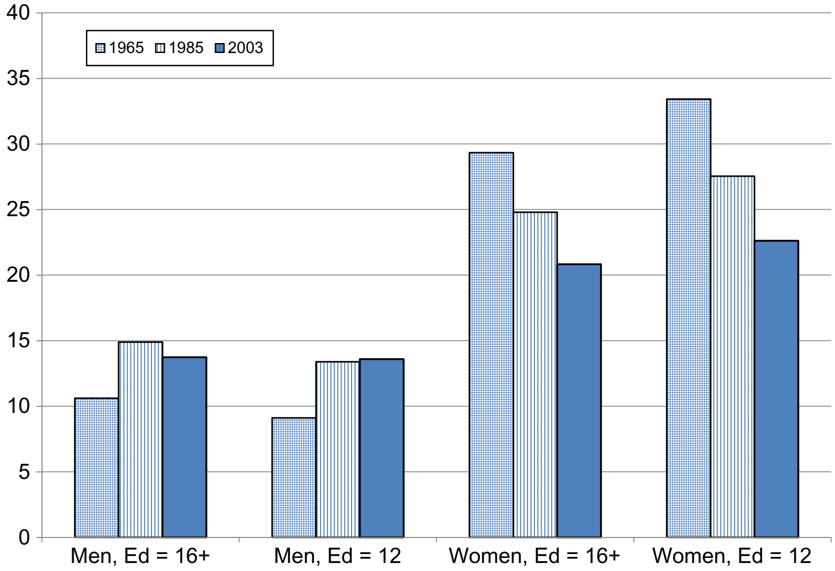


Fig. 10 Trends in nonmarket work hours: all, men, and women, by skill. Note: Figure shows the amount of time allocated to home production activities in 1965, 1985, and 2003 by sex and skill. The figure focuses on those with schooling levels of a bachelor's degree or more (Ed = 16+) and schooling levels of exactly a high school degree (ED = 12). Results in the figure come from tables V of Aguiar and Hurst (2007c). See text for additional details. Unlike the results in Fig. 7A–C, the results in this figure also adjust for the changing demographic composition over time within each sex-skill group. The demographic adjustment accounts for changing age distribution and family composition. The demographic adjustments made little difference to the broad time trends.

US time-use surveys to examine trends in nonmarket work and leisure over time. In classic books, Juster and Stafford (1985) and Robinson and Godbey (1999) harmonized the subset of the time-use data sets used by Aguiar and Hurst to explore trends in leisure and nonmarket work time during the 1960s, 1970s, and 1980s. Like Aguiar and Hurst (2007c, 2009), they also find large increases in leisure time for men and women during the 20-year period between 1965 and 1985. Contemporaneous to Aguiar and Hurst, Ramey and Francis (2009) harmonized the US time-use data and documented trends in leisure and home production for the population as a whole and for men and women separately. Like Aguiar and Hurst (2007c), Ramey and Francis (2009) also found a large decline in aggregate home production time for prime-age individuals between 1960 and the early 2000s. Ramey and Francis (2009), however, find that there was very little increase in leisure for either prime-age men or women during this time period.[f]

[f] See Ramey (2007) and Aguiar and Hurst (2007a) for a reconciliation of the differences in leisure trends between the two papers. A large part of the debate is whether eating while at market work is considered market work (Aguiar and Hurst) or leisure (Ramey and Francis).

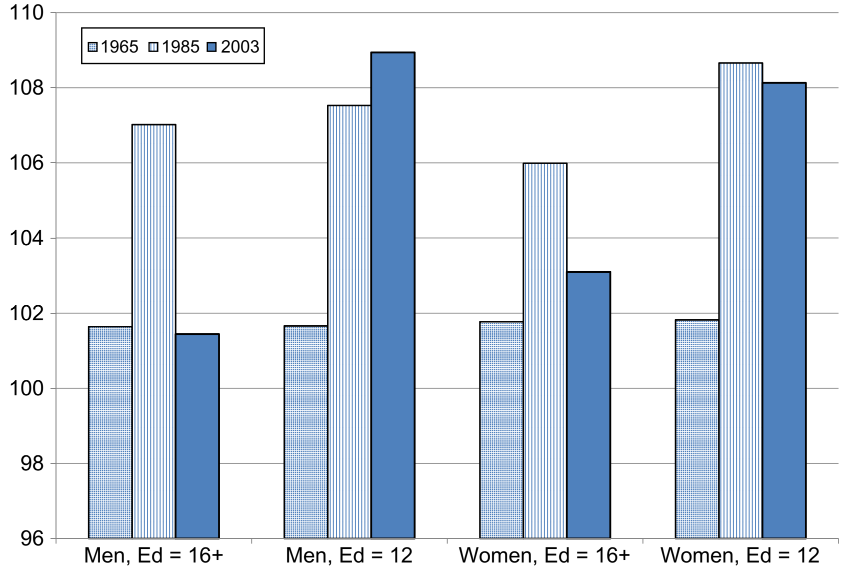


Fig. 11 Trends in leisure hours: all, men, and women, by skill. Note: Figure shows the amount of time allocated to leisure activities in 1965, 1985, and 2003 by sex and skill. The figure focuses on those with schooling levels of a bachelor's degree or more (Ed = 16+) and schooling levels of exactly a high school degree (ED = 12). Results in the figure come from tables V of Aguiar and Hurst (2007c). See text for additional details. Unlike the results in Fig. 7A–C, the results in this figure also adjust for the changing demographic composition over time within each sex-skill group. The demographic adjustment accounts for changing age distribution and family composition. The demographic adjustments made little difference to the broad time trends.

Additionally, Ramey and Francis (2009) incorporate the findings of Ramey (2009) into their analysis which allows them to compute trends in nonmarket work and leisure prior to 1965. This is a very ambitious task given that there are no nationally representative time diaries within the United States prior to 1965. The goal of Ramey (2009) is to use nonrepresentative time-use surveys conducted within the United States prior to 1965 to compute the amount of home production done in the United States for an average individual by weighting the nonrepresentative samples appropriately. Using this methodology, Ramey (2009) concludes that between 1900 and 1965, nonmarket work time for women fell by about 6 h per week, while nonmarket work time for men increased by about 7 h per week. Given the Ramey (2009) estimates, Ramey and Francis (2009) state that aggregate leisure increased by an additional 2 h per week for prime-aged individuals between 1900 and 1965.

In summary, there is ample evidence that home production has been declining in the aggregate and leisure has been increasing in the aggregate over long time periods.

5.2 Recent Trends in Time Use

One of the prominent downsides to harmonizing the different time-use surveys to compute long-run trends is that there is no guarantee that the data collection methods, sample frame, and time-use categorization remained constant over time. Changes in collection methods, sample frames, and categorization may cause the trends highlighted above to be mismeasured. The recent advent of the American Time-Use Survey (ATUS) helps to mitigate such issues. Since 2003, a nationally representative sample of individuals have been asked to record their time use using a consistently defined method and categorization procedure. Given the data have been in existence for 11 years now, it is possible to create time series trends using only within ATUS variation.

Using the sample described in the preceding section, Fig. 12 shows the trends in market work, nonmarket work, child care, and leisure over the 2003–2013 period. Each panel focuses on a different time-use category. Within each panel, four lines are shown.

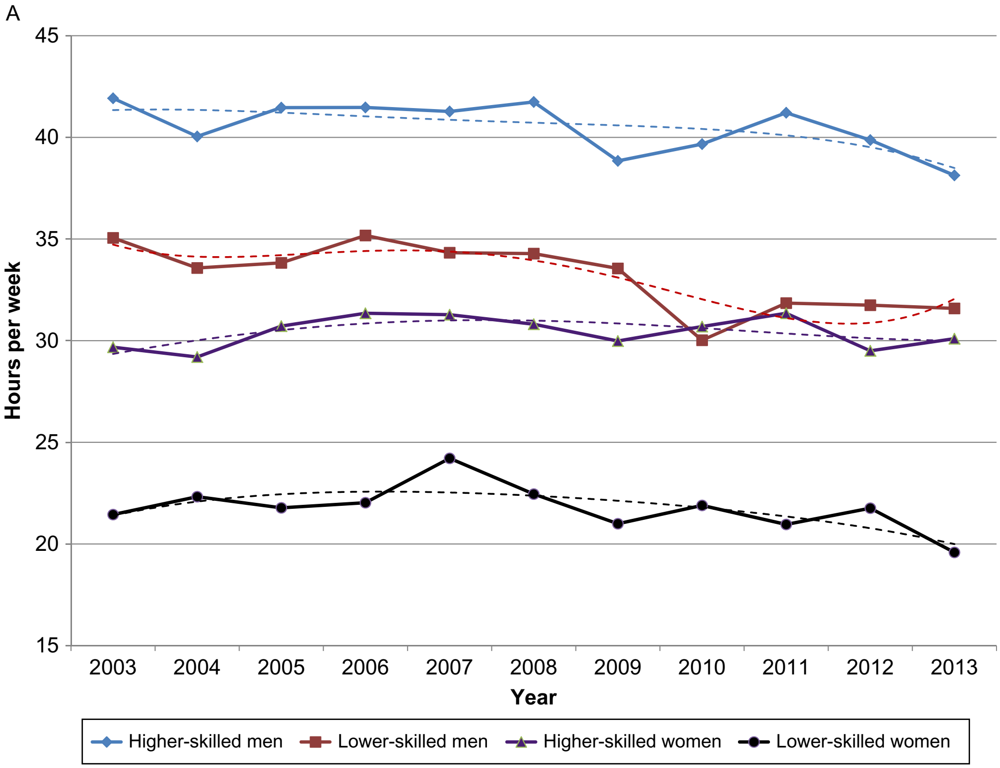


Fig. 12 ATUS trends by education and age. Note: Figure shows the trends in market hours (A), nonmarket work (B), child care (C), and leisure (D), per week worked for higher skilled men (*diamonds*), lower skilled men (*squares*), higher skilled women (*triangles*), and lower skilled women (*circles*) between 2003 and 2013. Data come from the American Time-Use Survey. The sample includes all individuals between the ages of 21 and 75 (inclusive) within the survey who had complete time diaries. Market work includes all time working on jobs for pay as well as any time commuting to work and any time spent at work associated with work meals and breaks. Nonmarket work includes activities such as cooking, cleaning, doing laundry, and shopping for groceries. Higher educated men are defined as those men with a bachelor's degree or higher. Lower educated men have years of schooling less than 16 years.

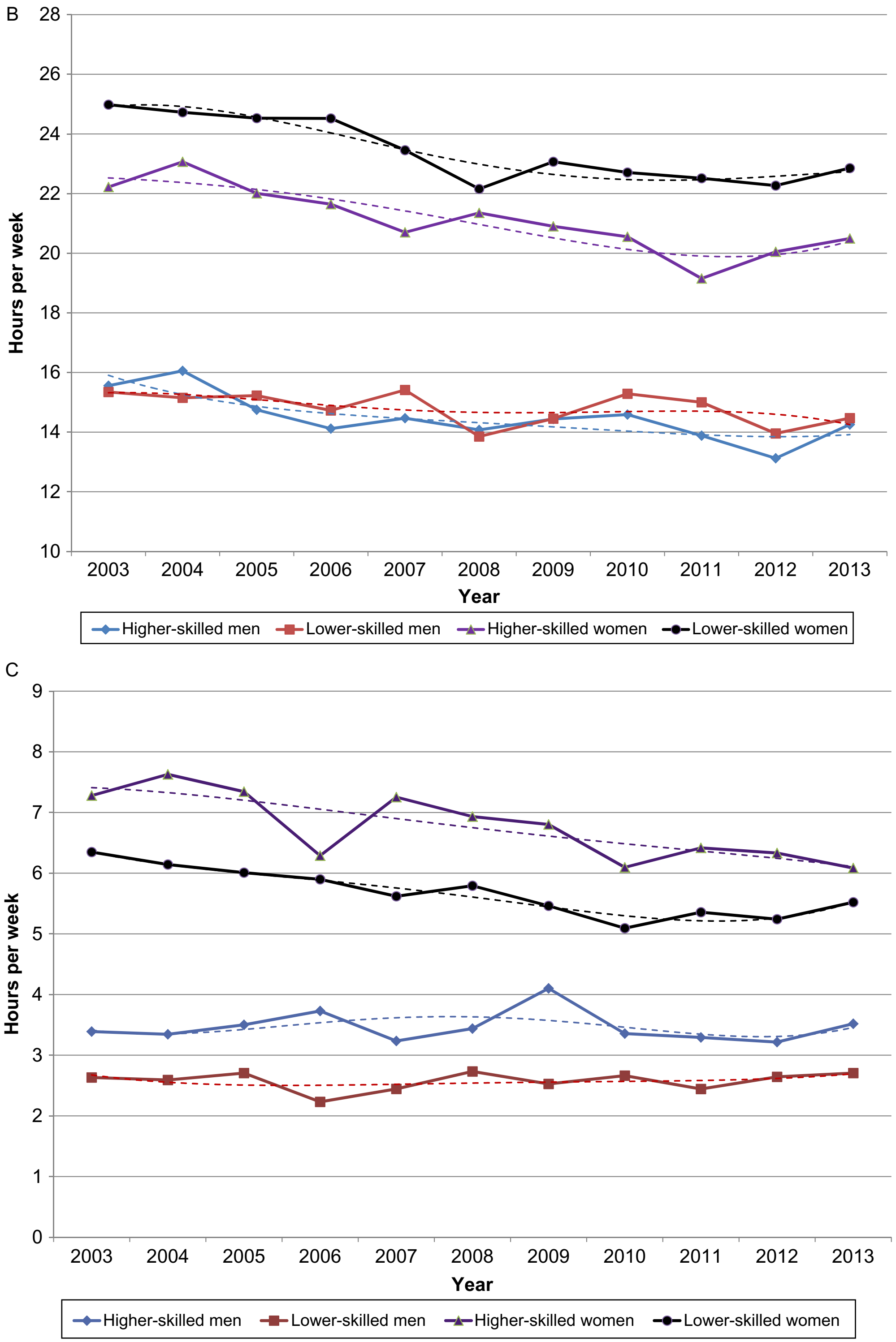

B
28
26
24
22
20
18
16
14
12
10
Hours per week
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Year
Higher-skilled men
Lower-skilled men
Higher-skilled women
Lower-skilled women
C
9
8
7
6
5
4
3
2
1
0
Hours per week
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Year
Higher-skilled men
Lower-skilled men
Higher-skilled women
Lower-skilled women

Fig. 12—Cont'd

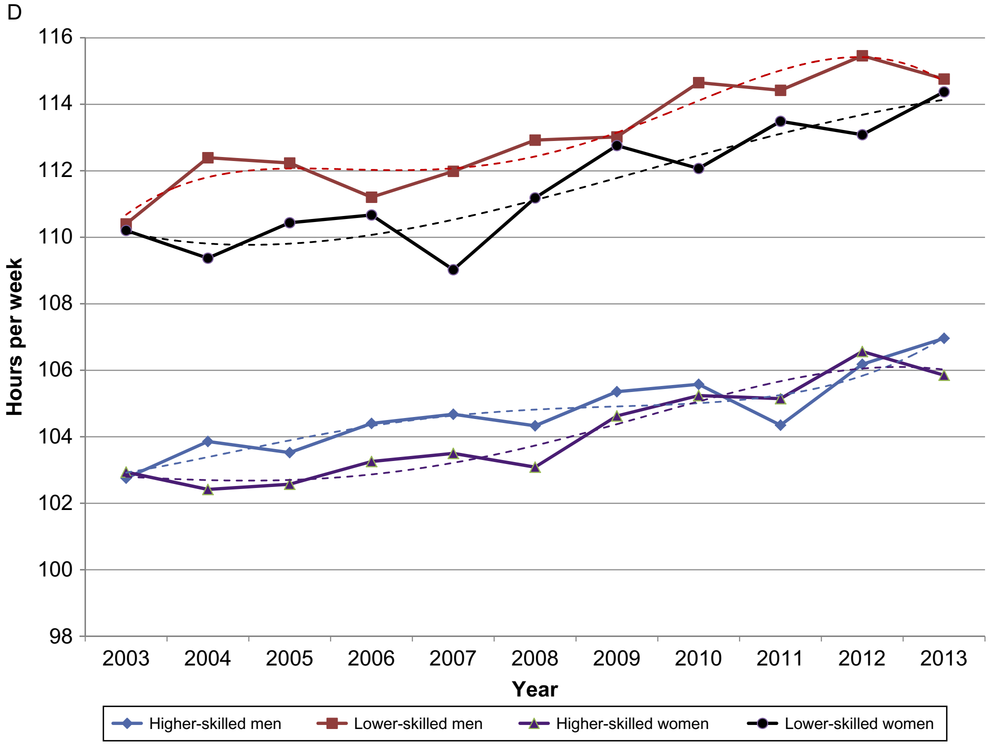


Fig. 12—Cont'd

Each line represents a sex-skill group pair. The data include all individuals between the ages of 21 and 75 who have all of their time use categorized by the ATUS. Fig. 13 is analogous to Fig. 12 except that the sample is restricted to individuals between the ages of 21 and 55.

Fig. 12A shows patterns similar to Figs. 5 and 6. During the last decade, all workers reduced the amount of time spent in market work with the declines being greater for those with less than at least a bachelors degree. Notice that the amount of time allocated to market work is higher in the ATUS relative to CPS totals documented in Figs. 5 and 6. The reason for this is that we are including time commuting to work and time spent at work during breaks and meals as being part of our market work measure. If we restrict our analysis to just time spent engaged in market work, the totals in the ATUS would be much closer to the market work totals reported in the CPS. Fig. 11A shows that the broad patterns are similar even restricting our analysis to those workers between the ages of 21 and 55 (as opposed to 21–75).

Figs. 12B and 13B show that home production has declined for all groups during the 2003–2013 period. For women, this just represents a continuation of the home production decline during the prior four decades. Notice that even within the ATUS, higher skilled women reduced their home production hours per week from about 22 h per week to about 19 h per week during the 2002–2013 period. This was made possible despite an overall decline in market work. As we show in the next section, a decline in market work is almost always associated with an increase in home production. What is also noticeable from Figs. 12B and 13B is that men actually reduced their nonmarket hours during this period as well. Again, this occurred despite their declines in market work hours. This

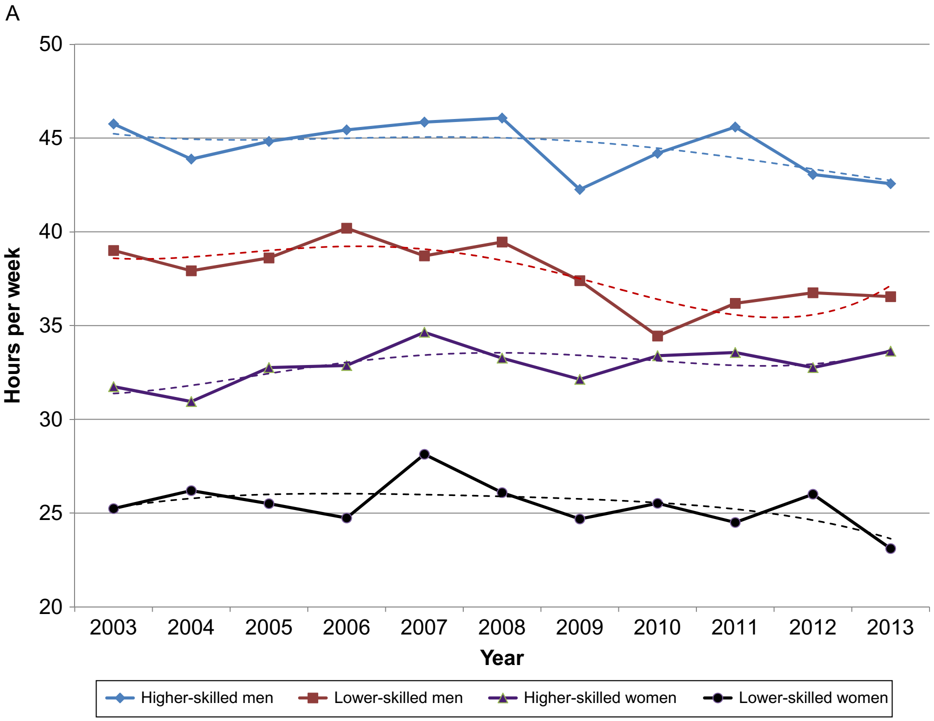


Fig. 13 ATUS trends by education and age: prime age. Note: Figure shows the trends in market hours (A), nonmarket work (B), child care (C), and leisure (D), per week worked for higher skilled men (*diamonds*), lower skilled men (*squares*), higher skilled women (*triangles*), and lower skilled women (*circles*) between 2003 and 2013. Data come from the American Time-Use Survey. The sample includes all individuals between the ages of 21 and 55 (inclusive) within the survey who had complete time diaries. Market work includes all time working on jobs for pay as well as any time commuting to work and any time spent at work associated with work meals and breaks. Nonmarket work includes activities such as cooking, cleaning, doing laundry, and shopping for groceries. Higher educated men are defined as those men with a bachelor's degree or higher. Lower educated men have years of schooling less than 16 years.

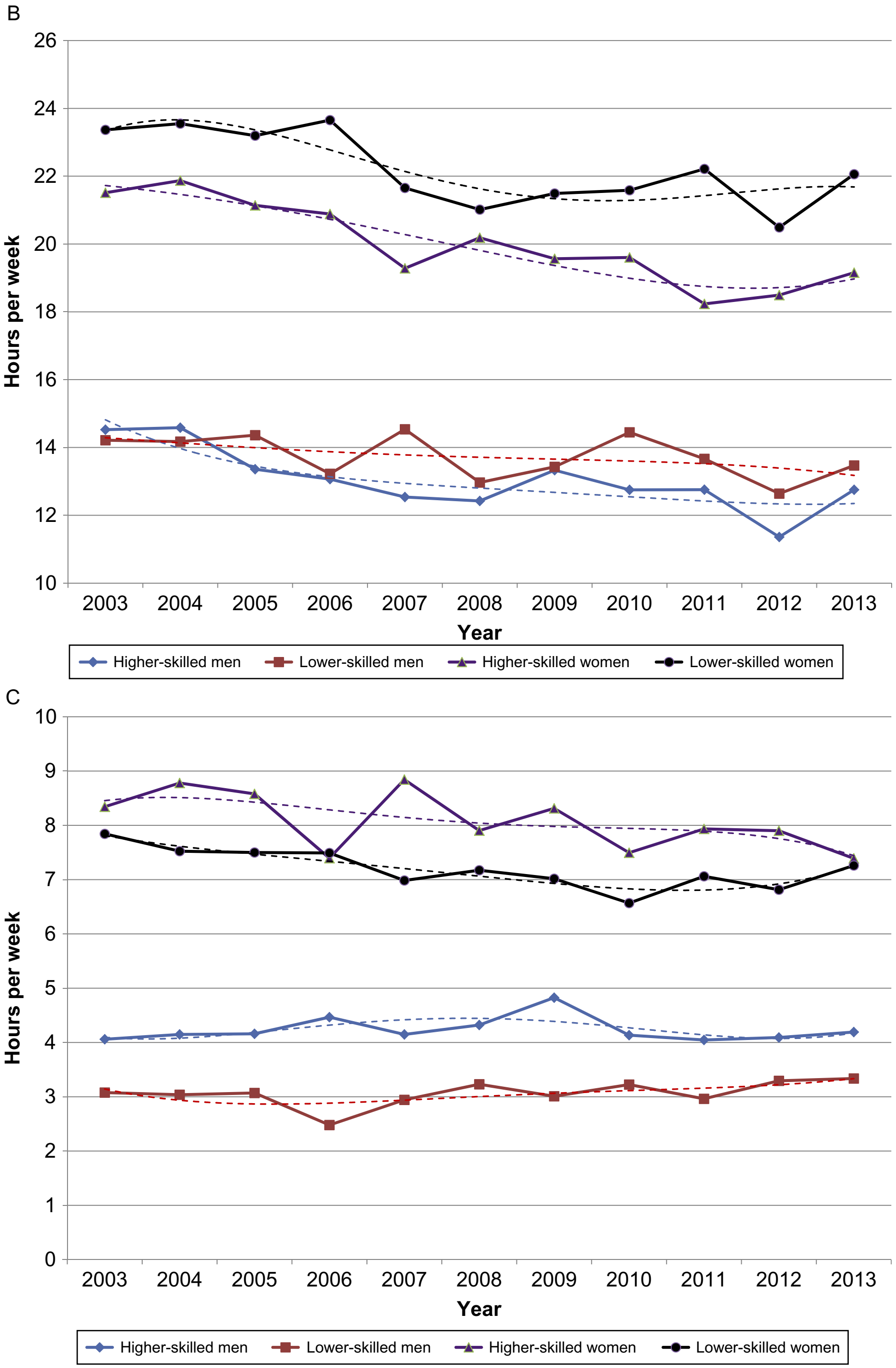

B
26
24
22
20
18
16
14
12
10
Hours per week
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Year
Higher-skilled men
Lower-skilled men
Higher-skilled women
Lower-skilled women
C
10
9
8
7
6
5
4
3
2
1
0
Hours per week
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Year
Higher-skilled men
Lower-skilled men
Higher-skilled women
Lower-skilled women

Fig. 13—Cont'd

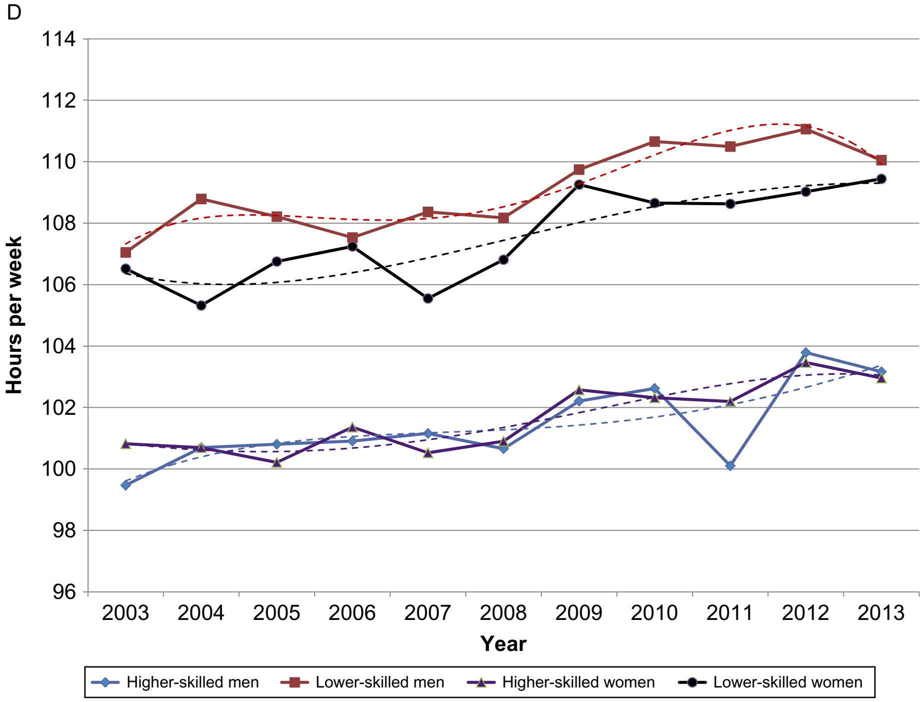


Fig. 13—Cont'd

recent trend is a slight reversal of the near constant nonmarket hours between 1985 and 2003 highlighted in the prior section.

Figs. 12C and 13C show that trends in child care also reversed slightly relevant to the trends over the prior 20 years. Both higher and lower skilled women reduced their child care time by about 1 h per week between 2003 and 2013. This increase reduced much of the gains in child care time that occurred between 1985 and 2003. For men, child care time was essentially flat during the last decade.

Figs. 12D and 13D show the trends in leisure for higher and lower skilled men and women between 2003 and 2013. All groups experienced an increase in time allocated to leisure during this period. What is noticeable is that the trends are nearly identical in terms of both levels and growth rates within a skill category. For example, high-skilled men and women again have nearly identical times allocated to leisure despite having dramatically different time allocated to market work, home production, and child care. Likewise, low-skilled men and women have nearly identical time allocated to leisure. Prime-aged lower skilled individuals increased their time allocated to leisure by roughly 3 h per week over the last decade. Prime-aged higher skilled individuals increased their leisure time by about 2 h per week during the last decade. Again, the recent time series results suggest a continuation of the increased leisure inequality trends that have been occurring during the prior few decades.

5.3 Business Cycle Variation in Time Use

In the prior section, we showed that leisure time increased while market work and home production time fell for all sex-skill groups during the last decade. However, it is hard to tease out the time series trends from the potential effects of the recent business cycle using time series data alone. As described in Aguiar et al. (2013), business cycle effects can be estimated using cross-region data.

We begin this section by documenting the business cycle effects on time use by exploiting cross-region variation in employment changes during the recent recession. Specifically, we estimate the following specification:

$$\Delta Time_{kt}^{j} = \alpha_0^j + \alpha_1^j \Delta Time_{kt}^{market} + \epsilon_{kt}^{j},$$

where $\Delta Time_{kt}^{market}$ is the average hour per week change in market hours across individuals in state k between period t and $t + s$ and $\Delta Time_{kt}^{j}$ is the average hour per week change in time spent on category j across individuals in state k between period t and $t + s$. To estimate these relationships, we use data for all individuals between the ages of 21 and 75 in the ATUS samples between 2007 and 2013. To increase power when computing means at the state level, we collapse the underlying data into multiyear samples. In particular, we create state level means for each time-use category in 2007–2008, 2009–2010, and 2011–2013. For each state, we compute $\Delta Time_{kt}^{j}$ by taking the difference in average time spent in category j in state k between the two adjacent time periods (2009–2010 vs 2007–2008 and 2011–2013 vs 2009–2010). As a result, we have 102 observations in the regression (two observations each for the 50 states plus the District of Columbia). The identification restriction for this exercise is that the underlying trends in time use for each category are similar across states. Therefore, the state variation is isolating only the business cycle variation in time use.[g]

Fig. 14 shows the cross-state relationship between market work changes and home production changes (A), child care changes (B), leisure changes (C), and job search (D). The change in market work within each state during the adjacent time periods (measured in hours per week) is on the x-axis. This stays the same across each of the four panels. On the y-axis of each panel is the respective change in the relevant activity, also measured in hours per week. According to Fig. 14A, as market work hours fall at business cycle frequencies, 36% is reallocated to home production ($\alpha_1^{nonmarket} = -0.36$ with a standard error = 0.04). As seen in Fig. 14C, a fall in market work of 1 h at business cycle frequencies leads to an increase in leisure of 0.44 h ($\alpha_1^{leisure} = -0.44$ with a standard error = 0.04). Taking the two together, 80% of the foregone time from a decline in

[g] See Aguiar et al. (2013) for a more complete discussion of the identification issues.

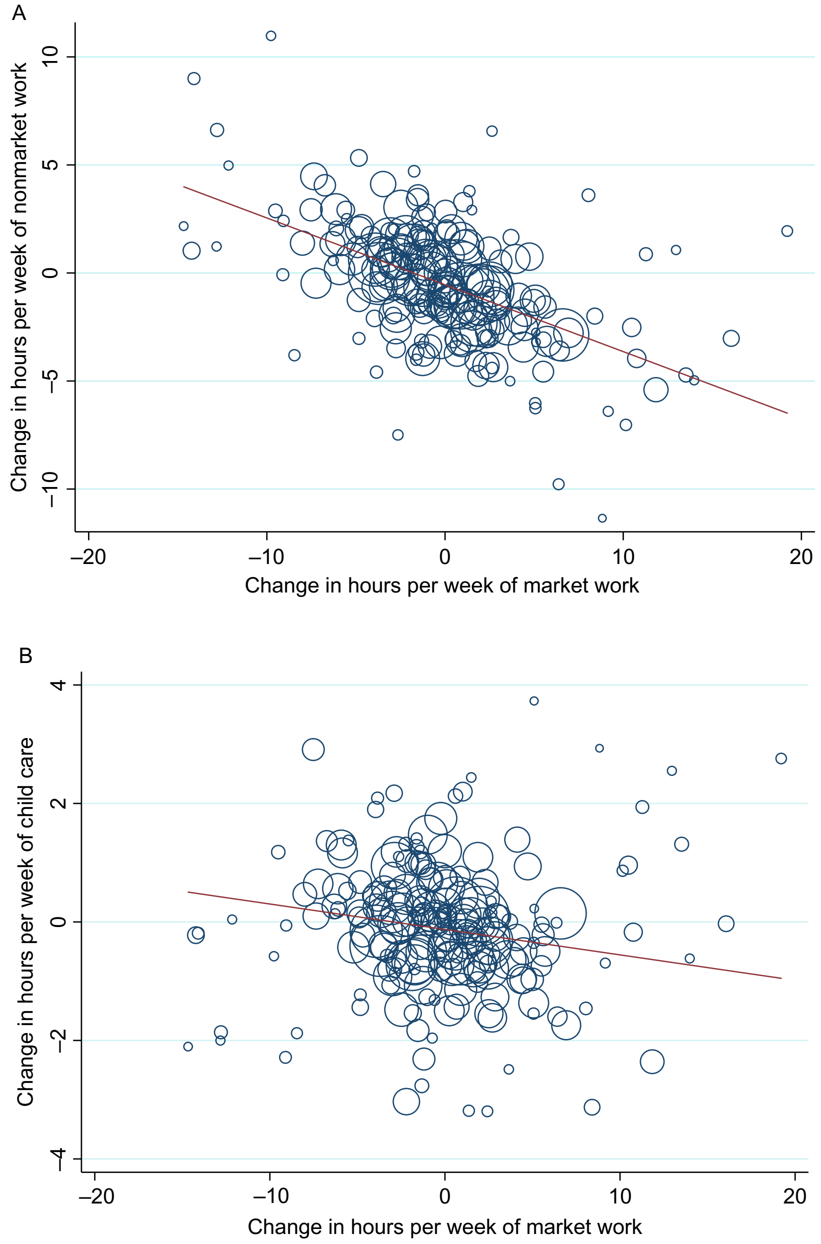


Fig. 14 Time allocation during the Great Recession. (A) Nonmarket hours, (B) child care, (C) leisure, and (D) job search. Note: Each panel shows change in market hours per week at the state level vs change in the indicated activity at the state level during the 2007–2013 period. For each state, three time-use observations are computed for each category: average time use in a given category pooled over years 2007 and 2008 (period 1), average time use in a given category pooled over years 2009 and 2010 (period 2), and average time use in a given category between 2011, 2012, and 2013 (period 3). The figure plots the change in time use between the first and the second period as well as the change in time use between the second and third time period. As a result, each state plus the District of Columbia is in the figure twice (for a total of 102 observations). The size of the *circle* represents the number of ATUS respondents within the state in the initial period from which the change is computed. The *line* is a weighted regression line through the scatter plots where the weights are the number of ATUS respondents within the state in the initial period from which the change is computed. The slope of the line is −0.31 with a standard error of 0.03 where the standard error is clustered by state.

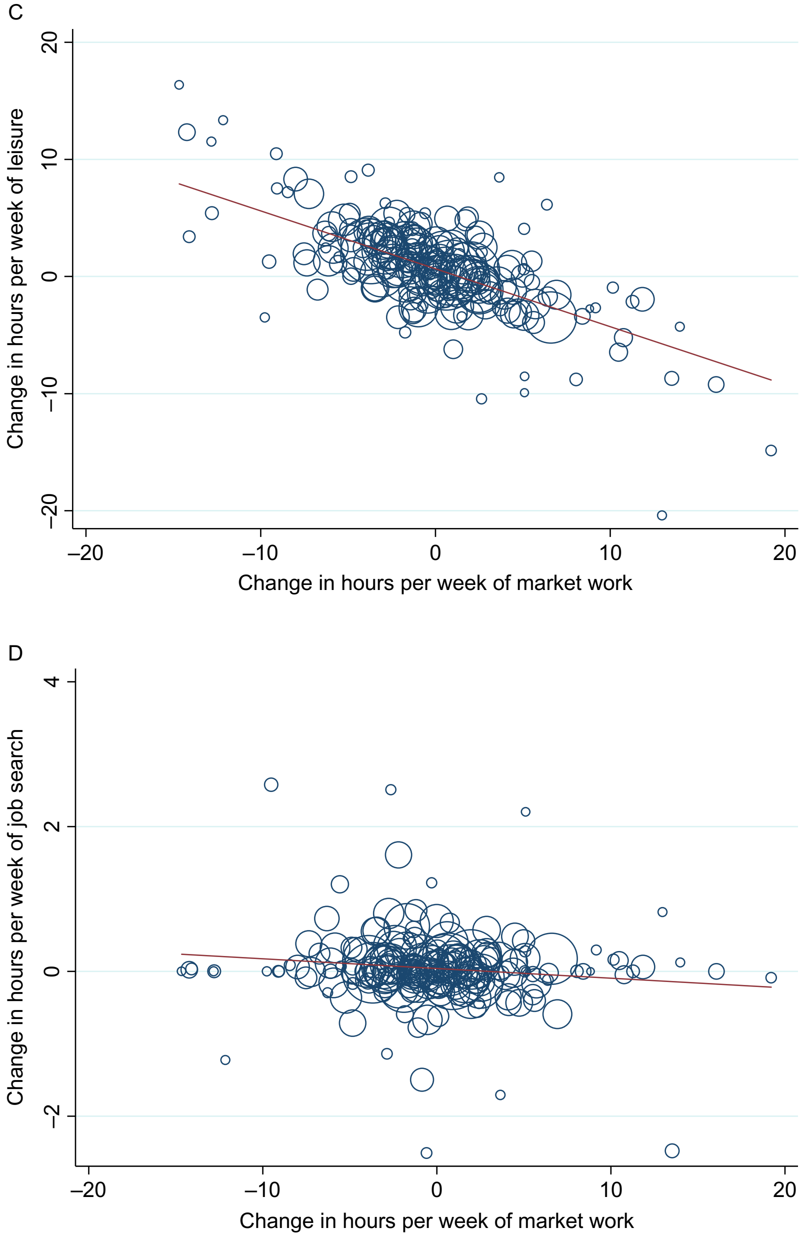


Fig. 14—Cont'd

market work is allocated to either leisure or home production. However, these findings complicate the interpretation of the time series trends shown in the prior sections. The fact that home production times fell for both high- and low-skilled men and women from the mid-2000s through 2013 despite the fact that the economy was in a recession may seem puzzling. If there were only business cycle factors driving the time series patterns, we would have expected home production times to increase as market work hours fell. The fact that home production times fell suggests that there was a large secular decline in home production time above and beyond the business cycle. This is not surprising given that home production times have been declining for decades.

Fig. 14B shows that child care time also increases in states as market work fell during the recession. Again, the time series patterns of time use suggest that during the recession child care time in the aggregate actually fell. The fact that aggregate time spent on child care activities fell despite the aggregate recession again suggests that there may have been a secular decline in child care time during the 2000s. If true, this would represent a reversal of the trends documented in Ramey and Ramey (2010), showing that time spent with children was increasing particularly among higher skilled parents.

While not formally extended in this chapter, Aguiar et al. (2013) show that investments in education, civic activities, and health care also absorb an important fraction of the decrease in market work hours (more than 10%), whereas job search absorbs around 1% of the decrease in market work hours (Fig. 14D). The latter finding is not surprising, given how little time unemployed spent searching for a job (Krueger and Mueller, 2010). The results suggest whether the job search measures in time-use surveys are designed to measure actual job search efforts of individuals looking for a job.

5.4 Time Use of the Unemployed

Another way to look at the effects of business cycle conditions on time use is to compare the time use of the unemployed relative to the employed. Such a comparison may suffer from composition differences across individuals. For example, individuals with a higher taste for leisure may be more likely to end up in the unemployment pool. Despite that limitation, we feel it is still informative to document the time use of individuals with different labor market status.

Table 1 shows the allocation of time in market work, nonmarket work, child care, leisure and other for men with at least 16 years of schooling (top panel) and men with less than 16 years of schooling. Each column represents a distinct labor market status. The first and second columns include men employed in the formal market sector (column 1) and men who are unemployment (column 2). The unemployed men are those individuals who are currently not working but who are actively seeking employment. Columns 3 and 4 include men who are out of the labor force. This category includes those who are disabled, retired, students, or who are otherwise not working and not seeking employment. We segment those out of the labor force into those under 63 and those 63 and over. The reason for this bifurcation is to identify potentially retired households. Most households over the age of 63 who are not attached to the labor force are retired.

Table 1 Time allocation by employment status: men

	More educated			
Activity	**Employed**	**Unemployed**	**NILF (age $<$ 63)**	**NILF (age $\geq$ 63)**
Leisure	100.27	121.47	127.80	134.11
Market work	47.70	1.98	0.47	0.11
Job search	0.09	9.37	0.58	0.00
Home production	12.73	23.64	21.42	25.26
Child care	3.59	4.25	2.70	1.71
Other	3.45	6.86	14.70	6.55
Observations	13,746	412	783	1,054

	Less educated			
Activity	**Employed**	**Unemployed**	**NILF (age $<$ 63)**	**NILF (age $\geq$ 63)**
Leisure	103.36	131.58	139.14	140.42
Market work	46.15	0.76	0.38	0.20
Job search	0.11	4.90	0.22	0.00
Home production	12.91	21.89	16.85	20.62
Child care	2.63	3.74	2.49	1.21
Other	2.72	4.59	8.71	5.31
Observations	22,319	1625	3603	3399

A few things are noticeable from Table 1. First, higher (lower) educated men who are unemployed still allocate roughly 2 (1) h per week to market work. All of this work, however, is outside the formal sector. This work includes side jobs for pay outside the formal sector. Second, higher educated unemployed men spend roughly 9 h per week in job search. The comparable number for lower educated men is 5 h per week. The number is essentially zero for employed men and men out of the labor force regardless of years of schooling. Third, like with the business cycle analysis discussed above, roughly 47% of the foregone difference in market work hours for higher skilled men (21/45) and 62% of foregone difference in market work hours for lower skilled men (28/45) are allocated to leisure. About 20–25% of the difference in work hours between unemployed and employed men—regardless of skill—is allocated to nonmarket work. The increase in leisure for lower skilled unemployed relative to the higher skilled unemployed is primarily due to differences in job search.

Table 2 shows similar patterns for women. The main difference between men and women is that lower educated women and higher educated women both have an increase in leisure time that represents roughly 45% of foregone differences in market work between the employed and unemployed. That is much smaller than the 62% of foregone work hours for lower educated men. Again, regardless of the analysis we perform—time series, life cycle, or business cycle—lower educated men take the most leisure.

Table 2 Time allocation by employment status: women

	More educated			
Activity	**Employed**	**Unemployed**	**NILF (age $<$ 63)**	**NILF (age $\geq$ 63)**
Leisure	99.56	115.96	112.79	128.24
Market work	40.96	0.78	0.19	0.17
Job search	0.10	4.74	0.11	0.00
Home production	17.75	29.40	30.79	29.27
Child care	5.36	7.68	14.89	2.06
Other	4.13	9.28	8.99	8.08
Observations	13,878	548	2,825	1,234

	Less educated			
Activity	**Employed**	**Unemployed**	**NILF (age $<$ 63)**	**NILF (age $\geq$ 63)**
Leisure	102.13	119.33	121.51	131.28
Market work	37.57	0.44	0.22	0.06
Job search	0.04	2.85	0.08	0.00
Home production	19.57	28.77	28.97	28.51
Child care	4.62	8.81	9.61	1.76
Other	3.90	7.07	7.32	6.23
Observations	22,665	2068	8878	5671

One final question we want to address is whether the long-term unemployed have different allocation of time relative to shorter term unemployed. If differences exist, it could represent either selection or potential duration dependence on time use. However, as seen in Table 3, there does not appear to be any differential time-use patterns between the short- and long-term unemployed. To measure the duration of unemployment, we bring in data from the individual's labor market status in their last interview of the CPS. As discussed above, the ATUS sample is drawn from the exiting rotation of the CPS. In the last interview of the CPS, an individual's current employment status is measured. If the individual is unemployed, it asks the duration of their unemployment spell. While the ATUS asks respondents of their current employment status, it does not ask them the duration of their unemployment spell if they were unemployed. By linking individuals across the two samples, we can get an imperfect measure of current unemployment duration.[h]

In Table 3, we restrict our sample to those individuals who are unemployed (not working and currently looking for job) in the ATUS who were either employed or unemployed in the CPS 3 months earlier.[i] We then estimate the following regression:

[h] There is no information on employment spells between the CPS and ATUS interviews.

[i] We restrict observations to having a 3-month gap between the ATUS and CPS. This was the overwhelming majority of ATUS respondents.

Table 3 Time use of the unemployed: duration dependence

Duration (weeks)	Leisure	Search	Home production	Child care
0–9	0.23	−0.70	0.34	0.37
	(1.57)	(0.77)	(1.29)	(0.69)
10–19	0.43	0.48	−0.80	−0.23
	(1.95)	(0.96)	(1.61)	(0.86)
20–29	−0.97	−2.16	2.23	1.95
	(2.51)	(1.23)	(2.08)	(1.10)
30–39	−1.53	1.83	0.04	0.59
	(2.61)	(1.29)	(2.16)	(1.15)
40–49	−5.64	2.86	−0.10	1.53
	(3.58)	(1.76)	(2.95)	(1.57)
50+	3.14	−1.23	1.11	−0.20
	(1.76)	(0.87)	(1.45)	(0.77)

Note: The sample consists of ATUS respondents between the ages of 21 and 62 who report being unemployed at time of ATUS interview and whose interview is 3 months after last CPS interview. The sample size is 2164. The omitted group consists of respondents who were employed at the time of the last CPS interview. The rows of the table report coefficients on dummy variables for being unemployed at the time of the CPS interview for a duration of 0–9 weeks, 10–19 weeks, etc. Other controls include age, age squared, marital status, a dummy indicating having a child, and a dummy indicate race=white.

$$Time^j_{it} = \beta^j_0 + \beta^j_1 UnempDur_{it} + \beta_2 X_{it} + \beta_4 D_t + \eta^j_{kt},$$

where $Time^j_{it}$ is the time use of individual i in time t on category j, $UnempDur_{it}$ is the duration of the respondent's unemployment spell as measured in the CPS 3 months earlier, X_{it} is a vector of individual-level controls, and D_t is a vector of 1-year time dummies. The X_{it} vector includes age, age squared, a marital status dummy, a dummy for whether the individual had a child, and a race dummy. The unemployment duration measure is a series of dummy variable indicating the length of the CPS unemployment spell: 0–9, 10–19, 20–29, 30–39, 40–49, and 50+ weeks. The omitted dummy in the regression is those individuals who were employed in their last CPS interview but are currently unemployed. As a result, the regression estimates how time use among the current unemployed differs by the duration of their CPS unemployment spell relative to the current unemployed who were working in their last CPS interview. If unemployment spells are persistent, those unemployed in the ATUS working in their last CPS interview will have shorter unemployment durations than those unemployed in the ATUS who were also unemployed in the CPS. It should be stressed that this is an imperfect measure of unemployment duration because we do not observe the individual's employment status in 3 months in between the CPS and ATUS.

The results in Table 3 show that there is no statistically significant relationship between time use and the duration of the unemployment spell in the CPS. However, standard errors of our estimates are large. As a result, we cannot rule out that time use evolves with the duration of unemployment. Additionally, as discussed above, there is some noise in the unemployment duration measure. Just because an individual was unemployed for 10 weeks in their last CPS interview does not mean they were unemployed for 22 weeks when we measure them in the ATUS. There is, on average, 12 weeks between an individual's CPS and ATUS interview. The individual could have found employment in that interval but because unemployed again by the start of the ATUS. We view this as suggestive evidence at best about the relationship between unemployment duration and time use.

5.5 Macro Implications of Time Use over the Business Cycle

One of the most important contributions of the economics of time is in improving our understanding of aggregate fluctuations. The first wave of dynamic general equilibrium models, pioneered by Kydland and Prescott (1982), assumed that total time is allocated into only two activities, market work and leisure. There are good reasons why introducing a third activity, time spent on home production, can make a difference for these models. First, when individuals derive utility both from market-produced goods and from home-produced goods, volatility in goods and labor markets can arise because of relative productivity differences between the two sectors, and not just because of productivity shocks in the market sector. Second, relative price changes cause households to substitute goods and time not only intertemporally between periods but also intratemporally between the market and the home sector. Intratemporal substitution introduces a powerful amplification channel which is absent from the standard real business cycle model. In fact, in his review of the home production literature Gronau (1997) writes that "…the greatest contribution of the theory of home production in the past decade was in its service to the better understanding of consumption behavior and changes in labor supply over the business cycle."

The first papers to introduce home production into the stochastic neoclassical growth model were Benhabib et al. (1991) and Greenwood and Hercowitz (1991). Benhabib et al. (1991) show that the real business cycle model with home production performs better than the standard real business cycle model along a number of dimensions. Specifically, in a calibrated version of their model, one of the main findings is that home production increases the volatility of labor and consumption relative to output. This is because home production introduces an additional margin of substitution toward which market work and market consumption can be directed following exogenous technology shocks. Second, the introduction of technology shocks in the home sector lowers significantly the correlation of productivity with labor hours. This is because technology shocks in the home sector shift the labor supply schedule and tend to generate a negative correlation between productivity and hours. This tends to offset the positive correlation induced by technology shocks in the market sector which shift the labor demand schedule.

However, the model also produces some notable discrepancies relative to the data. As Greenwood and Hercowitz (1991) show, the model produces a counterfactual negative correlation between investment in the market sector and investment in the home sector. This is because in a two-sector frictionless model, resources tend to flow to the most productive sector. In general, this implies that investment does not increase in both sectors simultaneously following a technology shock in one of the sectors. Greenwood and Hercowitz (1991) show that introducing highly correlated technology shocks between the home and the market sector and increasing the complementarity of time and capital in the production of home goods help address this discrepancy. Chang (2000) shows that adjustment costs in the accumulation of capital help resolve the investment anomaly when time and capital are substitutes in the production of home goods.

6. LIFE CYCLE VARIATION IN TIME USE

The economics literature typically analyzes life cycle patterns of consumption and work by appealing to models that emphasize only the intertemporal substitution of goods and time. However, as discussed above, intratemporal substitution between time and goods could be important for explaining the life cycle patterns of both time use and expenditures. In this section, we begin by documenting life cycle patterns in time use for both men and women of different schooling levels. We then briefly highlight recent research that has found evidence on the importance of intratemporal substitution in explaining life cycle profiles of expenditure.

6.1 Life Cycle Profiles of Time Use

When estimating the life cycle profiles of time use, one has to consider the potential that either time or cohort effects are driving the results. However, as is well known, colinearity prevents the inclusion of a full vector of time dummies, cohort dummies, and age dummies when estimating life cycle profiles. In particular, as discussed in Hall (1968), age, year, and cohort effects are identified in repeated cross sections up to a log-linear trend that can be arbitrarily allocated across the three effects. To isolate age profiles, additional assumptions are required.

In the remainder of this section, we proceed in two steps. First, we assess the extent to which cohort effects alter the life cycle profiles of market work using repeated cross-sectional data from the CPS between 1967 and 2013. Second, we then document the life cycle profiles of market work, home production, child care, and leisure using repeated cross sections from the ATUS between 2003 and 2007. For the latter analysis, we stop in 2007 to isolate periods before the Great Recession took place.

Fig. 13A–D uses the CPS data to show the life cycle patterns for market work for higher educated men, lower educated men, higher educated women, and lower educated women, respectively. As above, "higher educated" means having at least 16 years of schooling. Specifically, each figure shows the age coefficients (relative to age 25) from the following regression:

$$market\_hours^g_{it} = \beta^g_0 + \beta^g_{age} Age_{it} + \beta^g_c Cohort_{it} + \beta^g_t D^{norm}_t + \varepsilon^g_{it}, \tag{4}$$

where $market\_hours^g_{it}$ is market hours of household i during year t from group g, Age_{it} is a vector of 50 1-year age dummies (for ages 26–75) referring to the age of the household head, $Cohort_{it}$ is a vector of 1-year birth cohort dummies, and D^{norm}_t is a vector of normalized year dummies. Our approach is to attribute hours differences across households to age and cohort effects and use year dummies to capture cyclical fluctuations. Specifically, we restrict the year effects to average zero over the sample period. Henceforth, we refer to the year dummies with this restriction on their coefficients as normalized year dummies.

Each of the four panels in Fig. 15 contains three lines. The first line estimates the above equation as is using the CPS data from 1967 through 2013. These lines are represented with triangles on each of the four figures. The second line drops the cohort effects and does not restrict the year effects to sum to zero. Formally, we report the age coefficients from the following specification:

$$market\_hours^g_{it} = \beta^g_0 + \beta^g_{age} Age_{it} + \beta^g_t D_t + \varepsilon^g_{it}.$$

This specification is also estimated on the CPS data from 1967 through 2013. The second line is designated with squares on each of the figures. By comparing the first line to the second line, we can provide an assessment of the importance of omitting cohort effects when estimating life cycle profiles in market work off repeated cross sections. The third line on each figure—designated with the triangles—is the same as the second regression except restricted to the 2003–2007 period. By comparing the third line to the second, we can see the extent to which the life cycle profiles with no cohort effects and unrestricted time effects differ in the 2003–2007 period relative to the longer 1967–2013 period. This is important given that for the ATUS data, we will only be estimating life cycle profiles using the 2003–2007 period.

There are three interesting take aways from Fig. 15. First, the life cycle profiles of market work differ across sex-skill groups. For higher skilled men, market work hours per week increase by about 6–7 h between the ages of 25 and 31. Between 31 and 51, hours worked per week were roughly constant for these men. After the age of 51, market work hours declined steadily toward zero by age 75. For lower skilled men, market work hours did not increase as much between the ages of 25 and 31 (2–3 h per week). For these men, peak market hours worked per week occurred around 40 h per week. So lower skilled men start decreasing their hours worked per week much earlier than higher skilled men. The life cycle patterns for market work for higher skilled women is dramatically different relative to either lower or higher skilled men. Higher skilled women reduce their work hours per week by about 5 h between the ages of 25 and 35. These are the ages when higher skilled women leave the labor force to start families. However, by the early 40s, their market work hours per week are back to the levels in their mid-20s. Their hours remain high through their mid-50s before declining toward zero by age 75. Lower skilled women

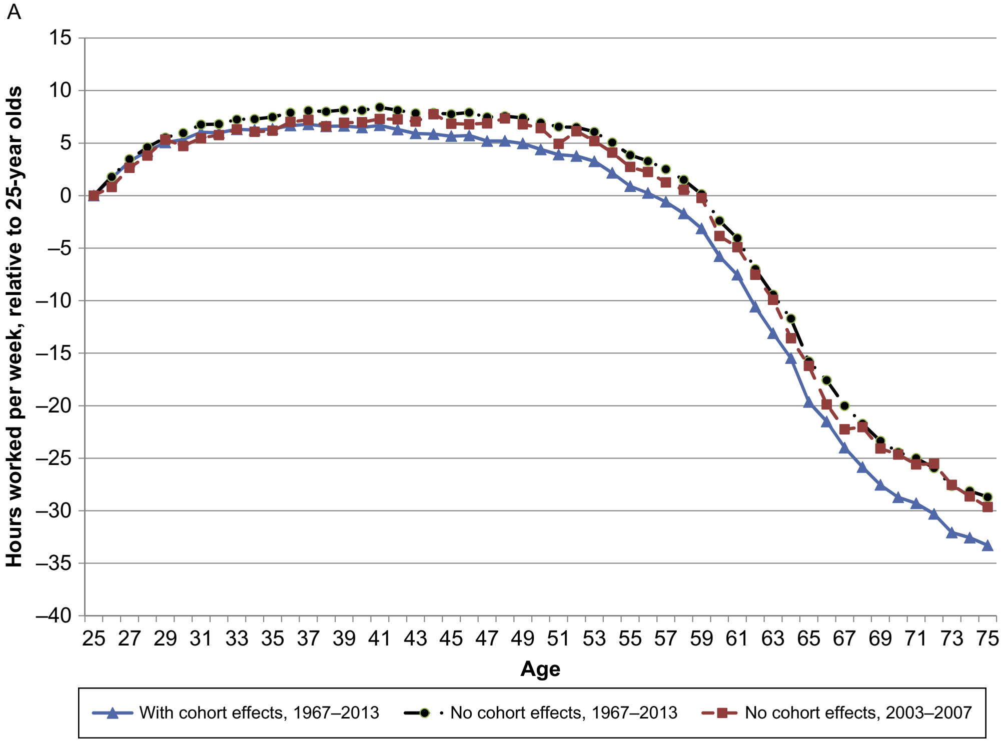


Fig. 15 Market hours over the life cycle. (A) More educated men, (B) less educated men, (C) more educated women, and (D) less educated women. Note: Figure shows the life cycle profile of market hours worked in the Current Population Survey (CPS) for men with at least 16 years of schooling (A), men with less than 16 years of schooling (B), women with at least 16 years of schooling (C), and women with less than 16 years of schooling (D). The *solid line with triangles* shows the life cycle profile using data from 1967 to 2013 controlling for 1-year cohort effects and normalized year effects. The normalized year effects are constrained to sum to zero across all years. The *dashed line with circles* shows the life cycle profile using data from 1967 to 2013 with no cohort effects but instead including year effects for each year separately. The *dashed–dotted line with squares* shows the life cycle profile using only data from 2003 to 2007 including year effects for each year separately.

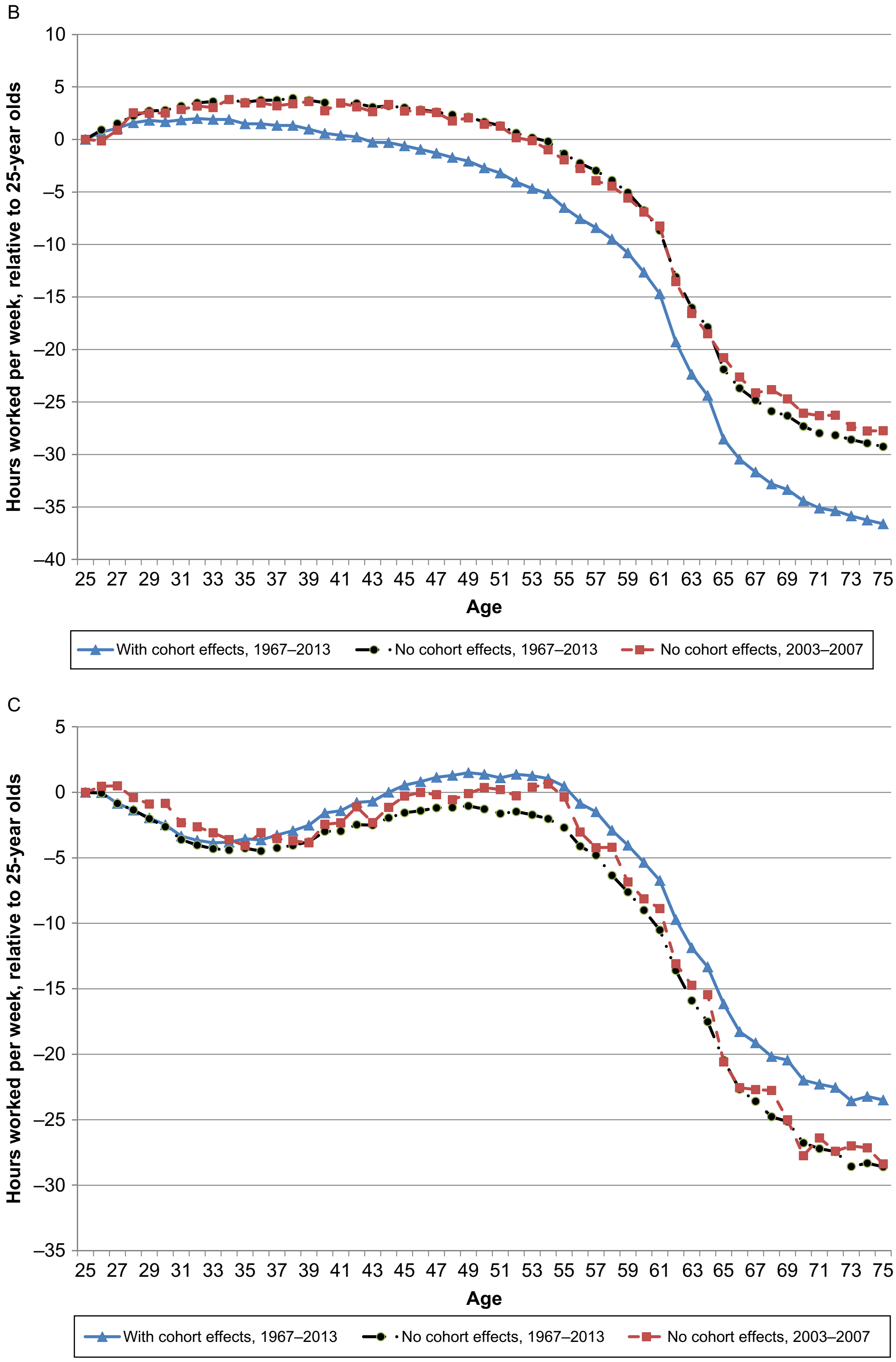

B
10
5
0
−5
−10
−15
−20
−25
−30
−35
−40
Hours worked per week, relative to 25-year olds
25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75
Age
With cohort effects, 1967–2013
No cohort effects, 1967–2013
No cohort effects, 2003–2007
C
5
0
−5
−10
−15
−20
−25
−30
−35
Hours worked per week, relative to 25-year olds
25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75
Age
With cohort effects, 1967–2013
No cohort effects, 1967–2013
No cohort effects, 2003–2007

Fig. 15—Cont'd

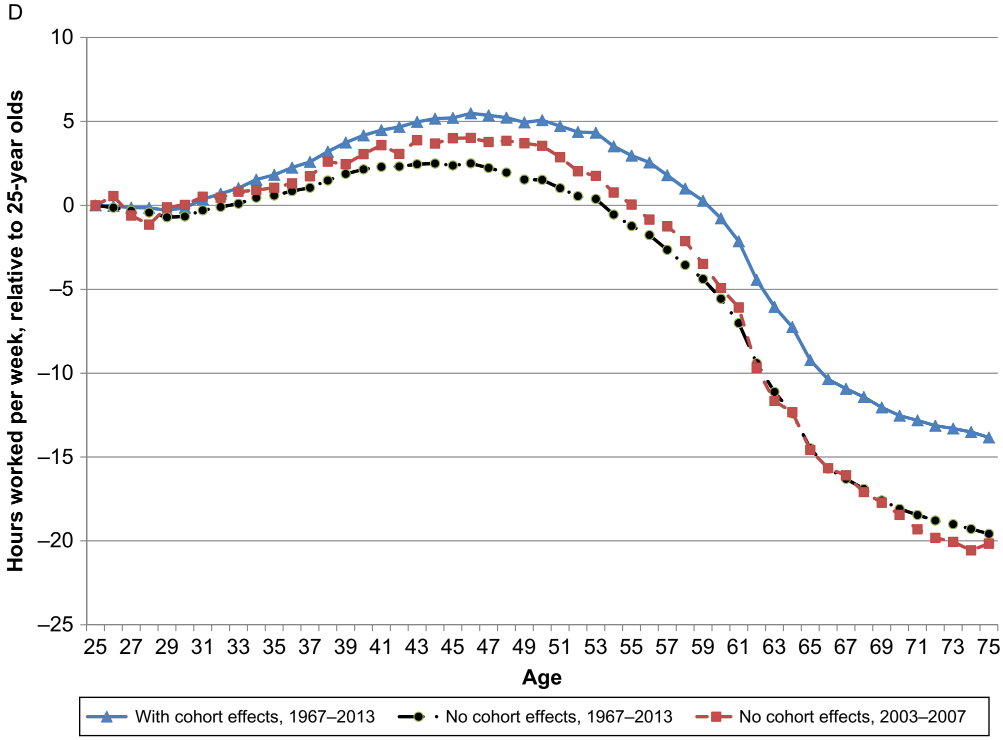


Fig. 15—Cont'd

have relatively low labor supply through their early 30s before increasing by roughly 3–5 h per week in their mid-40s.

The second thing to notice from Fig. 15 is that not controlling for cohort effects has only trivial effects on the life cycle profiles of market work for higher skilled men and women. This can be seen from the fact that the coefficients controlling for cohort effects (triangles) are nearly identical to the coefficients omitting the cohort effects (circles). When deviations exist, the differences are small. For example, controlling for cohort effects, higher educated men increase their hours worked per week by about 7 h per week between the ages of 25 and 40 and then decrease hours worked per week by about 41 h between 40 and 75. Without controlling explicitly for cohort effects, higher educated men increase their hours worked per week by about 8 h per week between ages 25 and 40 and then reduce hours worked per week by about 38 h between 40 and 75. The differences are slightly more pronounced for lower educated men and women. However, the life cycle patterns are for the most part quite similar regardless of whether or not one controls explicitly for cohort effects.

The final thing to notice from Fig. 15 is that life cycle profiles estimated from 1967 to 2013 with no cohort effects are again nearly identical as life cycle profiles estimated from 2003 to 2007 with no cohort effects. This fact holds for all sex-skill groups. This result gives us confidence that even though the ATUS data only start in 2003, the life cycle patterns we get from this period should be broadly consistent with the life cycle patterns over the past half century.

Fig. 16A plots the life cycle profiles of market work for higher educated men (diamonds), lower educated men (squares), higher educated women (triangles), and lower educated women (circles) using the 2003–2007 ATUS data. Instead of using 1-year age dummies, we regress hours per week in a given time-use category on a fourth-order polynomial in age. Using the coefficients from the fourth-order polynomial, we fit the predicted life cycle patterns for each time-use category. We use the fourth-order polynomial to smooth out some of the fluctuations over the life cycle in the 1-year age dummies given that the sample size of the ATUS is much smaller than the CPS. We then anchor the plots by taking the mean time use in each category for each sex-skill group at age 25.[j] This allows us to measure both the level and changes over the life cycle in hours per week allocated to a given activity.

Fig. 16A shows that the life cycle patterns in market work estimated of the cross section in the ATUS using 2003–2007 data are nearly identical to the patterns in Fig. 15A using CPS data. Higher educated men increase hours slightly from 25 to 40 before experiencing decline hours in their early 50s. Higher educated women decline their hours in market work between their mid-20s and mid-30s before increasing hours in market work through their early 50s. We view it as comforting that the life cycle patterns in market work in the ATUS are broadly similar with the life cycle patterns in the CPS.

Fig. 16B–D shows the life cycle patterns of time allocated to home production, child care, and leisure, respectively. Among younger individuals, lower educated women spend the most hours per week in nonmarket work. However, by the early 40s and throughout the remainder of the life cycle, the hours spent on home production for higher educated and lower educated women is nearly identical. All women, regardless of skill level, spend roughly 25 h per week in nonmarket work in their mid-40s. This number rises to about 30 h per week by age 65. Likewise, men spend nearly identical amounts in home production regardless of skill. As seen from Fig. 16B, the higher educated men and lower educated men lines are nearly on top of each other throughout most of the life cycle. Men spend about 12 h per week in home production in their mid-20s, about 15 h per week in their mid-40s, and about 20 h per week in their mid-60s. Between the ages of 40 and 70, the difference in home production hours per week between men and women narrows considerably. For all groups, as households age their time spent on home production increases.

Fig. 16C shows the life cycle patterns of time spent on child care for each group. A few things are noticeable from this figure. First, higher educated women have their peak in child care time around the age of 35. This is much later than the peak for lower

[j] When we report age 25 values, we actually take the mean for each sex-skill group for each category for ages 23–27. Again, we do this to help mitigate the measurement error given the smaller sample sizes within the ATUS.

educated women (around age 29). This reflects the fact that higher educated women have children later. Second, after the age of 29, higher educated women spend considerably more time in child care than lower educated women at every age. For example, at age 35, higher educated women allocate 17 h per week to child care. The comparable number is only about 10 h per week for lower educated women. Third, conditional on skill, men spend much less time on child care than do their female counterparts. Fourth, after the age of around 35, higher educated men spend much more hours per week in child care than lower educated women. Finally, higher educated men spend more time in child care at essentially every age. The uptick in time spent in child care in the 60s for higher educated men and women likely represents time spent with grandchildren.

Fig. 16D shows the life cycle patterns in leisure for all groups. Like the results above, lower skilled men experience the most leisure at every age of the life cycle. Higher educated men and women experience the least leisure at every age of the life cycle. However,

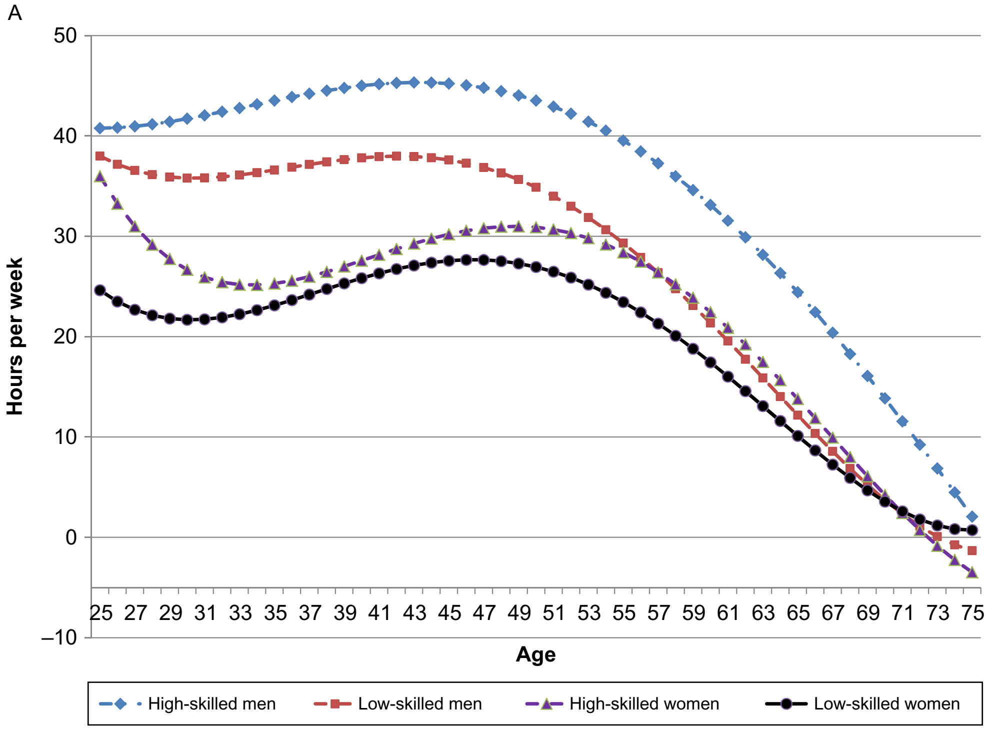


Fig. 16 Time allocation over the life cycle: ATUS data. (A) Market work, (B) nonmarket work, (C) child care, and (D) leisure. Note: Figure shows the life cycle profile of time allocation in the American Time-Use Survey (ATUS) by sex and skill group. The *line marked with diamonds* shows the pattern for men with at least 16 years of schooling. The *line marked with squares* shows the pattern for men with less than 16 years of schooling. The *line marked with triangles* shows the patterns for women with at least 16 years of schooling. The *line marked with circles* shows the patterns for women with less than 16 years of schooling. The profiles do not control for cohort effects but do include year effects for each year separately.

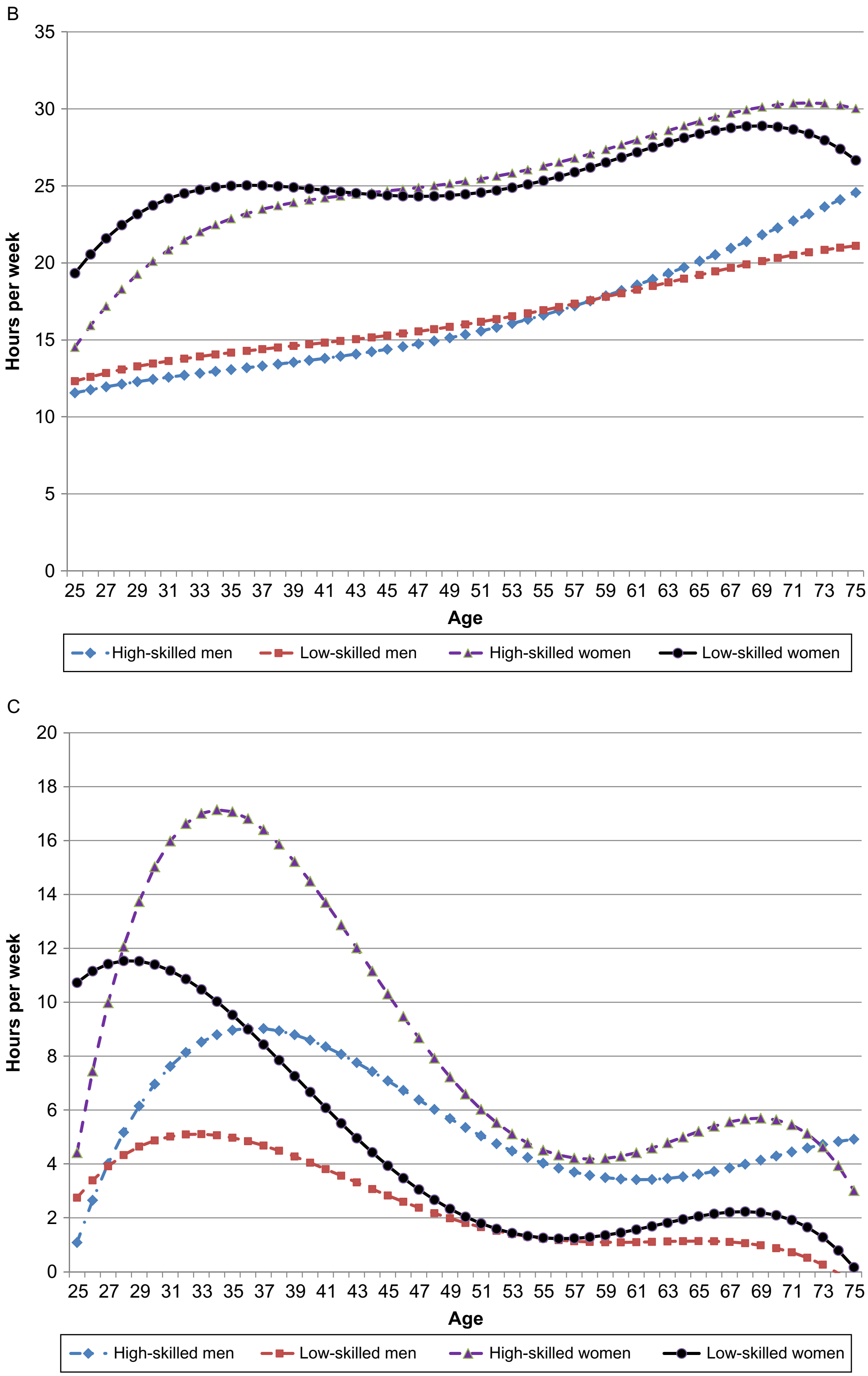

B
Hours per week
35
30
25
20
15
10
5
0
25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75
Age
High-skilled men
Low-skilled men
High-skilled women
Low-skilled women
C
Hours per week
20
18
16
14
12
10
8
6
4
2
0
25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75
Age
High-skilled men
Low-skilled men
High-skilled women
Low-skilled women

Fig. 16—Cont'd

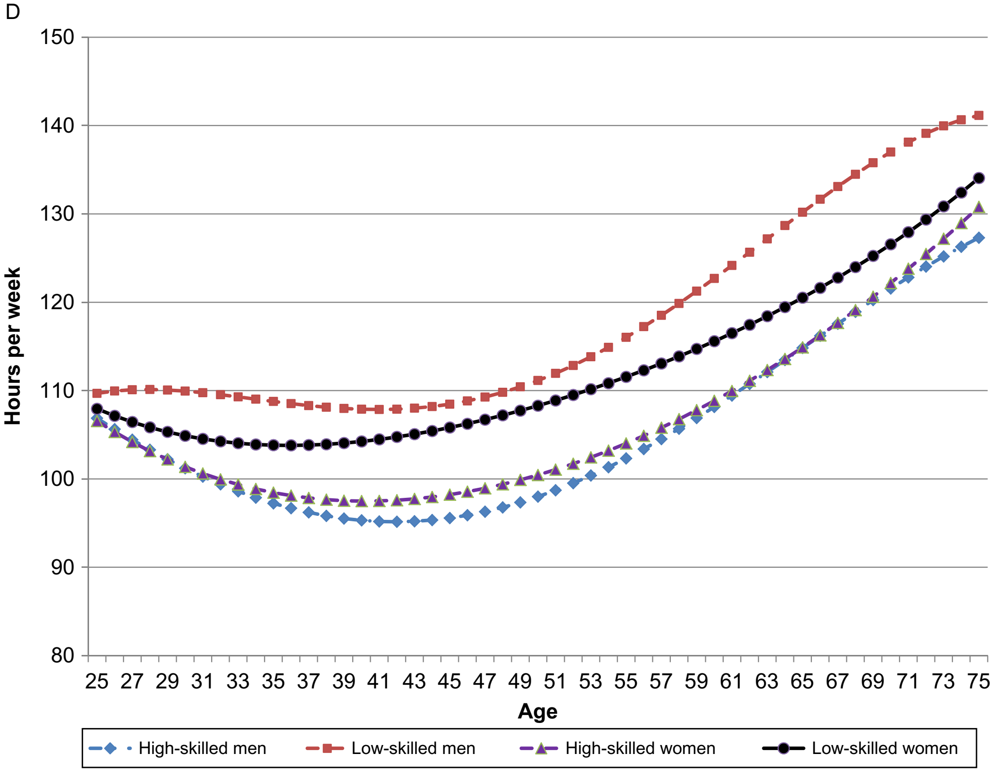


Fig. 16—Cont'd

one of the most striking facts from Fig. 16D is that despite the dramatic differences in market work, home production, and child care over the life cycle between higher educated men and women, their leisure times are nearly identical at every age. So, while the composition of work activities may differ between higher educated men and women, they are taking nearly identical amounts of leisure times. This is consistent with the time series evidence discussed above. Additionally, all households increase their leisure time dramatically after middle age. For example, higher educated men and women increase their weekly leisure time by about 35 h per week between the ages of 41 and 75. The increase is about 30 h per week for lower educated men and women.

6.2 The Importance of Intratemporal Substitution Between Time and Goods

The workhorse model of consumption over the life cycle, the permanent income hypothesis, posits that individuals allocate their resources in order to smooth their marginal utility of consumption across time (see, eg, Attanasio, 1999 for a review). If the marginal utility of consumption depends only on measured consumption, this implies that individuals will save early in their life cycle in order to maintain a smooth level of expenditures at retirement. During the last decade, there was a large amount of research that has showed that the substitution between time and expenditures is a first-order explanation as to why consumption varies over the life cycle.

The typical finding in the literature has been that consumption follows a hump-shaped pattern over the life cycle with consumption being low early in the life cycle, peaking at middle age and falling sharply at retirement. Some authors have argued that this life cycle profile represents evidence against the forward-looking consumption smoothing behavior implied by permanent income models, particularly since the hump in expenditures tracks the hump in labor income (as documented by Carroll and Summers, 1991). This view interprets expenditure declines in the latter half of the life cycle as evidence of poor planning. Other authors argue that the hump-shaped profile of consumption reflects optimal behavior if households face liquidity constraints combined with a need to self-insure against idiosyncratic income risks (see, for example, Zeldes, 1989; Deaton, 1991; Carroll, 1997; Gourinchas and Parker, 2002). Households build up a buffer stock of assets early in the life cycle, generating the increasing expenditure profile found during the first half of the life cycle. The decline in the latter half of the life cycle is then attributed to impatience once households accumulate a sufficient stock of precautionary savings.

In a recent paper, Aguiar and Hurst (2013) demonstrate that there is tremendous heterogeneity in the life cycle patterns of expenditures across different spending categories. In particular, some categories (eg, food and transportation) display the familiar hump-shaped profile over the life cycle, but other categories display an increasing (eg, entertainment) or decreasing (eg, clothing and personal care) profile over the life cycle. This heterogeneity cannot be captured by the standard life cycle model of consumption that emphasizes only the intertemporal substitution of goods and time. They show that home-produced goods (food) and work-related expenditures (clothing and nondurable transportation) account for the entire decline in total expenditures after middle age. Additionally, these same goods explain the overwhelming majority of the increase in the cross-individual dispersion in expenditures after middle age. The paper shows that failure to account for home-produced and work-related goods leads one to overestimate the amount of income risk faced by individuals.

A separate literature focused on the "retirement consumption puzzle." The literature found that that household expenditure falls discontinuously upon retirement. Banks et al. (1998) look at the consumption smoothing of British households around the time of retirement. Controlling for factors that may influence the marginal utility of consumption (such as family composition and age, mortality risk, labor force participation), they find that consumption falls significantly at retirement. Bernheim et al. (2001) find that total food expenditure declines by 6–10% between the preretirement and the postretirement period, which leads them to conclude that households do not use savings to smooth consumption with respect to predictable income shocks. Haider and Stephens (2007) use subjective retirement expectations as an instrument to distinguish between expected and unexpected retirements and find a decline in food expenditures ranging from 7% to 11% at retirement.

Aguiar and Hurst (2005) argue that tests of the life cycle model typically equate consumption with expenditure. However, as stressed by the model above, consumption is the output of a home production process which uses as inputs both market expenditures and time. As the above model highlights, individuals will substitute away from expenditures toward time spent on home production when the market price of time falls. Since retirees have a lower opportunity cost of time than their preretired counterparts, time spent on the production of commodities should increase during retirement. If this is the case, then the drop in expenditure does not necessarily imply a large decrease of actual consumption at retirement.

To test this hypothesis, Aguiar and Hurst (2005) explore how actual food consumption changes during retirement. Using data from the Continuing Survey of Food Intake of Individuals, a data set conducted by the US Department of Agriculture which tracks the dollar value, the quantity, and the quality of food consumed within US households, they find no actual deterioration of a household's diet as they transition into retirement. To test the hypothesis that retirees maintain their food consumption relatively constant despite the declining food expenditures, Aguiar and Hurst (2005) use detailed time diaries from the National Human Activity Pattern Survey and from the American Time-Use Survey and show that retirees dramatically increase their time spent on food production relative to otherwise similar nonretired households. That retirees allocate more time to nonmarket production has been also shown by Hurd and Rohwedder (2006) and Schwerdt (2005).

In light of these evidence, Hurst (2008) concludes that the retirement puzzle "has retired." That is, even though it is a robust fact that certain types of expenditures fall sharply as households enter into retirement, standard life cycle models with home production are able to explain this sharp fall because retirees spent more time producing goods.[k] Additionally, as we discuss in the next section, declines in expenditures are mostly limited to two types of consumption categories: work-related items (such as clothing and transportation expenditures) and food (both at home and away from home). When expenditures exclude food and work-related expenses, the measured declines in spending at retirement are either close to zero or even increasing.

A key parameter in whether household expenditures on a given good will increase or decrease as the household's opportunity cost of time falls is the elasticity of substitution between time and expenditures (σ from the theoretical discussion above) is greater than

[k] Hurst (2008) also discusses how health shocks that lead to early retirement can help reconcile the fact that actual consumption falls for a small fraction of households upon retirement.

or less than 1. In Aguiar and Hurst (2005) leisure goods are defined as goods for which the intratemporal elasticity between time and expenditures is less than 1. For these goods, spending increases when the opportunity cost of time falls (holding the marginal utility of wealth constant). For example, suppose that as individuals retire they play more golf. If the marginal utility of wealth was held constant during the retirement transition, golf would then be considered a leisure good. Conversely, Aguiar and Hurst argue that home-produced goods are goods for which the intratemporal elasticity between time and expenditure is great than 1 (holding the marginal utility of wealth constant). These goods may include groceries and cleaning services.

A large literature has developed to estimate the exact value of σ_i. Rupert et al. (1995) use home production time and food expenditure data from the Panel Study of Income Dynamics (PSID) to estimate σ for food. Most of their estimates point out for an elasticity that exceeds 1. Aguiar and Hurst (2007b) use data from the American Time-Use Survey. Assuming that the relevant opportunity cost of time is the marginal rate of technical substitution between time and goods in the shopping technology, they find a value of σ of around 1.8 for home-produced goods. Using PSID data, Gelber and Mitchell (2012) find that, in response to tax shocks, the elasticity of substitution between market- and home-produced goods is around 1.2 for single men and as high as 2.6 for single women. Finally, using consumer-level data on hours, wages, and consumption expenditure from the PSID and metro-level data on price indices p_i from the US BLS, Gonzalez Chapela (2011) estimates a life cycle model with home production and finds a value of σ in the production of food of around 2.

7. CONCLUSION AND DISCUSSION

The wealth of new data on measuring time use enable researchers to empirically investigate a variety of substantive questions in macroeconomics. Detailed diaries, linked to larger surveys, allow us to gain a better understanding of time series trends in market work, life cycle movements in household expenditures, and business cycle fluctuations in consumption and employment. This advances the agenda set forth in Gary Becker's Presidential Address. We conclude this chapter by highlighting some of the limitations of the existing time-use data and then discuss some directions for future research.

There are four major limitations to existing time-use surveys: (i) individual time-use data are not linked to individual data on expenditures; (ii) the data are from repeated cross sections, and do not contain a panel component; (iii) the data do not include measures of time use from multiple members of the same household; and (iv) the data do not measure detailed activities while at market work.

Researchers have worked around the lack of panel data by creating synthetic cohort data. Twenty-five-year-old white male high school graduates in year t of a time-use survey are, on average, the same individuals who are 26-year-old white male high school graduates in survey year $t + 1$. By tracking demographic groups across different years of cross-sectional data, synthetic panel data can be constructed. The synthetic cohort method also allows for a solution to the problem that time-use data and consumption data are measured in different surveys. If the samples are nationally representative, the consumption of 25-year-old white male high school graduates in year t from expenditure surveys can be merged with data for this same group in year t of the time-use surveys. The variation from the synthetic cohort method comes from variation across these demographic groups. Often this variation is enough to identify the questions of interest. But, the limitation is that lots of individual variation within a demographic group are thrown away when the synthetic panel method is used. Having panel data of time use—ideally in a survey which also measures expenditure—would allow researchers to exploit more variation to identify questions of interest. It would allow to compute changes in time allocation in response to, for example, demographic or employment status, while controlling for an individual's fixed characteristics. Moreover, multiple surveys would provide a better sense of how frequently an activity is undertaken.

Another major limitation of current time-use measurement is that we do not collect time-use information for multiple members of the same household. Many of the key questions that can be answered with time-use data can benefit from measuring the time use of multiple household members. If women start working more in the market, do their husbands work more at home? If one family member starts caring for an elderly parent, how is time use reallocated among additional family members? How do parents invest their time into their children? To really get a sense of the role of the family in explaining time series, life cycle, and business cycle variation in expenditure and labor supply, it is necessary to have time-use data that span multiple members of the same household.

Finally, no current nationally representative survey within the United States tracks in detail how individuals spend their time while at work. For example, within the American Time-Use Survey, time spent at market work is just one category. There is no additional detail provided about the tasks individuals perform while at work. It may be informative, for example, to know how much time individuals spend on the computer while at work vs in meetings. Or, alternatively, how much time an individual spends interacting with customers vs stocking shelves. How much time is spent in manual labor relative to time spent in cognitive activities? Making progress measuring how individuals allocate their time at work can help us to understand how the nature of work changes over time, over an individual's life cycle, and over the business cycle. As time-use surveys evolve, the type of questions researchers can answer will expand.

Nevertheless, the time-use data we now have available enable researchers to address many interesting macroeconomic questions. One line of research is obtaining a better understanding of labor supply, including how technological advances in nonmarket sectors shift labor force participation. Business cycle research can also benefit from incorporating data on time allocation. Particularly of interest is the time spent searching for employment, and the cyclical returns to job search. Time spent investing in children's human capital (viewed broadly) is also an active area of study. Time allocation is a key determinant of human capital accumulation, and it is important to quantify the return to time spent acquiring skills, on and off the job. More broadly, time-use surveys can shed light on how differences in the parental time allocated to child care influence the economic prospects of the next generation.

ACKNOWLEDGMENTS

We thank our discussant, Thibaut Lamadon, as well as the editors, John Taylor and Harald Uhlig, for helpful comments. The chapter also owes a debt to our collaborations with Loukas Karabarbounis. We thank Hilary Shi for excellent research assistance.

REFERENCES

Aguiar, M., Bils, M., 2015. Has Consumption Inequality Mirrored Income Inequality. AER 105 (9), 2725–2756.

Aguiar, M., Hurst, E., 2005. Consumption versus expenditure. J. Polit. Econ. 113, 919–948.

Aguiar, M., Hurst, E., 2007a. Comments on Valerie A. Ramey's "How much has leisure inequality really increased since 1965?" University of Chicago Booth Working Paper.

Aguiar, M., Hurst, E., 2007b. Lifecycle prices and production. Am. Econ. Rev. 97, 1533–1559.

Aguiar, M., Hurst, E., 2007c. Measuring trends in leisure: the allocation of time over five decades. Q. J. Econ. 122, 969–1006.

Aguiar, M., Hurst, E., 2009. The Increase of Leisure Inequality: 1965–2005. American Enterprise Institute Press.

Aguiar, M., Hurst, E., 2013. Deconstructing lifecycle expenditure. J. Polit. Econ. 121, 437–492.

Aguiar, M., Hurst, E., Karabarbounis, L., 2012. Recent developments in the economics of time use. Annu. Rev. Econ. 4, 373–397.

Aguiar, M., Hurst, E., Karabarbounis, L., 2013. Time use during the Great Recession. Am. Econ. Rev. 103, 1664–1696.

Attanasio, O., 1999. Consumption. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics. Amsterdam, New York: North Holland.

Banks, J., Blundell, R., Tanner, S., 1998. Is there a retirement-savings puzzle? Am. Econ. Rev. 88, 769–788.

Becker, G., 1965. A theory of the allocation of time. Q. J. Econ. 75, 493–517.

Becker, G., 1989. Family economics and macro behavior. Am. Econ. Rev. 78, 1–13.

Benhabib, J., Rogerson, R., Wright, R., 1991. Homework in macroeconomics: household production and aggregate fluctuations. J. Polit. Econ. 99, 1166–1187.

Bernheim, B.D., Skinner, J., Weinberg, S., 2001. What accounts for the variation in retirement wealth among U.S. households? Am. Econ. Rev. 91, 832–857.

Carroll, C., 1997. Buffer stock saving and the life cycle/permanent income hypothesis. Q. J. Econ. 112, 1–56.

Carroll, C., Summers, L., 1991. Consumption growth parallels income growth: some new evidence. In: Bernheim, D., Shoven, J. (Eds.), National Saving and Economic Performance. University of Chicago Press, Chicago.

Chang, Y., 2000. Comovement, excess volatility, and home production. J. Monet. Econ. 46, 385–396.

Deaton, A., 1991. Saving and liquidity constraints. Econometrica 59, 1221–1248.

Gelber, A., Mitchell, J., 2012. Taxes and time allocation: evidence from single women. Rev. Econ. Stud. 79, 863–897.

Gonzalez Chapela, J., 2011. Recreation, home production, and intertemporal substitution of female labor supply: evidence on the intensive margin. Rev. Econ. Dyn. 14, 532–548.

Gourinchas, P.O., Parker, J., 2002. Consumption over the life cycle. Econometrica 70, 47–89.

Greenwood, J., Hercowitz, Z., 1991. The allocation of capital and time over the business cycle. J. Polit. Econ. 99, 1188–1214.

Greenwood, J., Seshadri, A., Yorukoglu, M., 2005. Engines of liberation. Rev. Econ. Stud. 72, 109–123.

Gronau, R., 1997. The theory of home production: the past ten years. J. Labor Econ. 15, 197–205.

Guryan, J., Hurst, E., Kearney, M., 2008. Parental education and parental time with children. J. Econ. Perspect. 22, 23–46.

Haider, S., Stephens, M., 2007. Is there a retirement consumption puzzle? Evidence using subjective retirement expectations. Rev. Econ. Stat. 89, 247–264.

Hall, R.E., 1968. Technical change and capital from the point of view of the dual. Rev. Econ. Stud. 35, 35–46.

Hammermesh, D., Frazis, H., Stewart, J., 2005. Data watch: the American time use survey. J. Econ. Perspect. 19, 221–232.

Hurd, M., Rohwedder, S., 2006. Some answers to the retirement-consumption puzzle. NBER Working Papers 13929.

Hurst, E., 2008. The retirement of a consumption puzzle. NBER Working Papers 13789.

Juster, F.T., 1985. Preference for work and leisure. In: Juster, F.T., Stafford, F. (Eds.), Time, Goods, and Well-Being. University of Michigan Press, Ann Arbor.

Juster, F.T., Stafford, F. (Eds.), 1985. Time, Goods and Well-Being. University of Michigan Press, Ann Arbor.

Krueger, A., Mueller, A., 2010. Job search and unemployment insurance: new evidence from time use data. J. Public Econ. 94, 298–307.

Kydland, F., Prescott, E., 1982. Time to build and aggregate fluctuations. Econometrica 50, 1345–1371.

Mincer, J., 1962. Labor force participation of married women: a study of labor supply. In: Aspects of Labor Economics. (Eds.), Universities-National Bureau Committee for Economic Research, Princeton, NJ.

Ramey, V., 2007. How much has leisure really increased since 1965? University of California, San Diego Working Paper.

Ramey, V., 2009. Time spent in home production in the 20th century United States: new estimates from old data. J. Econ. Hist. 69, 1–47.

Ramey, V., Francis, N., 2009. A century of work and leisure. Am. Econ. J. Macroecon. 1, 189–224.

Ramey, G., Ramey, V., 2010. The rug rat race. Brookings Pap. Econ. Act. 41 (1), 129–176.

Robinson, J., Godbey, G., 1999. Time for Life: The Surprising Ways Americans Use Their Time. The Pennsylvania State University Press, University Park, Pennsylvania.

Rupert, P., Rogerson, R., Wright, R., 1995. Estimating substitution elasticities in household production models. Econ. Theory 6, 179–193.

Schwerdt, G., 2005. Why does consumption fall at retirement? Evidence from Germany. Econ. Lett. 89, 300–305.

Zeldes, S., 1989. Consumption and liquidity constraints: an empirical investigation. J. Polit. Econ. 97, 305–346.

CHAPTER 5

Who Bears the Cost of Recessions? The Role of House Prices and Household Debt

A. Mian[*,‡], **A. Sufi**[†,‡]
[*] Princeton University, Princeton, NJ, United States
[†] University of Chicago Booth School of Business, Chicago, IL, United States
[‡] NBER, Cambridge, MA, United States

Contents

Abstract

This chapter reviews empirical estimates of differential income and consumption growth across individuals during recessions. Most existing studies examine the variation in income and consumption growth across individuals by sorting on ex ante or contemporaneous income or consumption levels. We build on this literature by showing that differential shocks to household net worth coming from elevated household debt and the collapse in house prices play an underappreciated role.

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.005

Using zip codes in the United States as the unit of analysis, we show that the decline in numerous measures of consumption during the Great Recession was much larger in zip codes that experienced a sharp decline in housing net worth. In the years prior to the recession, these same zip codes saw high house price growth, a substantial expansion of debt by homeowners, and high consumption growth. We discuss what models seem most consistent with this striking pattern in the data, and we highlight the increasing body of macroeconomic evidence on the link between household debt and business cycles. Our main conclusion is that housing and household debt should play a larger role in models exploring the importance of household heterogeneity on macroeconomic outcomes and policies.

Keywords

Consumption, Recession, Distribution, Inequality, Housing, Household debt

JEL Codes

E21, E32, D12, D31

1. INTRODUCTION

Severe recessions are characterized by a large decline in household consumption. Consumption in real terms in the United States fell by almost 3% from the second quarter of 2008 to the second quarter of 2009. Consumption fell from 1929 to 1933 of the Great Depression by 18%. From 2008 to 2011, consumption fell by more than 5% in seven countries in the European Union, and by just less than 5% in Ireland and the United Kingdom.

Given the importance of consumption in household welfare, these sharp declines help explain why the study of recessions is a central pursuit of macroeconomics. One approach is to focus on the causes and implications of the aggregate decline in consumption. We believe, however, that an important pursuit of macroeconomic research should be to understand the *distribution* of the consumption decline across individuals. As the title of our chapter suggests, we want to focus on the question: who bears the cost of recessions? More specifically, which households see the largest drop in consumption during economic downturns?

This is an important question for several reasons. First, there has been an ongoing discussion within macroeconomics on the welfare cost of aggregate fluctuations, a debate instigated by the provocative exercise in Lucas (1987). Research since Lucas (1987) has shown that the distribution of income and consumption losses across individuals during recessions is an important factor in whether business cycles have large welfare consequences. For example, both Krebs (2007) and Krusell et al. (2009) use models with heterogeneity across households to argue that the welfare consequences of aggregate fluctuations are an order of magnitude larger than those calculated by Lucas (1987). Understanding both the distribution of consumption losses and their persistence helps reveal how harmful economic downturns are.

Another important reason to study the distribution of consumption growth during recessions is to evaluate the financial system. One of the central roles of the financial system is to efficiently allocate risk. A large body of research has focused on whether the data are consistent with full consumption risk sharing, when an individual consumption is not a direct function of idiosyncratic shocks received by the individual (eg, Cochrane, 1991). A focus on recessions is useful because it helps us evaluate whether risk sharing is present during times of steep declines in aggregate consumption. If it is not, then further analysis of the financial system and government insurance provision is warranted.

There are also important asset pricing implications from examining the distribution of consumption growth during recessions. Recessions tend to be times when asset prices decline. In representative agent consumption-based asset pricing models, a security's payments during recessions (ie, periods when marginal utility of consumption is high) is a central determinant of the value of the security. But as many researchers have noted, financial securities such as corporate equity are held disproportionately by high-income, wealthy individuals. Fluctuations in aggregate consumption may not be as useful in pricing financial assets as the fluctuations in consumption of individuals that tend to hold financial assets (eg, Malloy et al., 2009; Mankiw and Zeldes, 1991). Therefore, a central question in valuing financial assets is whether the consumption of individuals that hold financial assets is more or less cyclical than the rest of the population.

This review is split into three main parts. In the first part, we review the empirical literature on the cyclicality of income and consumption across individuals. We detail the exact time periods studied, data used, and conclusions of each study. Our primary focus is on research examining the cross-sectional differences in income growth and consumption growth across individuals during recessions. But we also cover ancillary empirical studies on consumption risk sharing and the evolution of consumption and income inequality over time. These latter two areas of research are related both from a theoretical and methodological perspective.

It becomes clear in our review of the literature that the role of wealth shocks, and in particular wealth shocks associated with housing, is largely absent. In Sections 3 and 4 of this chapter, we present empirical evidence on the importance of shocks to household net worth in explaining cross-sectional differences across US zip codes in consumption growth during the Great Recession. We begin this section by discussing both the advantages and disadvantages of zip code-level data. As a preview, the main advantage of zip code-level data is the ability to match high-quality administrative data on income, consumption, wealth, and demographics that naturally add up to aggregates used by most macroeconomists. The main disadvantage is that we can only estimate key parameters such as the elasticity of consumption with respect to net worth shocks at a slightly aggregated level.

Using zip code-level data, we show that variation in the decline in net worth coming from the collapse in house prices from 2006 to 2009, what we call the housing net worth shock, is a powerful predictor of consumption growth across zip codes. We utilize zip code-level administrative data on car ownership, new car purchases, and boat ownership, in addition to survey-based measures of the number of individuals living in a housing unit. By all of these measures, zip codes with a more negative housing net worth shock saw substantially lower consumption growth during the Great Recession. We also show that births declined by substantially more in zip codes hit harder by the housing crash.

The housing net worth shock in a zip code can be decomposed into the decline in house prices in the zip code, and the ex ante exposure of the zip code's wealth to a decline in house prices. We find that both matter. Prior to the recession, zip codes with a large exposure to the housing collapse saw a larger increase in house prices, homeowner borrowing, and consumption.

Motivated by these empirical results, we then review models of aggregate economic fluctuations that can best explain the link between house price shocks and the cross-sectional differences in consumption growth. While our empirical results focus mostly on the Great Recession, we also highlight macroeconomic evidence showing a strong link between household debt, house prices, and business cycles across many countries and time periods.

It is important also to note what this review chapter does not cover. Probably the biggest absence is a detailed review of quantitative macroeconomic models with heterogeneity across households (eg, the literature started by Bewley (1977) and Huggett (1993) among others). We cover some of this research as it has an important component, but we do not review it comprehensively. This literature has shaped our thinking in important ways, and our exclusion of this excellent body of research reflects the fact that there is already a must-read review of this literature by Heathcote et al. (2009). We highly recommend reading their review as a complement to this one.

2. WHO BEARS RECESSION RISK? EXISTING RESEARCH

2.1 Categorizing the Literature

As noted in Section 1, the cyclicality of consumption and income across the distribution of households in the economy plays a crucial role in important questions in macroeconomics. A critical input into any model of cross-sectional heterogeneity in risk exposure is a set of basic facts. As Guvenen et al. (2014) put it:

> *What is common to all of these theoretical and quantitative investigations is that they need to rely on empirical studies to first establish the basic facts regarding the cyclical nature of income risk. Unfortunately, apart from a few important exceptions discussed below, there is little empirical work on this question, largely because of data limitation.*

Our goal in this section is to review the empirical evidence on the cyclical nature of both income and consumption risk. While Guvenen et al. (2014) are correct that the evidence is limited, there are a number of important studies that can be used as a launching pad for further research.

There are five dimensions on which the extant literature can be categorized. First, does the study examine the cyclicality of *income* growth or *consumption* growth? Second, what data set is employed? Third, what time period is examined, and more specifically, does the study focus on one recession or a longer time series of cycles? Fourth, on what dimension are households sorted in the cross section when examining income growth or consumption growth across the distribution? And finally, does the study sort households based on ex ante characteristics, contemporaneous placement in the income or consumption distribution, or shocks received during the recession?

It is this last dimension around which we organize the rest of this section. In our view, the ideal empirical setting is one in which households can be sorted on some ex ante characteristic, and then tracked across cycles. More formally, define some period $\tau = 0$ to $\tau = T$ as an aggregate episode such as an expansion or recession. Following Guvenen et al. (2014), the empirical object of interest is:

$$f(H_{-1}^i) \equiv E[y_T^i - y_0^i | H_{-1}^i] \tag{1}$$

where H_{-1}^i is a characteristic of individual or group i measured before the episode in question and y_τ^i is log consumption or log income for individual or group i at τ.[a] The empirical object $f(H_{-1}^i)$ can be estimated in a flexible manner based on the number of groups.

For example, Guvenen et al. (2014) examine four recession periods between 1978 and 2013, and their primary specification uses average income over the five years prior to the recession as H_{-1}^i. They sort individuals into percentile bins based on this measure of H_{-1}^i, and they plot $y_T^i - y_0^i$ during each recession for each bin, where y is a measure of income. Such a plot allows us to see whether individuals with higher ex ante income levels see larger or smaller declines in income growth during recessions. As Guvenen et al. (2014) note, researchers must be cautious in estimating object (1) when H_{-1}^i is income or consumption. It is likely that income and consumption have mean-reverting properties. As a result, an estimation strategy that sorts on ex ante income and looks at subsequent income growth will tend to find negative effects of ex ante income on income growth. For example, one is likely to find that high-income individuals see larger declines in income growth. But in the presence of a mean-reverting income process, such a result would be partially mechanical.

Notice that availability of *panel* data is crucial for such an exercise. It is only possible if one can track the same households over time. We refer to the literature that estimates the object in (1) as sorting on ex ante characteristics.

[a] To minimize notation, y_τ^i represents log average consumption or log average income in the group whenever i is a group instead of an individual.

A related technique exploits panel data but sorts not on ex ante household characteristics but instead on shocks received during the recession. For example, a common assumption in quantitative models of heterogeneity is that some households become unemployed, and the probability of becoming unemployed is higher in recessions (eg, Krusell and Smith, 1999). A natural empirical object of interest in such a model is the decline in consumption among those individuals becoming unemployed during the recession:

$$f(S_T^i) \equiv E[y_T^i - y_0^i | S_T^i] \tag{2}$$

where S_T^i is a shock received during the recession such as unemployment or a decline in wealth. Once again, panel data are required to estimate this object. We refer to the literature that estimates the object in (2) as sorting on *contemporaneous shocks*.

Unfortunately, estimation of the objects in (1) and (2) requires panel data which are not widely available, especially on consumption. As a result, a third technique is to rely on repeated cross sections, where households in each cross section are sorted into percentiles of either the income or consumption distribution. This is common in studies, for example, that rely on income data from the Internal Revenue Service. Letting p be a percentile group of the distribution, these studies typically follow the object y_τ^p across time.

The drawback of this approach is that the evolution of y_τ^p over time depends both on changes in y for households that stay within group p, and changes in the composition of households in group p. Following Perri and Steinberg (2012), the change from any period $\tau = 0$ to $\tau = T$ in group p is:

$$y_T^p - y_0^p = \alpha\left(y_T^{p-stay} - y_0^{p-stay}\right) + (1-\alpha)\left(y_T^{p-in} - y_0^{p-out}\right) \tag{3}$$

The growth in an outcome for the p percentile of the distribution is a weighted average of the growth in the outcome for households that stay in the percentile group $\left(y_T^{p-stay} - y_0^{p-stay}\right)$ and the compositional change in the percentile group $p\left(y_T^{p-in} - y_0^{p-out}\right)$. Notice that the first term of this expression is almost identical to the object of interest in (1) where the sorting variable is the percentile of the distribution ex ante. We refer to research following (3) as following a *repeated cross-section* approach.

One obvious question is how good a proxy for object (1) is object (3). This depends primarily on movements across the distribution during episode being examined. To our knowledge, there is no comprehensive evaluation of this technique in the literature. The closest we could find is Perri and Steinberg (2012), who highlight different patterns in consumption growth depending on whether individuals or percentiles are tracked over time. We will discuss Perri and Steinberg (2012) in more detail below.

A final group of studies we discuss below are those that use empirical moments from data to calibrate quantitative macroeconomic models of household heterogeneity. As mentioned in Section 1, we do not do a comprehensive review of this literature.

But many of these studies contain significant empirical work that is related to the core questions of this review.

2.2 Sorting on Ex Ante Characteristics

The gold standard for evaluating the cross-sectional implications of recessions for earnings is the study by Guvenen et al. (2014). Using a very large data set from the US Social Security Administration, they are able to track the earnings of individuals from 1978 to 2011, which allows analysis of four recessions. Their main data set focuses on US working-age males over this time period. Given the panel structure of the data, they are able to estimate object (1) explicitly by sorting individuals into income bins prior to each recession and expansion.

The exact dates they use for the four recession periods are 1979–83, 1990–92, 2000–02, and 2007–10. They estimate a slight variation of object (1) above in order to avoid problems associated with those with zero earnings. The actual function they estimate is:

$$\log\left(E[y_T^i|H_{-1}^i]\right) - \log\left(E[y_0^i|H_{-1}^i]\right)$$

Figure 13 in their study reveals the central finding with respect to recessions. In all four recessions, there is a positive relation between ex ante income and income growth from the 10th percentile of the distribution to the 70th percentile. For all recessions except for 2000 to 2002, the positive relation continues to the 90th percentile. That is, for the majority of the income distribution and for all recessions, lower income individuals suffer more during recessions as measured by income growth. Notice this effect is probably understated given that mean-reversion would bias the coefficient in the opposite direction.

The pattern is less consistent at the upper tail of the distribution. The two latest recessions look remarkably similar at the very top of the income distribution. Individuals in the top 10% of the ex ante income distribution see the largest decline in income in the entire population. At the 99th percentile, income drops by a stunning 30%. This pattern is not present in the two earlier recessions. In sum, from the 10th percentile to 90th percentile of the ex ante income distribution, lower income individuals see a bigger decline in income during recessions. The results are less definitive above the 90th percentile, with the last two recessions showing the biggest decline among the very rich.

Perri and Steinberg (2012) use panel data from the Panel Study of Income Dynamics (PSID) to study disposable income growth during the 2007 to 2009 recession across the income distribution. They first show that the bottom 20% of the income distribution saw a sharp decline in earnings, falling by 30% relative to the median over the course of the recession. However, they also show that redistribution through taxes and transfer programs helped offset the decline in earnings. Disposable income, after taking into account taxes and transfers, declined the same amount for the rest of the population and households in the bottom 20% of the income distribution.

The above findings are based on comparing income for the bottom 20% of the income distribution in 2006 and 2008. But as mentioned above, a problem with such a methodology is that there may be significant compositional changes in the households in the bottom 20%. The study by Perri and Steinberg (2012) is one of the few that discusses this problem, and compares estimates when taking into account the compositional change. They show that 75% of the individuals in the bottom 20% of the income distribution in 2008 were in the bottom 20% in 2006. They also show that the households that enter the bottom 20% of the income distribution from 2006 to 2008 see a dramatic 53.4% decline in disposable income. Households that stayed in the bottom 20% saw a decline of 2% in disposable income. Households that left the bottom 20% from 2006 to 2008 saw a 110% rise in disposable income.

Heathcote and Perri (2015) utilize data from the Consumer Expenditure Survey (CEX) and the PSID to study the relation between ex ante household wealth and the change in consumption rates, or the consumption to income ratio, during the recession of 2007–09. The primary focus is on the PSID, and they sort households in year t based on the ratio of wealth to average consumption in years t and $t + 2$. Because the denominator includes future consumption, the analysis does not strictly sort on ex ante characteristics. Nonetheless, the spirit of the exercise is to sort on individuals on their wealth to consumption ratio prior to the recession.

Heathcote and Perri (2015) find a larger drop in the consumption rate for poor households from 2006 to 2008. The consumption rate out of income drops by almost 10 percentage points for the lower half of the wealth distribution, and only 4 percentage points for the upper half. The authors interpret the larger drop in consumption rates of the poor as evidence consistent with a larger rise in a precautionary savings motive. They conduct tests showing that the wealth shock from 2006 to 2008 was larger for rich households, and income prospects deteriorated equally for the rich and poor. Both of these factors strengthen the conclusion that the larger drop in consumption rates for poor households was due to precautionary savings.

Another strand of research focuses on variation in the cyclicality of consumption growth across households based on whether the household holds financial assets. The motivation behind this research is a canonical model of consumption-based asset pricing, where the key determinant of asset prices is consumption risk. Isolating who bears risk during recessions is not the central question of this literature. Nonetheless, it offers insight by showing how consumption growth is correlated with stock returns across households that own and do not own stocks.

Mankiw and Zeldes (1991) use the PSID from 1970 to 1984 and sort households based on whether they report a positive market value of stock holdings in 1984. Unfortunately, the PSID first asked this question in 1984, and so the authors must use an ex post measure of stock holding rather than an ex ante measure. They find that stockholders' consumption is more volatile and more highly correlated with the stock market than households that do not hold stock. They argue that this higher correlation "may be crucial to an ultimate resolution of this puzzle and other asset pricing anomalies." Households that hold stocks tend to have higher income and higher wealth than those that do not; therefore, this finding suggests that the cyclicality of consumption is highest for richer households.

A more recent paper by Malloy et al. (2009) uses the CEX from 1982 to 2004 to test whether long-run consumption growth of households who hold financial assets is more sensitive to asset price fluctuations. The CEX is typically used in the literature as a repeated cross section of respondents. However, there is a panel element because households are surveyed for four consecutive quarters. Malloy et al. (2009) utilize the panel dimension by calculating consumption growth for a group of households from period t to $t + 1$ as the average over each household's consumption growth from t to $t + 1$. The resulting group-level consumption growth rates have both a panel and repeated cross-sectional dimension, because households in the sample leave after four quarters.

Using consumption growth of stockholders vs nonstockholders, the authors first show that stockholder consumption growth has a higher sensitivity (or "beta") to aggregate consumption growth, especially at long horizons. They conclude based on this finding that "stockholders bear a disproportionate amount of aggregate consumption risk relative to nonstockholders and this burden increases in the long run …." Further, the authors find that the consumption growth of stockholders is more correlated with asset returns. Malloy et al. (2009) estimate Euler equations using asset returns and the consumption growth of stockholders, and they find that estimated risk aversion is much lower compared to using aggregate consumption growth.

2.3 Repeated Cross-Section Approach

Only a few of the existing data sets on income and consumption cover a panel of individuals that one can track over an extended period of time. As a result, many researchers use a "percentile-sorting" methodology, as described in Section 2.1. A classic example of this technique is the series of studies by Piketty, Saez, and Zucman using IRS tax returns to measure the evolution of income and wealth inequality in the United States (Piketty and Saez, 2003, 2006; Piketty and Zucman, 2014; Saez, 2015; Saez and Zucman, 2014). The primary focus of these studies is long-run trends in income and wealth inequality, not isolating who bears the cost of recessions.

However, Saez (2015) uses the same data to try to address the cyclicality of income across different groups. For example, he shows that average real income growth from 1993 to 2013 was 15.1%, and 62.4% for the top 1% of the distribution. Recall that this latter figure takes the income of the top 1% in 2013 and compares it to the income of the top 1% in 1993—it is not based on the same individuals. Saez then shows that the income of the top 1% falls considerably more in recessions, and increases significantly more during expansions. For example, the top 1% saw a decline in income of 31% and 36% during the 2001 recession and the Great Recession, respectively. The corresponding growth rates for average income are −12% and −17%. These figures include capital gains income, but the pattern is also present excluding capital gains, albeit less pronounced. This finding using repeated cross-sectional analysis confirms the panel data analysis of earnings in Guvenen et al. (2014): income fell the most for very high income individuals during the 2001 and the 2007–09 recessions.

Parker and Vissing-Jorgensen (2010) use the IRS data to show that more cyclical income growth of high income individuals is a recent phenomenon. They measure income as real pre-tax, pre-transfer income excluding capital gains. They examine the cross-sectional variation in income growth across the income distribution during recessions and expansions, and they find that the top 1% of the income distributions saw sharp declines during the last three recessions. However, the five recessions prior to the last three did now show this pattern. The authors also calculate an income "beta," which comes from the estimation of the following equation:

$$\Delta \ln Y_{i,t+1} = \alpha_i + \beta_i \Delta \ln Y_{t+1} + \epsilon_{i,t+1}$$

This specification tells us whether income of group i loads more heavily on changes in aggregate income. Parker and Vissing-Jorgensen (2010) show that the top 1% of the income distribution has a much higher β from 1982 to 2008 than in previous years. Further, the higher cyclicality of income of the very rich is robust to alternative measures of income from the Census.

Parker and Vissing-Jorgensen (2010) also examine the cyclicality of consumption across the ex ante expenditure distribution. They utilize the CEX and a methodology that is similar to Malloy et al. (2009). More specifically, they first sort households into groups based on expenditure level in quarter q, and they calculate the quarterly consumption growth for the group as the average of the quarterly growth rates of the households within the group. Recall that the CEX allows for such a calculation because households are surveyed for four consecutive quarters. They then calculate annual consumption growth for each group as the sum of the four quarterly growth rate figures. In this manner, the consumption growth measure has both a repeated cross section and panel dimension.

The authors estimate a similar equation as the income specification above to find a consumption β of each group on aggregate consumption and aggregate income. They find that the top 5% of the expenditure distribution has more cyclical consumption than the rest of the population. They find higher cyclicality using a number of measures including aggregate pre- and post-tax income from NIPA or aggregate consumption from NIPA. As with Malloy et al. (2009), the measures of consumption used by Parker and Vissing-Jorgensen (2010) are primarily expenditures on nondurable goods and services. Expenditures on durable goods, for example, are not included.

Meyer and Sullivan (2013a) focus on consumption inequality from 2000 to 2011 using the CEX. They convert expenditures into consumption for vehicles using a service flow equivalent, and they exclude housing outlays and spending on education. Their analysis is a pure repeated cross section; they do not take advantage of the panel element of the data. For each year, they sort households into percentiles based on the level of consumption, and they plot the log difference for each group relative to 2000. They find more cyclical consumption in the high consumption groups. For example, for the 90th percentile, consumption increased by 20% from 2000 to 2007 and then subsequently fell by 6% from 2007 to 2009. In contrast, consumption at the median increased by 16% from 2000 to 2007 before dropping by 4% from 2007 to 2009.

Most of the extant research on consumption growth variation across households relies on either the CEX or the PSID. Cynamon and Fazzari (2014) is an exception. They focus on consumption of the bottom 95% and top 5% of the income distribution, and they track these two groups over time. They estimate consumption by each of these two groups by estimating income and saving rates, and using the difference as the consumption rate. Their methodology takes aggregate savings and uses microeconomic estimates of savings rates to distribute savings to each of the two groups. Similarly, they distribute income to the two groups based on information from the Congressional Budget Office and the Piketty and Saez IRS data sets. By construction, total income, saving, and consumption add up to the aggregates from NIPA.

Cynamon and Fazzari (2014) show that the consumption to income ratio of the bottom 95% of the income distribution fell sharply during the Great Recession from 92% in 2007 to 87% in 2010. The consumption to income ratio rose sharply for the top 5% of the income distribution, which the authors argue is evidence that the top 5% smoothed consumption (income fell but consumption remained constant). The consumption to income ratio of the top 5% also increased substantially during the 2001 recession. Looking at the level of consumption, the authors show that consumption during the Great Recession deviated sharply from trend for both groups, with the magnitude of the decline being slightly larger for the bottom 95% of the income distribution.

2.4 Sorting on Shocks Received in the Recession

Unemployment increases sharply in recessions. For example, Davis and von Wachter (2011) show that the quarterly layoff rate rises by 129 basis points from 1990Q2 to 1991Q2, 85 basis points from 2000Q2 to 2001Q4, and 208 basis points from 2007Q3 to 2009Q1. Rather than sorting on ex ante characteristics, Davis and von Wachter (2011) sort individuals based on exposure to a mass layoff wave during recessions.

More specifically, the authors regard a worker as "displaced" in year y if he separates from his employer in y and the employer experiences a mass layoff in y. A mass layoff event is one where the employer meets the following criteria: 50 or more employees in year $y-2$, a contraction of employment between 30% and 99% from $y-2$ to y, employment in $y-2$ is no more than 130% of employment in year $y-3$, and employment in $y+1$ is less than 90% of employment in $y-2$. They utilize longitudinal SSA records from 1974 to 2008 to measure earnings, which is the same data used by Guvenen et al. (2014). Given the sample period, there are three main recession periods they analyze: the early 1980s, the early 1990s, and the early 2000s.

The central finding is that individuals losing a job during a mass layoff in a high unemployment environment (greater than 8%) lose 2.8 years of predisplacement earnings in present value terms. The loss is 1.4 year of predisplacement earnings if an individual loses a job in a mass layoff event when the unemployment rate is below 6%. The large loss of income when losing a job during a high unemployment rate environment is supported by a number of other studies. For example, Jacobson et al. (1993) show that job displacement in Pennsylvania in the early 1980s led on average to a near-term earnings loss of more than 50%. Losses persist for at least 10 years (Sullivan and Von Wachter, 2009). Topel (1991) finds that workers displaced from 1979 to 1984 who can find a new job face a 14% reduction in earnings. Davis and von Wachter (2011) have data only through 2008 and are therefore unable to measure the long-term consequences of the Great Recession. However, as they note, the existing research "suggests that workers who have experienced job displacement events since 2008 are likely to suffer severe and persistent earnings losses."

A related line of research examines the effects of graduating from college during a recession. Kahn (2010) uses data from the National Longitudinal Survey of Youth on students that graduated from college between 1979 and 1989. She uses both variation in the national unemployment rate and the state unemployment rates to identify the effect of graduation in a weak economy on wages. She finds "large, negative wage effects of graduating in a worse economy which persist for the entire period studied." More specifically, she finds an initial wage loss of 6–7% for a 1 percentage point increase in the unemployment rate. The wage loss dissipates over time, but wages remain 2.5% lower even 15 years after graduation. A related study by Oreopoulos et al. (2012) examines Canadian data. They find that a rise in unemployment rates by 5 percentage point implies an initial loss in earnings of about 9% that halves within 5 years and finally fades to 0 by 10 years.

Recessions are also times when there are substantial shocks to both housing and financial wealth. Such shocks typically have a strong cross-sectional component, given substantial cross-sectional variation across households in exposure to such shocks. Mian et al. (2013) exploit cross-sectional variation across US geographic units—counties or zip codes—in the exposure to housing net worth shocks during the Great Recession. The authors define the housing net worth as the decline in household net worth coming from the collapse in house prices. The variation across the country is large: the housing net worth shock was almost -50% in the bottom decile of county distribution, but

0% in the top. They show that the decline in consumer spending was larger in counties with a more negative housing net worth shock. Using zip code-level data on auto purchases, they also show that the marginal propensity to spend out of housing wealth is substantially larger for lower income and higher household leverage zip codes. Kaplan et al. (2015) and Stroebel and Vavra (2014) use a different data source and find results similar to Mian et al. (2013); consumption growth during the Great Recession is strongly correlated with house price growth across US cities. In the analysis below, we show a strong state-level correlation between house price growth and personal consumption growth during the Great Recession using new data from the Bureau of Economic Analysis.

Mian and Sufi (2010) sort counties by the change in the household debt to income ratio from 2002 to 2006, and then examine how the decline in new auto purchases and residential investment during the Great Recession is related to the previous increase in household debt. They find a negative relation: counties with large increases in the household debt to income ratio from 2002 to 2006 saw the largest decline in new auto purchases and residential investment during the Great Recession. Mian and Sufi (2010) also show that there is a strong relation between house price growth from 2006 to 2009 and the previous increase in household debt. As a result, while Mian and Sufi (2010) technically sort on an ex ante variable, it is best to view both the increase in ex ante household debt and ex post house price decline as reflecting similar underlying shocks. Bunn and Rostom (2014) use microeconomic data on British households and find that individuals with higher debt had lower subsequent consumption growth after 2007. Andersen et al. (2014) use individual data on Danish households and find a strong negative correlation between precrisis leverage and the change in nonhousing consumption during the crisis. Baker (2014) uses data from an online financial services firm and finds that highly indebted households in the United States during the Great Recession displayed a larger elasticity of consumption with respect to negative income shocks. All of these studies imply a close connection between consumption growth during recession and household balance sheets.

One concern with sorting individuals based on a shock received in the recession is that the decline in income of consumption could be related an omitted variable driving both the shock and the outcome of interest. For example, in Davis and von Wachter (2011), one worry is that the individuals laid off during a recession are lower quality workers, which partially explains the high earnings displacement. Or in Mian et al. (2013), the worry is that some omitted variable drives both the collapse in house prices and the collapse in consumption in a given county. In the section below, we will extend the results in Mian et al. (2013) and discuss why we believe this is unlikely to be the case. The studies examining college graduation are less exposed to this criticism given that the timing of college graduation is less likely to be driven by an omitted variable correlated with a recession occurring.

2.5 Results from Quantitative Models

As mentioned in Section 1, we focus in this review chapter on studies focusing on the cross-sectional variation across individuals in consumption growth and income growth during recessions. There is a large body of research employing quantitative models and calibration to assess the importance of business cycle fluctuations, and the review chapter by Heathcote et al. (2009) covers these studies in detail. We did, however, want to highlight some of the empirical findings from this literature, as they are related to the core question of who bears the cost of recessions.

Storesletten et al. (2001, 2004) use the PSID to make two main points. First, they argue that innovations to the idiosyncratic component of an individual's income process is highly persistent. And second, they argue that idiosyncratic earnings risk is counter-cyclical. They support these arguments with a number of results. For example, they show that the cross-sectional standard deviation of earnings increases substantially during recessions just as the cross-sectional mean of earnings falls. Further, they show that an age cohort of individuals who have lived through more contractionary periods have higher cross-sectional dispersion in earnings even as they age.[b] The authors use these facts to motivate a quantitative model where the welfare losses associated with business cycle fluctuations are substantially larger than those implied by Lucas (1987).

2.6 Summarizing the Literature

A few points emerge when looking at the research as a whole.

First, there remains a need for more panel data, especially when it comes to consumption. Of the studies reviewed, only two sort on ex ante characteristics, and then track consumption through a recession for the same units. One of these is based on county-level data, not individual-level data. On this same point, we need more research detailing whether sorts on contemporaneous placement within the distribution biases results in a meaningful way. This related to the discussion in Section 2.1 about people moving into and out of groups.

Second, the results on consumption growth in recessions are mixed. Researchers using the CEX tend to find that consumption of the rich is more cyclical and falls more in the Great Recession. Researchers using the PSID find that the poor see substantial declines in consumption, and one study shows that the decline in the consumption rate is much larger for the poor than the rich. It is difficult to reconcile the different findings because different data are used and different points in the distribution are analyzed. It may be that the perfect consumption data (the analogous data of the SSA on income)

[b] Guvenen et al. (2014) use SSA data to argue that variance of idiosyncratic earnings shocks is not counter-cyclical. Instead, it is the left skewness of shocks that is strongly counter-cyclical.

would show the same nonmonotonicity shown in Guvenen et al. (2014) where the very rich see the largest decline in consumption, but the moderate rich see less of a decline than the poor.

Third, researchers that sort on shocks received in the recession find long-lasting effects. This is especially true when sorting on whether one loses a job in a recession.

2.7 Related Areas of Research

There are two areas of research that are related to the core question of this chapter: consumption risk sharing and consumption inequality. We do not present a comprehensive review of these areas here, but we want to mention a few studies that are related to the issue of who bears recession risk. Full consumption risk sharing is the idea that an individual's consumption is insured against idiosyncratic shocks. Cochrane (1991) is one of the earliest contributions. He uses the PSID to test whether idiosyncratic shocks such as illness, unemployment, or forced moving affect consumption. He finds that involuntary job loss in particular has a large effect on consumption growth. However, given a limited sample, he concludes that "many of the variables examined here do not yield a robust rejection of the theory."

Attanasio and Davis (1996) examine how consumption responds across groups in response to changes in the hourly wage structure of US workers in the 1980s. They note the "extreme scarcity of longitudinal data sources with high-quality information on both earnings and consumption …." Given this problem, they instead construct synthetic panels of individuals based on earnings information from the Current Population Survey and consumption data from CEX. They form groups of men based on age and education, and they examine how relative movements in wages for each group affects consumption growth during the 1980s. Their core finding, best illustrated in their Fig. 2, shows a strong relation between relative wage movements and consumption growth. They conclude that their findings represent "a spectacular failure of between-group consumption insurance, a failure not explained by existing theories of informationally constrained optimal consumption behavior."

Schulhofer-Wohl (2011) argues that it is crucial to take into account heterogeneity in risk preferences when conducting tests of consumption risk sharing. For example, he shows that if less risk-averse households have more procyclical incomes (as would be expected if individuals sort into occupations based in part on risk tolerance), standard tests of consumption risk sharing will tend to reject full risk sharing even if it is present. He uses the PSID to show that accounting for such heterogeneity leads to a failure to reject full consumption risk sharing.

Another related area of research is on the evolution of consumption inequality over the past 50 years in the United States. There is an enormous literature on this subject, which includes contributions by Aguiar and Bils (2015), Attanasio et al. (2004, 2012), Heathcote et al. (2010), Krueger and Perri (2006), Slesnick (2001), and Meyer and Sullivan (2013b). For the purposes of this chapter, we want to highlight the data used and controversy regarding whether there in fact has been an increase in consumption inequality. The earlier studies in this literature found that consumption inequality measured using the CEX does not track the increase in income inequality. However, later studies argued that this may be due to reporting biases associated with the CEX. Attanasio et al. (2012) in particular make adjustments in the use of the CEX and look also at evidence from the PSID to find that consumption inequality did in fact rise over the past 30 years.

This debate was particularly fruitful because it helped researchers understand better the advantages and disadvantages of the CEX, which is the main data set on consumption used in the literature. Our approach in our own research has been to rely on administrative data collected by private companies. We will return to some of the issues associated with the CEX in the next section.

3. ZIP CODE-LEVEL CONSUMPTION MEASURES

3.1 Toward Administrative Measures of Consumption

Following the discussion in Section 2.1, a key goal of empirical research in macroeconomic fluctuations is to estimate loadings on ex ante factors that predict the decline in consumption across individuals during economic downturns. Substantial progress has been made on the factors that predict a decline in *income* during recessions, in large part due to advances in the administrative data on earnings that have become recently available. But less progress has been made on *consumption*. The key limitation is the lack of individual-level panel data that very accurately measure consumption.

As noted above, researchers have primarily used data from two surveys: the CEX and the PSID. However, these two data sets have important limitations. First, neither data set is an ideal panel. The CEX data set only tracks the same individuals for four quarters, and the PSID is only conducted once every two years. This makes a comprehensive analysis of consumption growth during a recession difficult.

Second, both data sets are based on surveys rather than administrative data. The CEX in particular has been criticized as a method for studying cross-sectional variation in consumption across individuals (Attanasio et al., 2004; Cantor et al., 2011). Studies have outlined many problems with the CEX, such as underreporting by high-income households and a low response rate. Further, according to Cantor et al. (2011), these problems are becoming worse over time.[c] More generally, there is an extensive body of research

[c] The debate on the quality of CEX data is not settled. For example, Bee et al. (2012) argue that the CEX performs well once adjustments are taken into account. Attanasio et al. (2012) argue that adjustments can be done to the CEX which makes cross-sectional comparisons across individuals more accurate. We do not wish to wade too deep into this debate; our goal is to point out that alternative measures of consumption can help.

showing nonclassical error in measures of consumption using survey data.[d] One of the most striking examples from the literature comes from Sweden. Koijen et al. (2014) measure actual car purchases from registry data vs responses to a survey. They find underreporting in the survey of 30%—that is, 30% of the households that actually buy a car do not report the purchase to the survey. This underreporting is worse for low income, poor, and older households.

To be clear, we are not saying we should never use survey-based measures of consumption. Adjustments to the measures can be made, or measurement errors may be less relevant for certain questions. However, we believe a promising avenue is to follow the income literature which increasingly relies on administrative data, such as the Social Security Administration data used in Guvenen et al. (2014).

Some progress on this front has been made very recently. For example, Baker (2014) uses data from a large online personal finance website that connects users' financial accounts. Because the website has bank and credit card accounts, Baker (2014) can measure consumption using administrative data on transactions and withdrawals. Using these data, he finds that highly indebted households are more sensitive to income fluctuations. In particular, a one-standard deviation increase in the debt to asset ratio increases the elasticity of consumption by approximately 25%. Two studies use administrative data from personal finance websites to study the effect of the 2013 government shutdown on consumption and borrowing (Baker and Yannelis, 2015; Gelman et al., 2015).

3.2 Zip Code-Level Data on Consumption

The approach we have used in our own research is to use administrative data that are aggregated by some geographical unit (Mian et al. (2013)). We have utilized two measures: new auto sales at the zip code level from R.L. Polk, and purchases from debit card and credit card transactions at the county level from MasterCard Advisers. The county identified in the latter data set reflects the county of the store where items are purchased, not the county of residence of the buyer. Both of these data sets are based on actual transactions as opposed to survey responses. The R.L. Polk data are based on the universe of new vehicle registrations. The MasterCard data are limited to transactions where MasterCard is the servicer, but Mian et al. (2013) show that aggregate spending using the MasterCard data closely tracks aggregate spending from Census retail sales.

In the work that follows we introduce three new measures of consumption. The first is also from R.L. Polk, and reflects the total number of vehicles registered to individuals living in a zip code. We have the total number of vehicles broken down by model year, which we use to depreciate the older vehicles using average used car prices reported in Jacobsen and van Benthem (2015). The final zip code-level variable is the total number of vehicles in units of the current year (where older vehicles are depreciated before adding up to the total).

[d] See for example the cites in Koijen et al. (2014).

The second measure of consumption is registered recreational boats. The boat data come from Merchant Vessels of the United States, a data file of merchant and recreational vessels documented by the US Coast Guard. Code of Federal Regulations (2001) requires any vessel of at least five net tons which engages in the fisheries on the navigable waters of the United States or in the Exclusive Economic Zone, or coastwise trade to be registered each year, with some minor exemptions. Most vessels longer than 25 ft measure five net tons.

The data contain variables for the general service type of the vessel and the registration status. We take the number of all boats with recreational service type and valid registration to calculate the amount of boat consumption. Roughly 70% of vessels with valid registrations are recreational vessels in the data. We use the postal code of the managing owner's address to allocate boats to zip codes.

The third measure of consumption comes from the American Community Survey. The ACS provides zip code-level statistics for every five-year wave of surveys conducted. We use the 2005–09 wave and the 2008–12 wave. Two of the five years overlap (2008 and 2009), which will mechanically bias us away from finding large changes in a zip code. We refer to each wave by their midpoint year: the earlier wave is the 2007 wave and the later wave is the 2010 wave. The specific measure of consumption we use from this survey is the number of individuals over 16 per housing unit. We calculate the change in the number of adults living per housing unit from 2007 to 2010 as a measure of per capita consumption of housing services.

There is an additional measure we use that is better interpreted as a measure of welfare than consumption: the number of child births. We have this data set only for the state of California, which collects information on the number of births for residents of zip codes.

3.3 Advantages and Disadvantages of Zip Code-Level Data

The main advantage of administrative zip code-level data on consumption is that it is measured very well and therefore closely tracks aggregates. For example, total new purchases of vehicles according to R.L. Polk almost perfectly match total purchases according to Census retail sales. In an ideal world, we would be able to decompose aggregate macroeconomic consumption data into each individual. The zip code-level administrative data sets are the closest we have at this point to this ideal.[e]

Relative to the ideal individual-level panel data set, what are the disadvantages of using data at the zip code level? The first major disadvantage is that any aggregation procedure smooths differences across the population. If we are interested in how differences in ex ante factors or ex post shocks affect consumption growth during recessions, we must have sufficient variation across zip codes to estimate parameters. The extent of this

[e] Baker (2014) shows that data he employs from a personal finance website closely track aggregates once weights are taken into account to adjust for differences in the characteristics of individuals who use the website.

problem depends crucially on how people sort across zip codes. If, for example, there is a large degree of heterogeneity across individuals but no heterogeneity across zip code averages, then analysis using zip code-level data has no variation to exploit. This would be the case if people randomly sorted across zip codes.

How big of a problem is this in the United States? As of 2000, there were 220 million individuals living in the United States over the age of 16, and approximately 31,000 zip codes, which gives an average number of adults per zip code of about 7000. However, the zip code-level population is heavily skewed to the right. The median number of adults per zip codes is 2225 and there are 21,732 at the 90th percentile. There are 43,377 at the 99th percentile.

How much variation do we lose by using zip code-level data? To answer this question, we need data for some variable at both the individual level and zip code level. One important sorting variable in the literature is income. IRS data are available at both the individual level (in the public use file) and the zip code level. Fig. 1 plots the distribution of the individual- and zip code-level data. For the zip code-level data, we calculate the adjusted gross income per return, and then we look at points in the distribution when weighting by the total number of returns. As Fig. 1 shows, the zip code-level distribution is smoothed, especially at the tails of the distribution. The 99th percentile in the individual-level data is almost $400,000. The 99th percentile zip code has an average AGI of $250,000. But even with this smoothing, there is substantial variation in average income across zip codes.

The second potential problem is movers. This is the same criticism that applies to repeated cross-sectional analysis based on percentiles already discussed in Section 2.1, Eq. (3). Over long periods, people moving across zip codes present a serious problem. It is not obvious how large a problem moving is when studying a two- to three-year window such as the Great Recession.

We are unaware of estimates that document the likelihood of people moving across zip codes during a two- to three-year period. The Census provides information on annual mover rates, which averaged 12.5% from 2006 to 2009. However, most of these movers stayed within the same county. From 2006 to 2009, annual moving rate to another county was 3.9%.

We do not yet have sufficient evidence to assess how moving across zip codes affects the cross-sectional variation of consumption growth using zip code-level data. For example, if movers tend to move to other zip codes that are similar based on the sorting variable in question, such movement may not be a concern.[f]

[f] The closest study we could find to answering this question is Yagan (2014) who examines whether workers bore the incidence of local labor demand shocks during the Great Recession even after they move to less affected areas. He finds that even workers that migrate to less impacted areas see "unusually small employment gains." This does not answer the key question we have in mind: do individuals that move end up in zip codes with a similar shock as the one they left.

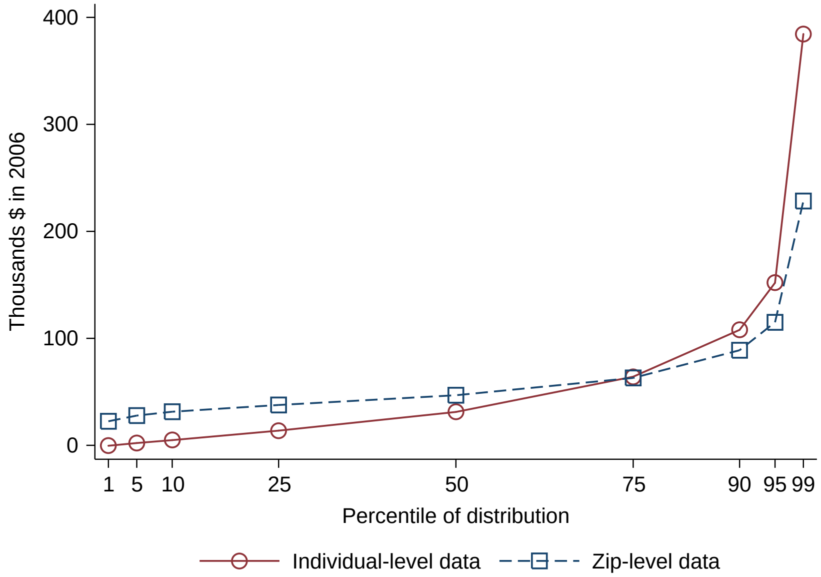


Fig. 1 Zip code vs individual data: distribution of 2006 income. This figure compares zip code-level and individual-level data on 2006 adjusted gross income from the IRS. For the zip code-level distribution, we calculate the AGI per return for each zip code, and then we look at points on the distribution when weighting by the total number of returns in the zip code. As the figure shows, zip code-level data smooth the distribution, and the smoothing is substantial at the tails of the distribution.

4. HOUSING NET WORTH SHOCK AND THE GREAT RECESSION

In this section, we review and build on evidence on the importance of shocks to net worth coming from the collapse in house prices during the Great Recession. In particular, Mian et al. (2013) show that variation across US counties in the housing net worth shock—or the percentage decline in household net worth coming from the collapse in house prices—is correlated strongly with consumption growth. Using new measures of consumption at the zip code level, we build on this study and discuss possible interpretations of the results.

4.1 Housing Net Worth Shock: Definition

Following Mian et al. (2013), we define the housing net worth shock in a zip code as:

$$\text{HNW shock}_z \equiv \frac{p_{z,2009} - p_{z,2006}}{p_{z,2006}} * \frac{H_{z,2006}}{F_{z,2006} + H_{z,2006} - D_{z,2006}} \tag{4}$$

where $p_{z,t}$ is the median price of an owner-occupied unit in zip code z at time t, $H_{z,2006}$ is the value of housing assets owned by residents in zip code z in 2006, $F_{z,2006}$ is the value of financial assets held by residents in zip code z in 2006, and $D_{z,2006}$ is the book value of debt outstanding for residents of zip code z in 2006. This definition follows from the decomposition of the percentage change in net worth from 2006 to 2009, where we isolate the percentage change in net worth coming from the collapse in house prices.[g]

As definition (4) illustrates, the cross-sectional variation across zip codes in the housing net worth shock during the Great Recession is driven by two factors: house price growth in zip code z during the Great Recession and the ex ante housing wealth to net worth ratio. The latter captures the effect of leverage. All else equal, zip codes that have more leverage have a higher housing wealth to net worth ratio. This amplifies the effect of house price growth on total net worth. Throughout, we define the housing wealth to total net worth ratio as the second term:

$$\text{Housing wealth to net worth ratio}_z \equiv \frac{H_{z,2006}}{F_{z,2006} + H_{z,2006} - D_{z,2006}} \tag{5}$$

As we show below, the housing wealth to net worth ratio is a strong predictor of consumption growth across zip codes during the Great Recession, with high housing wealth to net worth ratio zip codes seeing a larger decline in consumption.

The main sample restriction we must make is due to the fact that not all zip codes have accurate house price indices available. In particular, we use CoreLogic data on house prices, which cover approximately 6600 of the 31,000 zip codes in the United States. These zip codes account for 65% of the adult population of the United States as of 2000. The main difference between zip codes with and without house price data is population density. The zip codes not in the sample are much more likely to be in rural areas. Rural areas do not have a sufficient number of housing transactions to construct house price indices.

In Fig. 2, we sort zip codes into five quintiles based on the housing net worth shock during the Great Recession. The quintiles are population weighted, so each quintile contains the same number of people. We then plot the housing net worth shock across the distribution. By construction, the housing net worth shock is more negative in the lower quintiles.

Fig. 2 shows a large degree of heterogeneity across zip codes in the United States in the housing net worth shock. In the first quintile, the collapse in house prices reduced household net worth by almost 30%. In the fifth quintile, there is almost no change in household net worth coming from the collapse in house prices.

Table 1 shows summary statistics for the sample of zip codes. The summary statistics weight zip codes by total adult population as of 2000. As mentioned above, the housing net worth shock can be decomposed into house price growth from 2006 to 2009 and the housing wealth to net worth ratio as of 2006. As the summary statistics show, there

[g] The exact methodology used to construct all variables in definition (4) is in Mian et al. (2013).

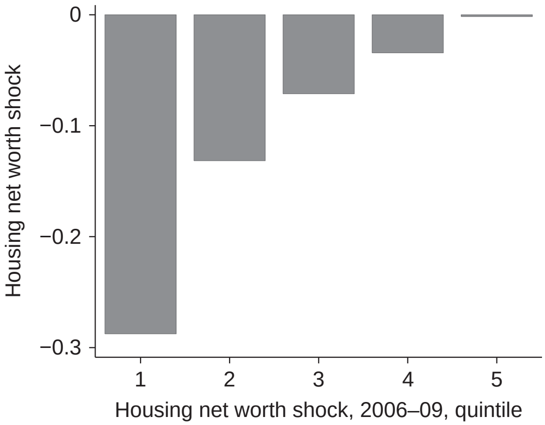


Fig. 2 Housing net worth shock from 2006 to 2009 across zip codes. This figure plots the housing net worth shock across the housing net worth shock distribution. Each quintile contains 20% of the adult population.

Table 1 Summary statistics

	N	Mean	SD	10th	90th
Housing net worth shock					
Housing net worth shock, 2006–09	6689	−0.105	0.111	−0.263	−0.005
Housing wealth to net worth, 2006	6689	0.456	0.218	0.212	0.767
House price growth, 2006–09	6689	−0.200	0.153	−0.430	−0.017
Outcomes					
Growth in autos registered, 2006–10	6689	−0.135	0.084	−0.234	−0.036
Growth in new auto sales, 2006–09	6689	−0.382	0.149	−0.582	−0.199
Change in adults per housing unit, 2007–10	6686	−0.016	0.095	−0.125	0.091
Growth in boats registered, 2006–10	3204	−0.102	0.196	−0.348	0.150
Growth in number of births, 2006–10	855	−0.087	0.106	−0.202	0.038
Fraction of homeowners underwater, 2011	6300	0.337	0.158	0.143	0.565
Ex ante patterns					
House price growth, 2002–06	6689	0.456	0.318	0.090	0.910
Change in cash-out refinancing share	6689	0.022	0.047	−0.025	0.081
Median household income, 2000	6689	48.705	16.657	31.031	70.821
Wage growth, per tax return, 2002–06	6689	0.106	0.068	0.030	0.189
AGI growth, per tax return, 2002–06	6689	0.186	0.124	0.064	0.345

This table presents summary statistics for zip codes in our sample. The sample is restricted to zip codes for which CoreLogic house price growth data are available. The housing net worth shock is the decline in net worth driven by the decline in house prices, or the product of house price growth from 2006 to 2009 and the 2006 housing wealth to net worth ratio. Registered autos reflect the total number of autos registered to residents of a zip code, where autos are depreciated according to their model year. Data on births are available only for California. For registered recreational boats, only zip codes with at least 10 registered boats in 2006 are included.

is substantial variation across zip codes in both measures. The average housing wealth to net worth ratio is 0.46, but it is 0.21 at the 10th percentile and 0.77 at the 90th percentile.

Table 1 also shows summary statistics for measures of consumption growth. There is a sharp decline in both new auto purchases and in the stock of vehicles during the Great Recession. Recall that our measure of registered autos depreciates older cars so that the units are in terms of new cars. The average number of adults per housing units declined by 0.016 from 2007 to 2010. The number of births fell sharply during the Great Recession.[h]

4.2 Housing Net Worth Shock and Consumption Growth

Figs. 3 and 4 show the strong correlation between consumption growth and the housing net worth shock across zip codes. Zip codes with the most negative housing net worth shock from 2006 to 2009 see a 20% decline in registered autos from 2006 to 2010 and a 50% decline in new auto purchases. For zip codes in the top quintile, the respective growth was −10% and −25%.

Similar results are found in Fig. 4. Registered recreational boats fall more in zip codes with the most negative housing net worth shock. The number of adults living per housing unit increases the most in these same zip codes. The number of births also falls more in zip codes with the most negative housing net worth shock.

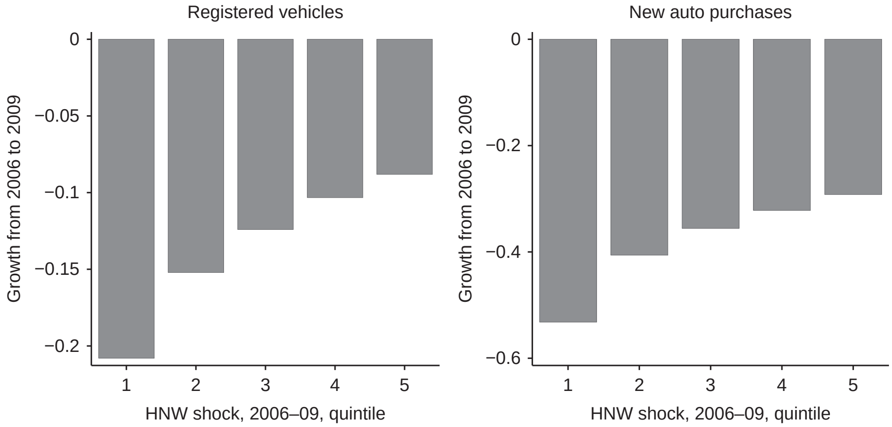


Fig. 3 Housing net worth shock from 2006 to 2009 and consumption growth. This figure plots measures of consumption growth across the housing net worth shock distribution. Each quintile contains 20% of the adult population.

[h] For auto and boat registrations, we use 2006–10 as the time period of examination. We do so because the registration data are not updated immediately, especially when registrations expire. For births, we use 2006 to 2010 under the assumption that parental decisions on births that are made in 2009 are realized in 2010.

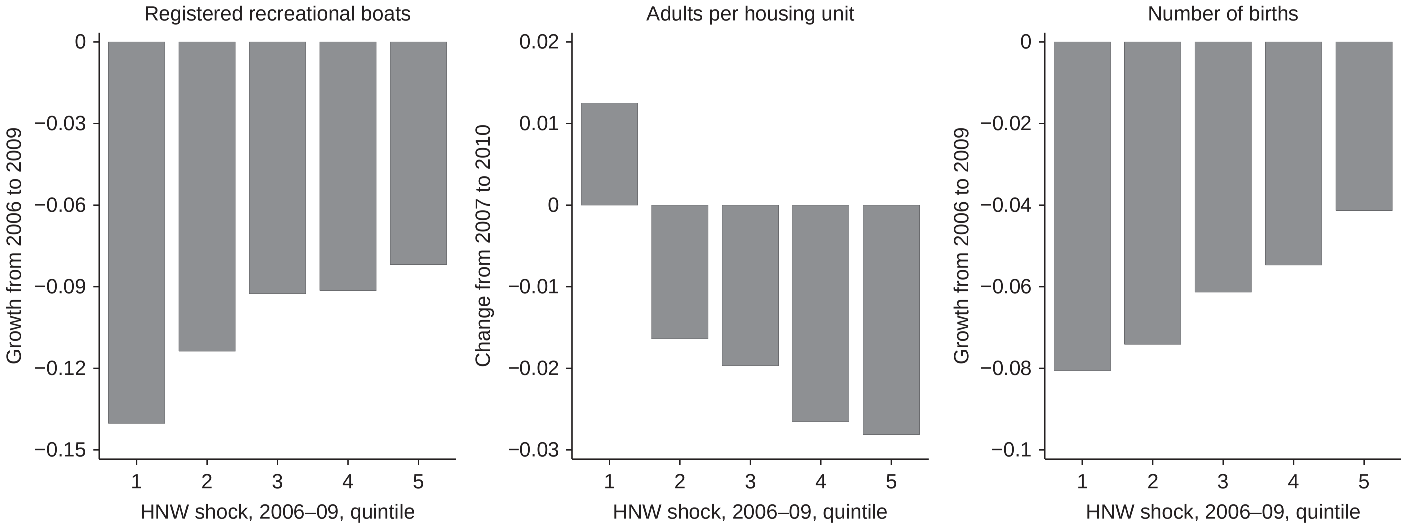


Fig. 4 Housing net worth shock from 2006 to 2009 and other measures of consumption growth. This figure plots recreational boats registered, the number of adults per housing unit, and births across the housing net worth shock distribution. Each quintile contains 20% of the adult population. Births are only available for California, and as a result we resort zip codes into quintiles based on California only for the right panel.

Table 2 presents estimates of the elasticity of consumption growth with respect to the housing net worth shock. Recall that the housing net worth shock is defined as the percentage decline in net worth coming from the collapse in house prices. The estimated elasticities are on the order of 0.2–0.8 depending on the measure of consumption. For births, the estimated elasticity is 0.12.

The magnitude in column 5 requires additional information to interpret. Moving from the 90th to 10th percentile of the distribution of housing net worth shocks is a −0.25 movement. The coefficient estimate in column 5 implies such a move leads to a 0.04 increase in the number of adults living per housing unit. This may sound small, but the variation in the number of adults living per housing unit is also small—the standard deviation in the change in number of adults living per unit is only 0.095, and so this effect is almost 1/2 a standard deviation.

In Table 3, we decompose the housing net worth shock from 2006 to 2009 into its two components: house price growth from 2006 to 2009 and the housing wealth to net worth ratio as of 2006. Panel A presents OLS estimates. The first column shows that the two components are strongly correlated: zip codes with a high housing wealth to net worth ratio in 2006 experienced a larger decline in house prices during the Great Recession. The powerful correlation makes it difficult to separate the independent effects on consumption. The OLS estimates in columns 2 through 6 show that both components matter, especially for registered autos and new autos.

In panel B, we include county fixed effects in the estimation. This is useful because the majority of the variation in house price growth across zip codes is driven by county-level

Table 2 Housing net worth shock from 2006 to 2009 and consumption growth

	(1)	(2)	(3)	(4)	(5)
	Registered auto growth 2006–09	**New auto sales growth 2006–09**	**Registered boat growth 2006–09**	**Adults per unit change 2006–09**	**Births growth 2006–09**
Housing net worth shock, 2006–09	0.381** (0.019)	0.788** (0.030)	0.239** (0.052)	−0.150** (0.036)	0.116** (0.026)
Constant	−0.095** (0.003)	−0.299** (0.006)	−0.080** (0.007)	−0.031** (0.003)	−0.035** (0.009)
Observations	6689	6689	3204	6686	856
R^2	0.250	0.342	0.012	0.031	0.027

This table presents zip code-level regressions relating measures of consumption growth during the Great Recession to the housing net worth shock from 2006 to 2009. The housing net worth shock is the percentage decline in net worth coming from the collapse in house prices. Births are only available for California. Standard errors are clustered by county.
**,* Coefficient statistically different than zero at the 1% and 5% confidence level, respectively.

Table 3 Decomposition of housing net worth shock effect on consumption growth

Panel A: OLS

	(1)	(2)	(3)	(4)	(5)	(6)
	House price growth 2006–09	**Registered auto growth 2006–09**	**New auto sales growth 2006–09**	**Registered boat growth 2006–09**	**Adults per unit change 2006–09**	**Births growth 2006–09**
House wealth to net worth, 2006	−0.297**	−0.034*	−0.147**	−0.019	0.061**	−0.057**
	(0.022)	(0.013)	(0.022)	(0.021)	(0.015)	(0.014)
House price growth, 2006–09		0.274**	0.464**	0.147**	−0.031	0.054
		(0.020)	(0.034)	(0.037)	(0.028)	(0.037)
Constant	−0.064**	−0.065**	−0.222**	−0.064**	−0.050**	−0.007
	(0.011)	(0.004)	(0.008)	(0.010)	(0.006)	(0.016)
Observations	6689	6689	6689	3204	6686	856
R^2	0.179	0.291	0.360	0.013	0.029	0.033

Panel B: County fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)
	House price growth 2006–09	**Registered auto growth 2006–09**	**New auto sales growth 2006–09**	**Registered boat growth 2006–09**	**Adults per unit change 2006–09**	**Births growth 2006–09**
House wealth to net worth, 2006	−0.090**	−0.073**	−0.219**	−0.087**	0.063**	−0.051**
	(0.010)	(0.008)	(0.018)	(0.026)	(0.012)	(0.013)
House price growth, 2006–09		0.093**	0.230**	0.150	−0.052*	0.017
		(0.021)	(0.036)	(0.091)	(0.023)	(0.037)
Observations	6689	6689	6689	3204	6686	856
R^2	0.893	0.556	0.686	0.258	0.443	0.128

This table presents zip code-level regressions relating measures of consumption growth during the Great Recession to the components of the housing net worth shock: the 2006 house wealth to net worth ratio and house price growth from 2006 to 2009. Births are only available for California. Panel B includes county fixed effects. Standard errors are clustered by county.

**,* Coefficient statistically different than zero at the 1% and 5% confidence level, respectively.

variation.[i] For example, the across-county standard deviation in house price growth from 2006 to 2009 is 0.13, whereas the within-county standard deviation is only 0.05. This is a robust feature of house price variation in the United States—the majority of the variation is across county (or across city) as opposed to within county. In contrast, more of the variation in the housing wealth to net worth ratio as of 2006 is driven by within-county differences. The within-county standard deviation in the housing wealth to net worth ratio is 0.16, whereas the across county standard deviation is 0.12.

In an ideal setting, we could test whether the ex ante housing wealth to net worth ratio as of 2006 in a zip code predicts a stronger decline in consumption, given an identical decline in zip code-level house prices. In other words, we could hold the asset price shock constant and estimate whether larger exposure to the shock has a strong effect on consumption growth.

Unfortunately, even with county fixed effects, house price growth and the housing wealth to net worth ratio are correlated. As column 1 of panel B shows, the regression coefficient remains negative and statistically significant when including county fixed effects, but the coefficient is smaller with a smaller *t* statistic. In columns 2 through 6 we include county fixed effects, and the results indicate a much stronger loading of consumption growth on the housing wealth to total net worth ratio as of 2006. That is, controlling for house price growth, zip codes with a larger share of net worth in housing see a bigger decline in consumption.

Fig. 5 plots the fraction of homeowners underwater on their mortgage as of 2011 across the 2006 housing wealth to net worth ratio. As it shows, homeowners in zip codes

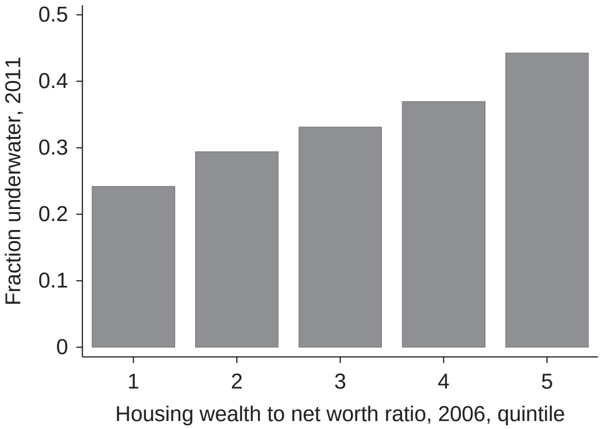


Fig. 5 Housing wealth to net worth ratio and underwater homeowners. This figure plots the fraction of homeowners underwater in 2011 (total mortgage balance greater than the value of the home) across the distribution of the 2006 housing wealth to net worth ratio. Each quintile contains 20% of the population.

[i] There are 1021 counties in our sample, so an average of about 7 zip codes per county.

with a higher housing wealth to net worth ratio as of 2006 were almost twice as likely to find themselves underwater on their mortgage when house prices fell.

4.3 Exploring Zip Codes with High 2006 Housing Wealth to Net Worth Ratio

We have so far kept silent on the underlying economic mechanism connecting the 2006 housing wealth to net worth ratio to consumption growth during the Great Recession. Table 4 begins our discussion. Each cell in Table 4 is a regression coefficient from a separate regression of the 2006 housing wealth to net worth ratio on factors leading up to the Great Recession. The first column presents OLS estimates, and the second column presents county fixed effects specifications.

Zip codes with a higher housing wealth to net worth ratio as of 2006 experienced higher house price growth from 2002 to 2006. The second row uses as a dependent variable the change in the cash-out refinancing share. This variable is defined as the average share of mortgages refinanced with cash taken out (ie, the mortgage balance was increased) from 2003 to 2006 minus the average share from 2000 to 2002. As the coefficient estimates show, households with a high housing wealth to total net worth ratio as of 2006 saw a large increase in the share of mortgages refinanced with cash taken out from 2003 to 2006.

Fig. 6 shows the result over time. We split the sample into five population-weighted quintiles based on the 2006 housing wealth to net worth ratio, and then plot the share of

Table 4 Understanding variation in 2006 housing wealth to net worth ratio

	Housing wealth to net worth ratio, 2006	
	(1)	(2)
House price growth, 2002–06	0.313** (0.041)	0.499** (0.106)
Δ Cash-out refinancing share	2.495** (0.271)	3.894** (0.205)
Wage growth, 2002–06	−0.626** (0.071)	−1.179** (0.090)
AGI growth, 2002–06	−0.825** (0.038)	−1.078** (0.051)
ln(Adjusted gross income, 2002)	−0.244** (0.025)	−0.339** (0.014)
Fraction subprime, 2002	0.702** (0.075)	0.957** (0.057)
County fixed effects	No	Yes

This table presents coefficients from zip code-level univariate regressions of the housing wealth to net worth ratio in 2006 on various zip code-level characteristics. Each cell is from a separate regression. The regression specifications in column 2 include county-fixed effects. Standard errors are clustered by county.
**,* Coefficient statistically different than zero at the 1% and 5% confidence level, respectively.

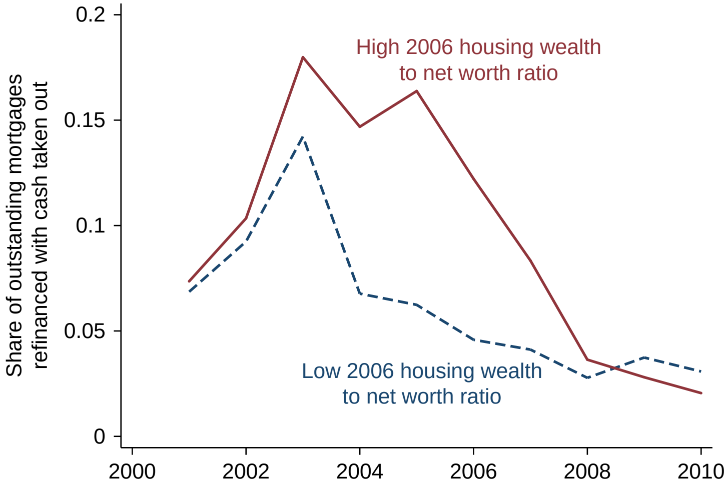


Fig. 6 Cash out refinancing share, by 2006 housing wealth to net worth ratio. For this figure, we sort zip codes into population-weighted quartiles based on the 2006 housing wealth to net worth ratio. We plot the highest and lowest quartile below. The cash out refinancing share is the share of outstanding mortgages refinanced where cash is taken out in a given year.

mortgages refinanced with cash out for the top and bottom quintile. There is a spike in cash-out refinancing in 2003 for both quintiles. But the share drops sharply for low ratio zip codes but remains elevated for high ratio zip codes through 2006. Starting in 2007, the cash-out refinancing share drops sharply for high ratio zip codes.

In Fig. 7, we utilize individual-level data from Equifax (described in detail in Mian and Sufi, 2011) to plot debt growth of existing homeowners. More specifically, we condition the sample on people who already owned their home in 1998, and then we plot total debt for the high and low 2006 housing net wealth to net worth ratio zip codes. It shows evidence consistent with Fig. 6: existing homeowners borrowed much more aggressively against their homes during the housing boom, especially in 2005, 2006, and 2007.

Taken together, the evidence in Table 4 and Figs. 5 and 7 supports the following narrative. Zip codes with a high 2006 housing wealth to net worth ratio in 2006 saw a large relative increase in house prices from 2002 to 2006, and a large increase in borrowing by homeowners. Fig. 8 shows evidence that consumption growth was also higher from 2002 to 2006, which builds on evidence in Mian and Sufi (2014). New auto sales and registered autos increased by more in high housing wealth ratio zip codes during the housing boom, and then fell sharply during the recession.[j]

[j] McCully et al. (2015) use survey data on home equity withdrawal and car purchases, and they argue that cash extracted through home equity withdrawal was not directly used for car purchases. They investigate other channels for the correlation between cash-out refinancing and auto purchases.

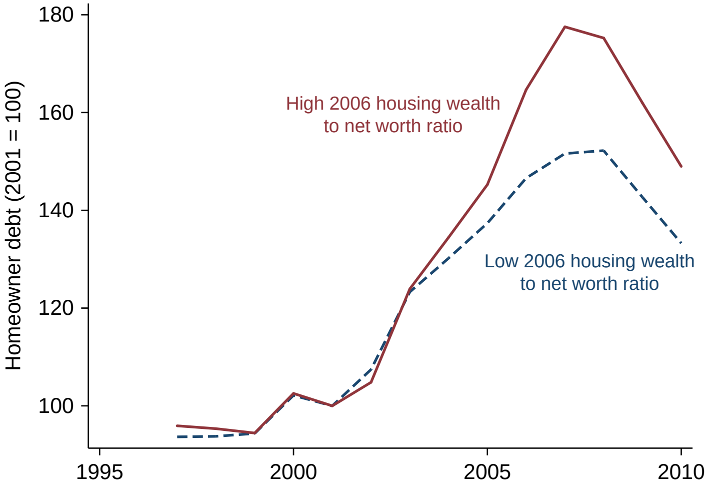


Fig. 7 Total debt of homeowners, by 2006 housing wealth to net worth ratio. This figure uses individual-level data on people who were homeowners as of 1998. We sort individuals into groups based on the housing net worth to total net worth ratio of the zip code they live in 2006. We then track total debt of homeowners in the top and bottom quartile.

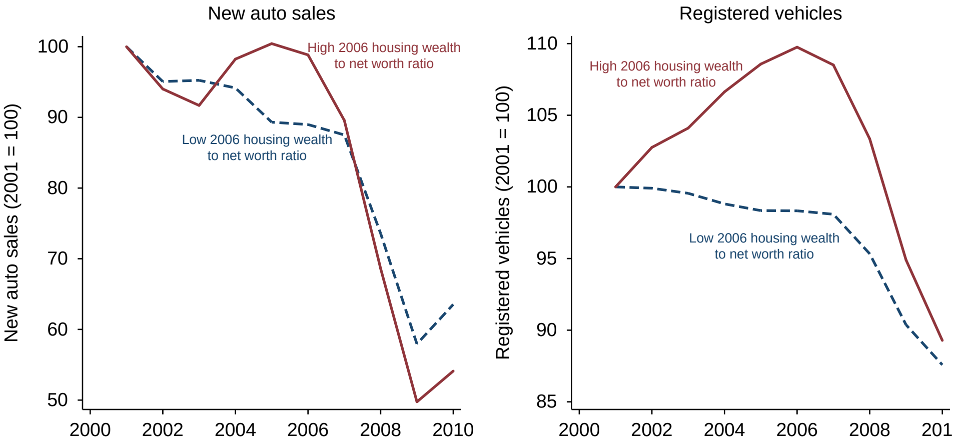


Fig. 8 Consumption of car services, by 2006 housing wealth to net worth ratio. For this figure, we sort zip codes into population-weighted quartiles based on the 2006 housing wealth to net worth ratio. We plot the highest and lowest quartile below.

4.4 Identification of House Price Effects

Our preferred interpretation of the evidence is that cross-sectional variation in exogenous house price growth from 2002 to 2006 across zip codes drove the cross-sectional variation in debt growth and consumption growth by existing homeowners across zip codes. When house prices fell, households in these same zip codes were forced to cut back sharply on consumption. The cut back in consumption was likely due to both an increase in desired savings and a tighter constraints on borrowing. This is the essence of the argument we have built in a series of studies (Mian et al., 2013; Mian and Sufi, 2011, 2014).

It is crucial to note that the causal channel we have highlighted in our previous research is related to zip codes that saw both a *boom* and *bust* in house prices. A zip code-level regression of house price growth from 2006 to 2009 on house price growth from 2002 to 2006 yields a negative coefficient with a t-statistic of 57 and an R^2 of 0.33. In our view, it is difficult to exploit exogenous variation across zip codes in the collapse in house prices independent of the boom in house prices. The boom and bust should be considered together.

The primary concern with an interpretation in which house price movements were the causal factor is that the cross-sectional variation across zip codes in the house price boom and bust was due to omitted factors that may have simultaneously driven house prices, borrowing, and consumption patterns. The most worrisome alternative explanation in our view is a shock to permanent income or productivity during the early part of the 2000s differentially affecting zip codes where house prices rose the most that subsequently reversed during the Great Recession.

In an attempt to rule out such alternative explanations for the patterns we have shown above, our previous work has relied on instruments for cross-sectional variation in house price growth. In Mian and Sufi (2011), we use two instruments: one based on across-MSA variation and the other on within-MSA variation. The across-MSA instrument is housing supply elasticity of a city according to Saiz (2010). The logic behind this instrument is that there was a national shock to housing demand from 2002 to 2006 (driven by increased credit availability or stronger preferences for housing services). This national shock translated into higher house price growth in inelastic housing supply MSAs. The within-MSA instrument is based on the interaction of MSA housing supply elasticity and zip code-level credit scores. Mortgage originations for home purchase pushed up house prices disproportionately in low credit score neighborhoods, even though these neighborhoods saw no evidence of stronger wage or income growth.

The third and fourth rows of Table 4 examine the regression coefficients of zip code-level housing wealth to total net worth ratio as of 2006 on income growth from 2002 to 2006. Echoing results from our previous research, there is no evidence of a contemporaneous positive income shock in these zip codes. In fact, income fell in relative terms in zip codes that saw the most house price growth and borrowing.[k]

The across-city housing supply elasticity instrument has been criticized by Davidoff (2013) and Davidoff (2014). In particular, Davidoff (2014) argues that supply constraints are correlated with demand growth. One implication of this argument is that perhaps the

[k] In Mian and Sufi (2014), we also utilize an instrument proposed by Charles et al. (2014) which is based on how quickly house prices accelerated in some cities during the 2000–05 period.

relative increase in borrowing and spending in inelastic housing supply cities during the 2002–06 period was driven by differential demand shocks. Davidoff (2014) does not address the evidence in Mian and Sufi (2011) and Mian and Sufi (2014) that contemporaneous measures of permanent income shocks are uncorrelated with housing supply elasticity. Further, he does not address the within-city or within-county evidence that shows that borrowing and spending growth were strongest among zip codes within inelastic cities seeing a relative *decline* in observable measures of income growth. The results described in Table 4 confirm this evidence: zip codes with the highest housing wealth to net worth ratio as of 2006 saw a relative *decline* in adjusted gross income and wage growth from 2002 to 2006, at the same time they were seeing the largest increase in borrowing and consumption of car services.

There are additional facts about high housing wealth to net worth ratio zip codes: they are poorer and have lower credit scores. The final two rows of Table 4 show that 2006 average adjusted gross income is lower in high housing wealth to net worth zip codes, and the fraction of individuals with a credit score below 660 was higher. These results are similar with or without county fixed effects. Fig. 9 shows the relation between income and the housing wealth to net worth ratio. In zip codes in the highest quintile of the housing wealth to net worth ratio as of 2006, adjusted gross income was on average $40,000. It was almost $90,000 in the lowest quintile.

This is important because many researchers sort on ex ante income or wealth when examining cross-sectional variation in consumption growth during the Great Recession. Lower consumption growth among low income individuals will reflect partially the differential effect of house price declines on low-income individuals.

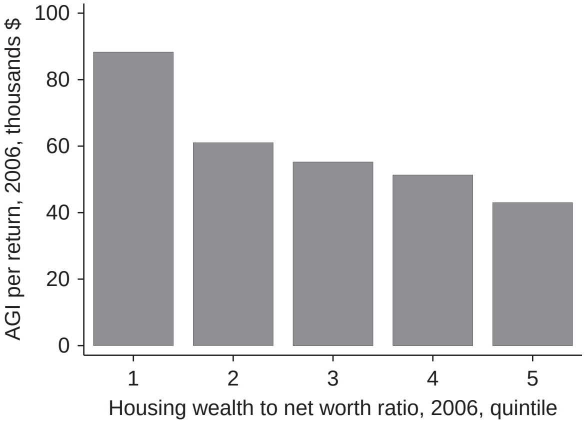


Fig. 9 Average 2006 income by 2006 housing wealth to net worth ratio. This figure plots the average 2006 adjusted gross income across the housing wealth to net worth ratio distribution. Each quintile contains 20% of the adult population.

5. MODELS MOST CLOSELY RELATED TO THESE FACTS

The importance of housing and household debt in the Great Recession has spurred a large body of theoretical research exploring the interaction of household balance sheets and consumption. The first wave of macroeconomic models that most closely match the facts presented above are Eggertsson and Krugman (2012), Guerrieri and Lorenzoni (2015), Midrigan and Philippon (2011), and Huo and Ríos-Rull (2016).

In Guerrieri and Lorenzoni (2015), agents receive idiosyncratic uninsurable productivity shocks that generate a wealth distribution. There is a borrowing constraint households face, and households that have received the worst realization of productivity shocks end up closest to the constraint. The aggregate shock that precipitates a recession in the model is an exogenous tightening of the borrowing constraint. Such a tightening generates a decline in consumption across much of the distribution. Even unconstrained agents cut back on consumption when the constraint tightens due to a precautionary savings motive.

However, the drop in consumption is largest among households closest to the constraint. Guerrieri and Lorenzoni (2015) provide a figure showing heterogeneity in the consumption response depending on proximity to the constraint (Fig. 6 in the current version). The most constrained agents cut consumption by more than 10%, whereas the least constrained agents do not adjust consumption.

In Eggertsson and Krugman (2012), there are only two types of agents where an assumption on preferences generates the variation. Constrained agents have a low discount factor and consume up to their borrowing limit every year. Their consumption is therefore pinned down by the borrowing limit. Unconstrained agents lend to the constrained agents, and the interest rate governs their intertemporal consumption allocation decision.

As in Guerrieri and Lorenzoni (2015), the shock that generates the recession is a tightening of the borrowing constraint. In response to the tightening, the constrained agents cut back on consumption. If the interest rate is free to adjust, the unconstrained agents boost consumption as the interest rate falls. However, in the presence of nominal rigidities and the zero lower bound on nominal interest rates, the sharp decline in consumption of the constrained agents pushes the economy into recession. Once again, the key cross-sectional pattern in the model is that the decline in consumption is largest for the constrained agents.

How do these models relate to the empirical findings above? One interpretation is that the tightened borrowing constraint in the model reflects collapsing house prices during the Great Recession. Indeed, as Eggertsson and Krugman (2012) put it: "there are many reasons to expect the borrowing limit to depend to some extent on current condition, for example if the collateral value of the borrowers assets depend on current market conditions (such as the price of houses)…." Under this interpretation, the model prediction that the cut-back in consumption is largest in zip codes hardest hit by the house price collapse fits well with the data. Mian et al. (2013) show that counties hit hard by the housing crash saw a larger decline in home equity loan availability, credit card limits, and mortgage refinancing volume.

In Midrigan and Philippon (2011), the key friction is a cash-in-advance constraint where both government-issued cash and private credit can be used by households to spend. Private credit for consumption must be collateralized by housing wealth, and the key parameter of interest is the fraction of housing wealth that can be borrowed against, which the authors call θ. The authors introduce heterogeneity by having islands that start identical, but then receive differential positive shocks to their own θ. The positive differential shocks to θ are meant to explain the differential rise in household debt during the housing boom. But then, θ reverts back to its preboom level, which generates a boom and bust on islands that saw the biggest rise and fall in the collateral constraint.

Midrigan and Philippon (2011) calibrate the model with assumptions on nominal rigidities, labor market rigidities, and collateral constraints that come from cross-sectional analysis of US states during the Great Recession. Their key result is that states with a larger run-up in household debt see a bigger decline in consumption during the Great Recession.

Huo and Ríos-Rull (2016) build a model in which the fundamental shock is a tightening of financial conditions facing households. A key contribution of their study is a more serious consideration of housing wealth. In their model, the reduction in credit availability to households triggers a large drop in house prices, which then in turn depresses consumption. The model contains additional frictions such as difficulty in expanding production of tradable goods and labor market frictions preventing a dramatic decline in wages. In their model, the households that see the most negative consumption growth are those with most of their wealth tied up in housing. This is consistent with the patterns we have shown above, where the decline in consumption is largest in zip codes where housing wealth was a large fraction of total net worth prior to the recession.

A second wave of studies extends the above models to explain more broadly how excessive borrowing may lead to economic downturns. Korinek and Simsek (2014) begin with a framework similar to Eggertsson and Krugman (2012), with preference differences generating a set of borrowers and lenders. Borrowers can choose any level of debt in the initial period of the model, but then a borrowing constraint is imposed in the second period of the model. If the imposed borrowing constraint is sufficiently tight, borrowers must cut spending considerably, generating a similar cross-sectional relation as in Eggertsson and Krugman (2012). Korinek and Simsek (2014) show that in the presence of nominal rigidities that generate aggregate demand externalities, borrowing in the initial period may be excessive relative to the optimum a social planner would choose.

The reasoning is that the severe cutback in consumption by borrowers in the second period generates a reduction in consumption by other individuals in the economy (the "aggregate demand externality"), given the presence of nominal rigidities such as the zero lower bound on interest rates. Borrowers do not take into account their effect on the consumption of other individuals if the borrowing constraint is binding, which implies that they borrow more than is socially optimal in the initial period.

Farhi and Werning (2015) explore economies in which nominal rigidities are present in both goods and labor markets, and the economies are subject to the zero lower bound on nominal interest rates or fixed exchange rate regimes. They show in this setting that the distribution of wealth affects aggregate demand and output, and agents may borrow too much ex ante. They apply this framework to a number of situations including credit booms and capital flows into an open economy.

Justiniano et al. (2014) build a quantitative model to show that an expansion in credit supply, as opposed to a loosening of borrowing constraints, is more consistent with house price and mortgage debt patterns witnessed from 2000 to 2006. In particular, an expansion in credit supply predicts an increase in house prices, an increase in mortgage debt to GDP ratios, a decline in interest rates, and a flat debt to collateral value ratio. However, Justiniano et al. (2014) do not explore how the expansion in credit supply affects consumption. Favilukis et al. (2015) build a quantitative model to explore dynamics in house price movements, and find that the housing boom from 2000 to 2006 was due to the relaxation of financing constraints and a decline in the risk premium associated with housing assets.

The models described utilize as exogenous shocks an expansion or contraction in borrowing limits or credit supply, and then explore the effects of these shocks on consumption and the aggregate economy. They generate cross-sectional predictions in who should experience the biggest drop in consumption during recessions. However, their primary focus is explaining the decline in aggregate economic activity. In contrast, the recent study by Berger et al. (2015) is focused on what governs the individual consumption response to a change in house prices. The build a model in which households derive utility from both the consumption of nonhousing goods and housing services and from bequeathing wealth to their children. They also face borrowing constraints and income uncertainty. Their main analytical result is to show that the individual elasticity of nondurable consumption to an unexpected and permanent change in house prices can be characterized by the following sufficient statistic:

$$\eta_{it} = \frac{\frac{dC_{it}}{C_{it}}}{\frac{dP}{P}} = MPC_{it} * (1-\delta) * \frac{P * H_{i,t-1}}{C_{it}} \tag{6}$$

where MPC_{it} is the marginal propensity to consume out of transitory income shocks for individual i at time t, δ is the depreciation rate of housing assets, P is the price of housing (assumed to be constant prior to the shock), $H_{i,t-1}$ is the amount of housing assets held by individual i at time $t-1$, and C_{it} is consumption of individual i at time t. The authors provide evidence from the PSID that the marginal propensity to consume out of transitory shocks and housing asset holdings are not strongly correlated.

Eq. (6) is derived in a partial equilibrium framework, and both sides of the equation depend on endogenous variables such as house price growth and the marginal propensity to consume. Nonetheless, it is a useful statistic for understanding the cross-sectional variation across individuals in consumption growth in response to the same decline in house prices. If one can isolate exogenous shocks to house price growth, then Eq. (6) implies that the effect on consumption growth will be larger for individuals with a larger marginal propensity to consume out of temporary income shocks and a higher housing wealth to consumption ratio.

Our zip code-level data described above is insufficient to test this equation explicitly, because we cannot measure either the marginal propensity to consume out of housing wealth or total consumption for households living in a zip code. Nonetheless, the results shown in Tables 3 and 4 are broadly supportive of Eq. (6). Panel B of Table 3 shows that consumption growth during the Great Recession was lower in zip codes with a high housing wealth to net worth ratio after controlling for house price growth. In other words, for the same decline in house prices, high housing wealth to net worth zip codes saw lower consumption growth.

While we cannot know for sure, it seems likely that zip codes with a high housing wealth to net worth ratio also have a high housing wealth to consumption ratio. Further, as Table 4 shows, high housing wealth to net worth ratio zip codes have lower income and lower credit scores. We know from a large body of research that lower income and lower credit score individuals have higher marginal propensities to consume and borrow out of income, borrowing availability, or house price shocks.

We look forward to more empirical research that explicitly tests the elasticity of consumption growth with respect to house price growth using the sufficient statistic approach of Berger et al. (2015). As a final note, while Berger et al. (2015) use a fully rational framework to derive Eq. (6), it is useful to note that marginal propensities to consume may vary across the population due to behavioral biases such as hyperbolic discounting in addition to borrowing constraints (see Harris and Laibson, 2001).

6. AGGREGATE EVIDENCE ON HOUSEHOLD DEBT

The primary focus of this review chapter is on who bears the cost of aggregate economic downturns. However, the theoretical studies discussed above—in particular, Eggertsson and Krugman (2012), Guerrieri and Lorenzoni (2015), Farhi and Werning (2015), Huo and Ríos-Rull (2016), and Korinek and Simsek (2014)—have the additional implication that elevated household debt combined with a credit supply shock and nominal rigidities may precipitate recessions. That is, these studies have both *cross-sectional* and *aggregate* implications.

Testing whether elevated household debt and a collapse in house prices cause aggregate economic downturns is a difficult exercise. For example, Beraja et al. (2015) argue that inferences about the determinants of aggregate business cycles using cross-region variation is potentially misleading because aggregate shocks cannot always be identified in the cross section. However, there is an increasing body of evidence that suggests a robust correlation between increases in household debt and subsequent economic growth. One approach is to look at cross-sectional variation across more aggregated geographical units, such as countries or states within a country. Fig. 10 utilizes new state-level personal consumption expenditure data from the BEA, and it shows a very strong relation between house price growth and consumption growth across states during the Great Recession. The estimated elasticity is 0.21 and has a *t* statistic of 7. There may be skepticism of whether we can interpret this relation as a causal effect. But previous research has attempted to isolate exogenous variation in the boom and bust in house prices, and it robustly predicts a boom and bust in household spending.

Similar analyses have been conducted across countries for the Great Recession. Glick and Lansing (2010) look across countries and find that "countries with the largest increases in household leverage tended to experience the fastest rises in house prices over the same period. These same countries tended to experience the biggest declines in household consumption once house prices started falling." An analysis by Leigh et al. (2012) confirms these findings. They find that countries with a larger run-up in household debt fueled by house price growth from 2002 to 2006 see the largest decline in

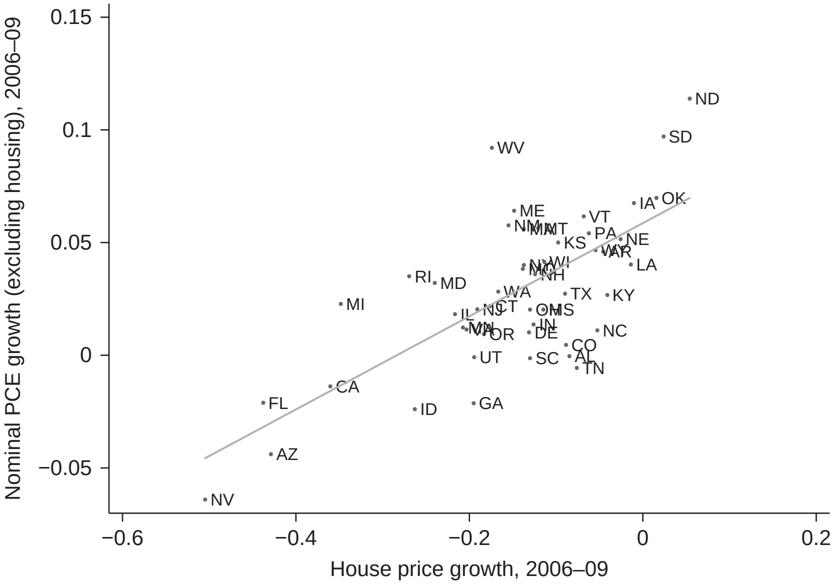


Fig. 10 Consumption growth and house price growth across states, 2006–09. This figure plots consumption growth from 2006 to 2009 against house price growth from 2006 to 2009 for states. We use BEA state-level data on personal consumption growth and house price growth data from CoreLogic.

consumption during the bust. Further, they do a more systematic analysis of past episodes from 1980 to 2011, and they find that housing busts lead to a larger decline in consumption when a large run-up in private debt precedes the bust.

The evidence goes beyond the Great Recession. In a series of studies, Jordà et al. (2013, 2014a); Jordà et al. (2014b) and Schularick and Taylor (2012) examine how growth in credit predicts financial crises and recession severity in a long historical panel of advanced countries. Schularick and Taylor (2012) estimate regressions showing that credit growth predicts financial crises, whereas Jordà et al. (2013) show that recessions preceded by a large run-up in credit tend to be more severe. Jordà et al. (2014b) extend the work in these previous two studies using novel data splitting credit into household and firm debt. They find that mortgage debt and real-estate booms predict financial crises in the post-World War II era, and they find that recessions preceded by rapid growth in mortgage debt tend to be deeper with slower recoveries. Dell'Ariccia et al. (2012) examine the characteristics of sharp increases in the bank credit to GDP ratio across a panel of countries from 1970 to 2009, and present descriptive evidence on the nature of booms and whether they lead to busts or financial crises.

Perhaps the strongest evidence in support of the aggregate implications of the theoretical frameworks by Eggertsson and Krugman (2012), Guerrieri and Lorenzoni (2015), Farhi and Werning (2015), Huo and Ríos-Rull (2016), and Korinek and Simsek (2014) comes from Mian et al. (2015). They show robust predictive power of increases in the household debt to GDP ratio on subsequent economic growth in a panel of 30 mostly advanced economies over the last 40 years. Further, they attempt to isolate *credit supply*-driven increases in household debt, and show these credit supply-driven booms predict lower subsequent economic growth. The predictive power of household debt increases is strongest in economies with fixed exchange rate regimes, supporting the importance of nominal rigidities. They also show evidence of a global household debt cycle: increases in the global household debt to GDP ratio predict subsequent lower global output growth.[1]

7. CONCLUSION

Shocks to household net worth coming from the collapse in house prices was an important determinant of consumption growth during the Great Recession. Further analysis of this pattern reveals that exposure to the housing crash as of 2006, which we measure as the 2006 housing wealth to total net worth ratio, was an important driver of this relation. Our review of the existing literature shows that very few studies focus on exposure to housing

[1] In contrast, Cecchetti et al. (2011) estimate country-level panel regressions relating economic growth from t to $t+5$ to the *level* of government, firm, and household debt in year t. They use a longer window of five years because it "reduces the potential effects of cyclical movements and allows [them] to focus on the medium-term growth rate." They do not find strong evidence that the *level* of private debt forecasts growth.

shocks when discussing the cross-sectional variation across households in consumption growth over the business cycle. We believe that more work is needed. Studies using macroeconomic data at the country level over longer time horizons find a robust relation between increases in household debt and subsequent economic downturns, which suggests that the strong relation between household balance sheets and output growth is present even outside the Great Recession.

We now have a solid body of theoretical work that can be used to motivate further empirical analysis of the effects of house prices and household debt on consumption. We believe the most important leap forward on these questions will be made if researchers are able to obtain administrative panel data on consumer spending that mirrors the high-quality panel income data from the Social Security Administration. As discussed in Mian et al. (2013), estimating differences in the marginal propensity to consume out of income and wealth shocks requires high-quality microeconomic data on consumption. Obtaining such high-quality data is even more important given the increasing focus of research on differences in the population in the marginal propensity to consume.

ACKNOWLEDGMENTS

The research presented here was supported by funding from the Initiative on Global Markets at Chicago Booth, the Fama-Miller Center at Chicago Booth, and Princeton University. We thank Ruediger Bachmann, John B. Taylor, Harald Uhlig, and seminar participants at the Handbook of Macroeconomics conference at the University of Chicago for valuable feedback. We thank Seongjin Park, Jung Sakong, and Xiao Zhang for excellent research assistance. Any opinions, findings, or conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of any other institution.

REFERENCES

Aguiar, M., Bils, M., 2015. Has consumption inequality mirrored income inequality? Am. Econ. Rev. 105 (9), 2725–2756.

Andersen, A.L., Duus, C., Jensen, T.L., 2014. Household debt and consumption during the financial crisis: evidence from Danish micro data. Danmarks Nationalbank Working Papers.

Attanasio, O., Battistin, E., Ichimura, H., 2004. What really happened to consumption inequality in the US? National Bureau of Economic Research.

Attanasio, O., Davis, S.J., 1996. Relative wage movements and the distribution of consumption. J. Polit. Econ. 104 (6), 1227–1262.

Attanasio, O., Hurst, E., Pistaferri, L., 2012. The evolution of income, consumption, and leisure inequality in the US, 1980-2010. National Bureau of Economic Research.

Baker, S.R., 2014. Debt and the consumption response to household income shocks.

Baker, S.R., Yannelis, C., 2015. Income changes and consumption: evidence from the 2013 federal government shutdown. Available at SSRN 2575461.

Bee, A., Meyer, B.D., Sullivan, J.X., 2012. The validity of consumption data: are the consumer expenditure interview and diary surveys informative? National Bureau of Economic Research.

Beraja, M., Hurst, E., Ospina, J., 2015. The aggregate implications of regional business cycles.

Berger, D., Guerrieri, V., Lorenzoni, G., Vavra, J., 2015. House prices and consumer spending. National Bureau of Economic Research.

Bewley, T., 1977. The permanent income hypothesis: a theoretical formulation. J. Econ. Theory 16 (2), 252–292.

Bunn, P., Rostom, M., 2014. Household debt and spending. Bank Engl. Q. Bull Q3, 304–315.

Cantor, D., Schneider, S., Edwards, B., 2011. Redesign options for the consumer expenditure survey. Westat.

Cecchetti, S.G., Mohanty, M.S., Zampolli, F., 2011. The real effects of debt. BIS Working Paper.

Charles, K.K., Hurst, E., Notowidigdo, M.J., 2014. Housing booms, labor market outcomes, and educational attainment. University of Chicago Working Paper.

Cochrane, J.H., 1991. A simple test of consumption insurance. J. Polit. Econ. 99 (5), 957–976.

Cynamon, B.Z., Fazzari, S.M., 2014. Inequality, the great recession, and slow recovery.

Davidoff, T., 2013. Supply elasticity and the housing cycle of the 2000s. Real Estate Econ. 41 (4), 793–813.

Davidoff, T., 2014. Supply constraints are not valid instrumental variables for home prices because they are correlated with many demand factors. Available at SSRN 2400833.

Davis, S.J., von Wachter, T., 2011. Recessions and the costs of job loss. Brook. Pap. Econ. Act. 1–72.

Dell'Ariccia, G., Laeven, L., Igan, D., Tong, H., 2012, June. Policies for macrofinancial stability: how to deal with credit booms. IMF Staff Discussion Note.

Eggertsson, G.B., Krugman, P., 2012. Debt, deleveraging, and the liquidity trap: a Fisher-Minsky-Koo approach. Q. J. Econ. 127 (3), 1469–1513.

Farhi, E., Werning, I., 2015. A theory of macroprudential policies in the presence of nominal rigidities. Working Paper.

Favilukis, J., Ludvigsson, S.C., Van Nieuwerburgh, S., 2015. The macroeconomic effects of housing wealth, housing finance, and limited risk-sharing in general equilibrium. Working Paper.

Gelman, M., Kariv, S., Shapiro, M.D., Silverman, D., Tadelis, S., 2015. How individuals smooth spending: evidence from the 2013 government shutdown using account data. National Bureau of Economic Research.

Glick, R., Lansing, K.J., 2010. Global household leverage, house prices, and consumption. FRBSF Econ. Lett. 2010 (1), 1–5.

Guerrieri, V., Lorenzoni, G., 2015. Credit crises, precautionary savings, and the liquidity trap.

Guvenen, F., Ozkan, S., Song, J., 2014. The nature of countercyclical income risk. J. Polit. Econ. 122 (3), 621–660.

Harris, C., Laibson, D., 2001. Hyperbolic discounting and consumption. In: Forthcoming Proceedings of the 8th World Congress of the Econometric Society.

Heathcote, J., Perri, F., 2015. Wealth and volatility. National Bureau of Economic Research.

Heathcote, J., Perri, F., Violante, G.L., 2010. Unequal we stand: an empirical analysis of economic inequality in the United States, 1967-2006. Rev. Econ. Dyn. 13 (1), 15–51.

Heathcote, J., Storesletten, K., Violante, G.L., 2009. Quantitative macroeconomics with heterogeneous households. Annu. Rev. Econ. 1 (1), 319–354.

Huggett, M., 1993. The risk-free rate in heterogeneous-agent incomplete-insurance economies. J. Econ. Dyn. Control 17 (5), 953–969.

Huo, Z., Ríos-Rull, J.V., 2016. Financial frictions, asset prices, and the great recession. Working Paper.

Jacobsen, M.R., van Benthem, A.A., 2015. Vehicle scrappage and gasoline policy. Am. Econ. Rev. 105 (3), 1312–1338.

Jacobson, L.S., LaLonde, R.J., Sullivan, D.G., 1993. Earnings losses of displaced workers. Am. Econ. Rev. 83, 685–709.

Jordà, O., Schularick, M., Taylor, A.M., 2013. When credit bites back. J. Money Credit Bank. 1538-461645 (s2), 3–28.

Jordà, O., Schularick, M., Taylor, A.M., 2014a. Betting the house. Federal Reserve Bank of San Francisco Working Paper.

Jordà, ‘O., Schularick, M., Taylor, A.M., 2014b. The great mortgaging: housing finance, crises, and business cycles. National Bureau of Economic Research.

Justiniano, A., Primiceri, G.E., Tambalotti, A., 2014. Credit supply and the housing boom. FRB of Chicago Working Paper.

Kahn, L.B., 2010. The long-term labor market consequences of graduating from college in a bad economy. Labour Econ. 17 (2), 303–316.

Kaplan, G., Mitman, K., Violante, G., 2015. Consumption and house prices in the great recession: model meets evidence. Working Paper.

Koijen, R., Van Nieuwerburgh, S., Vestman, R., 2014. Judging the quality of survey data by comparison with "truth" as measured by administrative records: evidence from Sweden. In: Improving the Measurement of Consumer Expenditures. University of Chicago Press.

Korinek, A., Simsek, A., 2014. Liquidity trap and excessive leverage. Working Paper.

Krebs, T., 2007. Job displacement risk and the cost of business cycles. Am. Econ. Rev. 97 (3), 664–686.

Krueger, D., Perri, F., 2006. Does income inequality lead to consumption inequality? Evidence and theory. Rev. Econ. Stud. 73 (1), 163–193.

Krusell, P., Mukoyama, T., Şahin, A., Smith, A.A., 2009. Revisiting the welfare effects of eliminating business cycles. Rev. Econ. Dyn. 12 (3), 393–404.

Krusell, P., Smith, A.A., 1999. On the welfare effects of eliminating business cycles. Rev. Econ. Dyn. 2 (1), 245–272.

Leigh, D., Igan, D., Simon, J., Topalova, P., 2012. Dealing with household debt. In: World Economic Outlook. International Monetary Fund, Washington, DC, pp. 89–124.

Lucas, R.E., 1987. Models of business cycles, vol. 26. Basil Blackwell Oxford, Hoboken, NJ, USA.

Malloy, C.J., Moskowitz, T.J., Vissing-Jørgensen, A., 2009. Long-run stockholder consumption risk and asset returns. J. Financ. 64 (6), 2427–2479.

Mankiw, N.G., Zeldes, S.P., 1991. The consumption of stockholders and nonstockholders. J. Financ. Econ. 29 (1), 97–112.

McCully, Brett, Pence, K.M., Vine, D.J., 2015. How Much Are Car Purchases Driven by Home Equity Withdrawal? Evidence from Household Surveys. Finance and Economics Discussion Series 2015–106. Board of Governors of the Federal Reserve System, Washington. http://dx.doi.org/10.17016/FEDS.2015.106.

Meyer, B.D., Sullivan, J.X., 2013. Consumption and income inequality and the great recession. Am. Econ. Rev. 103 (3), 178–183.

Meyer, B.D., Sullivan, J.X., 2013. Consumption and income inequality in the US since the 1960s. Working Paper.

Mian, A., Rao, K., Sufi, A., 2013. Household balance sheets, consumption, and the economic slump. Q. J. Econ. 128 (4), 1687–1726.

Mian, A., Sufi, A., 2010. Household leverage and the recession of 2007-09. IMF Econ. Rev. 58 (1), 74–117.

Mian, A., Sufi, A., 2011. House prices, home equity-based borrowing, and the US household leverage crisis. Am. Econ. Rev. 2132–2156.

Mian, A., Sufi, A., 2014. House price gains and US household spending from 2002 to 2006. National Bureau of Economic Research.

Mian, A.R., Sufi, A., Verner, E., 2015. Household debt and business cycles worldwide. National Bureau of Economic Research.

Midrigan, V., Philippon, T., 2011. Household leverage and the recession. NYU Working Paper.

Oreopoulos, P., von Wachter, T., Heisz, A., 2012. The short-and long-term career effects of graduating in a recession. Am. Econ. J. Appl. Econ. 4 (1), 1–29.

Parker, J.A., Vissing-Jorgensen, A., 2010. The increase in income cyclicality of high-income households and its relation to the rise in top income shares. National Bureau of Economic Research.

Perri, F., Steinberg, J., 2012. Inequality and redistribution during the Great Recession. Federal Reserve Bank of Minneapolis.

Piketty, T., Saez, E., 2003. Income inequality in the United States, 1913-1998. Q. J. Econ. 118 (1), 1–39.

Piketty, T., Saez, E., 2006. The evolution of top incomes: a historical and international perspective. Am. Econ. Rev. 96 (2), 200–205.

Piketty, T., Zucman, G., 2014. Capital is back: wealth-income ratios in rich countries 1700-2010. Q. J. Econ. 129 (3), 1255–1310.

Saez, E., 2015. Striking it richer: the evolution of top incomes in the United States (updated with 2014 preliminary estimates). University of California, Berkeley.

Saez, E., Zucman, G., 2014. Wealth inequality in the United States since 1913: evidence from capitalized income tax data. National Bureau of Economic Research.

Saiz, A., 2010. The geographic determinants of housing supply. Q. J. Econ. 125 (3), 1253–1296.

Schularick, M., Taylor, A.M., 2012. Credit booms gone bust: monetary policy, leverage cycles, and financial crises, 1870-2008. Am. Econ. Rev. 102 (2), 1029–1061. http://dx.doi.org/10.1257/aer.102.2.1029.

Schulhofer-Wohl, S., 2011. Heterogeneity and tests of risk sharing. J. Polit. Econ. 119 (5), 925–958.

Slesnick, D.T., 2001. Consumption and Social Welfare: Living Standards and Their Distribution in the United States. Cambridge University Press, UK.

Storesletten, K., Telmer, C.I., Yaron, A., 2001. The welfare cost of business cycles revisited: finite lives and cyclical variation in idiosyncratic risk. Eur. Econ. Rev. 45 (7), 1311–1339.

Storesletten, K., Telmer, C.I., Yaron, A., 2004. Cyclical dynamics in idiosyncratic labor market risk. J. Polit. Econ. 112 (3), 695–717.

Stroebel, J., Vavra, J., 2014. House prices, local demand, and retail prices. National Bureau of Economic Research.

Sullivan, D., Von Wachter, T., 2009. Job displacement and mortality: an analysis using administrative data. Q. J. Econ. 124 (3), 1265–1306.

Topel, R., 1991. Specific capital, mobility, and wages: wages rise with job seniority. J. Polit. Econ. 145–176.

Yagan, D., 2014. Moving to opportunity? Migratory insurance over the Great Recession.

CHAPTER 6

Allocative and Remitted Wages: New Facts and Challenges for Keynesian Models

S. Basu[*,†], **C.L. House**[†,‡]
[*] Boston College, Chestnut Hill, MA, United States
[†] NBER, Cambridge, MA, United States
[‡] University of Michigan, Ann Arbor, MI, United States

Contents

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.05.001

Abstract

Modern monetary business-cycle models rely heavily on price and wage rigidity. While there is substantial evidence that prices do not adjust frequently, there is much less evidence on whether wage rigidity is an important feature of real world labor markets. While real average hourly earnings are not particularly cyclical, and do not react significantly to monetary policy shocks, systematic changes in the composition of employed workers and implicit contracts within employment arrangements make it difficult to draw strong conclusions about the importance of wage rigidity. We augment a workhorse monetary DSGE model by allowing for endogenous changes in the composition of workers and also by explicitly allowing for a difference between allocative wages and remitted wages. Using both individual-level and aggregate data, we study and extend the available evidence on the cyclicality of wages and we pay particular attention to the response of wages to identified monetary policy shocks. Our analysis suggests several broad conclusions: (i) in the data, composition bias plays a modest but noticeable role in cyclical compensation patterns; (ii) empirically, both the wages for newly hired workers and the "user cost of labor" respond strongly to identified monetary policy innovations; and (iii) a model with implicit contracts between workers and firms and a flexible allocative wage replicates these patterns well. We conclude that price rigidity likely plays a substantially more important role than wage rigidity in governing economic fluctuations.

Keywords

User cost of labor, Composition bias, Sticky wages, Real wage cyclicality

JEL Classification Codes:

E24, E3, E31, E32

1. INTRODUCTION

Since at least Hume ((1742), sluggish adjustment of wages and prices has been thought to be central for understanding the monetary transmission mechanism. This is certainly true in modern New Keynesian models, of either the textbook variety or in medium-scale models that attempt to match economic data.[a] Loosely speaking, models with nominal rigidities reproduce many of the patterns featured in partial-equilibrium settings, and can make demand-determined output fluctuations consistent with both basic business cycle facts and observed reactions to changes in monetary policy. Beyond monetary nonneutrality, it is now understood that models with nominal rigidities also behave differently in response to real shocks. For example, models with nominal rigidities can

[a] For a classic textbook treatment, see Woodford (2003). The canonical medium-scale models are due to Christiano et al. (2005) and Smets and Wouters (2007).

create business-cycle comovements in response to intertemporal shocks (such as news about future technology, uncertainty, or financial frictions that change expected capital returns), even when flexible-price models would not display such comovements.[b]

It is thus important to understand the extent and importance of nominal price and wage rigidities. While we discuss both wage and price rigidities in this survey, we focus mostly on wage rigidity, for several reasons. First, Christiano et al. (2005, CEE henceforth) and many successors have found that wage rigidity is quantitatively more important than price rigidity for explaining the effects of monetary shocks and for explaining cyclical fluctuations more generally. Second, attempts to decompose the cyclical behavior of the "labor wedge"—the gap between the marginal product of labor and the marginal rate of substitution between consumption and leisure—typically find that sluggish wage adjustment accounts for a large fraction of the observed cyclical behavior of the total wedge. That is, the wage markup appears more cyclical than the price markup.[c] Third, there is broad agreement among researchers on the basic empirical facts regarding price rigidity, but there is no such consensus regarding the nature of wage rigidity. Following the initial work of Bils and Klenow (2004), a large number of recent papers have investigated the frequency and magnitude of price changes and the pass-through from costs to prices. By contrast, there are fewer studies of wage rigidity, and the ones that exist often do not relate their results to macroeconomic models. Part of the reason for the greater uncertainty regarding wage behavior is that wages are harder to measure, and it is difficult to know whether observed wages are allocative.

While there is a tendency to discuss price and wage rigidity as independent phenomena, this is incorrect at the macroeconomic level. In some cases, one might be able to take the *microeconomic* rates of wage and price rigidity—for example, exogenous Poisson hazard rates for adjusting wages or prices—as independent parameters. But as macroeconomic models make clear, the inertia of the aggregate price level—the extent of *macroeconomic* price rigidity—depends heavily on the rigidity of wages. Since most models assume that target prices are set as a constant markup on nominal marginal costs, the inertia of the price level depends on sluggish adjustment of marginal cost. Wages are the largest component of the marginal cost of producing real value added, and thus wage stickiness naturally reinforces price stickiness. Indeed, in most medium-scale models, wage stickiness is essential for obtaining price level inertia and thus, for example, persistent real effects of nominal shocks. Similarly, wage setting, for example by monopoly unions, will also be influenced by expectations of future price inflation. In another chapter in this *Handbook*, Taylor (2016) discusses microeconomic evidence on staggered wage and price setting and its implications for macroeconomic models.

[b] See, for example, Basu and Bundick (2012). The basic issues were pointed out by Barro and King (1984).

[c] See, for example, Galí et al. (2007). However, their conclusion depends sensitively on the wage measure used. See, for example, Bils et al. (2014).

Our main purpose in this survey is to discuss several definitions and measures of wage stickiness and cyclical wage adjustment and then ask what they imply for wage and price rigidity in a prototypical medium-scale macroeconomic model. Although this survey concentrates on wage measurement, we believe that it is difficult to assess the implications of data without reference to theory. Thus, we construct a medium-scale DSGE model based on CEE, but with more extensive modeling of different concepts of wages. The model distinguishes between four concepts that we discuss later: average earnings, average earnings adjusted for labor force composition, wages of new hires, and the user cost of labor. These distinct wage concepts behave in different ways in response to the monetary shock we study, so we can use the model to predict the behavior of these different concepts of wages to a monetary policy shock, which is our measure of an archetypal nominal aggregate shock.

Our use of a model to motivate the measurement has the effect of focusing attention on the wage and price statistics that we believe are most relevant for macroeconomics. These are the responses of prices and wages to identified aggregate shocks—that is, conditional correlations—rather than average business-cycle correlations or the average frequency of wage or price change. When there are both idiosyncratic and aggregate shocks, micro wages and prices may change frequently for reasons unrelated to aggregate fluctuations, but they may change only slowly in response to aggregate shocks. To concentrate attention on the statistics that matter most for macroeconomics, we focus on the responses of nominal wages and prices to a monetary shock, which is the standard example of a nominal aggregate shock. Our focus on monetary shocks does not reflect a judgment that these shocks cause a significant fraction of business-cycle fluctuations. On the contrary, most of the available evidence suggests that monetary shocks account for a relatively small fraction of output volatility. But because they are identified using a consensus set of restrictions, and because these shocks would be neutral absent some nominal rigidities, they provide a valuable opportunity for assessing the performance of macroeconomic models.

To preview our findings, we argue that recent research provides suggestive evidence that the conceptually correct measure of the allocative wage is strongly procyclical. This finding contrasts sharply with typical estimates in the macro literature, which often claim that the real wage is roughly acyclical. We then discuss the implications of the new facts about wages for models with nominal rigidities, and find that these models struggle to explain the empirical facts regarding the effects of monetary policy shocks. We show that standard DSGE models can be augmented with realistic features to reproduce many of the wage patterns found in the recent literature but that these features typically pose serious problems for the ability of DSGE models of the monetary business cycle to match estimated reactions of other variables to monetary shocks. We argue that additional evidence on measured adjustment for allocative wages and on propagation mechanisms for monetary models is needed to reconcile the micro data with our understanding of the monetary transmission mechanism. The search for this evidence should be a high priority for future research.

This survey is structured as follows. After an initial overview of different concepts of "the wage" and a survey of the relevant literature, we take the workhorse model of CEE and extend it along several dimensions. The model enables us to define several different concepts of wages in a precise manner, and to derive predictions for their behavior in response to a monetary policy shock. Importantly, the model produces output corresponding to each wage concept, even though only one is perceived by workers and firms as the allocative wage. Thus, we are able to use the model to predict the behavior of both allocative and nonallocative wages in response to a monetary shock. We then relate the model implications to existing micro data studies on wages and prices, and indicate points where further evidence is needed.

With the concepts from the model in mind, we discuss evidence on wage cyclicality, drawing mostly on research using US micro data. We focus on three issues that are crucial for interpreting the data but are generally not incorporated into macroeconomic models: composition bias in aggregate wage measures, the distinction between wages of new hires and wages of workers in continuing employment, and the distinction between spot wages and the expected time path of wages. In order to draw our statistics from a common source and ensure that they are comparable to one another, we construct the composition-corrected wages, the wages of new hires and the user cost of labor using micro data from the National Longitudinal Survey of Youth 1979 (henceforth NLSY). We then show how the different measures of real wages respond to a monetary policy shock. Our main conclusion is that real wages, correctly defined and measured, are quite procyclical, in contrast to average hourly earnings, which are basically acyclical.

We conclude by confronting the model we have developed with the empirical evidence based on micro data. We find that the allocative wage needs to be quite flexible in order to match the behavior of the real wage we estimate from the micro data. However, in order to match the behavior of average hourly earnings, it is useful to combine the flexible allocative wage with a remitted (observed) wage that changes only infrequently, also in line with evidence from micro data. A model with sticky prices, flexible wages, and implicit labor contracts comes closest to matching the impulse responses of key variables to a monetary policy shock. However, a standard medium-scale DSGE model with flexible wages struggles to match the estimated high persistence of the output response to a monetary shock. Many recent models have used wage stickiness, justified in a variety of ways, as an important propagation mechanism for shocks. With the micro data indicating that wage flexibility is a better assumption than wage stickiness, macroeconomists need to search for new propagation mechanisms in order to match the observed persistence of output fluctuations, especially to monetary shocks.

2. DEFINING "THE WAGE"

Macroeconomic models are typically populated by a large number of identical worker-consumers, who supply labor along the intensive margin in a spot market. In this setting, it is easy to define the wage: it is the current payment at time *t* for an extra unit of labor supplied in the same period. If the world were as simple as the model, "the wage" would be easy to measure. Unfortunately, nearly all of the assumptions about the labor market noted above are violated in important ways in the data, making the effort to measure wage behavior far more complicated.

First, workers are not homogeneous. This obvious fact would not necessarily create a problem for measuring wage cyclicality if the hours of workers of different types increased and decreased in synchronized fashion. Then one could define a representative worker as a worker with human capital equal to the weighted average of the human capital of all workers in the population, and show that the average wage we observe in the data is also the wage commanded by the representative worker's fixed bundle of human capital characteristics. Unfortunately, the composition of the labor force changes over the cycle. Stockman (1983) conjectured and Solon et al. (1994) confirmed that the hours of low-paid workers are more cyclical than average. Hence, low-paid workers account for a larger share of labor payments in booms than in recessions. Thus, the cyclical behavior of the aggregate (average) real wage is not the cyclicality of the wage paid to a representative worker with fixed human capital characteristics, which is the implicit or explicit concept in almost all business-cycle models. As we shall see, correcting for this composition bias shows that the wage paid to a representative worker with fixed characteristics is considerably more procyclical than the average wage in the data. This is also the conclusion of an important early paper by Bils (1985).

Second, since most workers are in long-term relationships with their employers, the labor market is not a spot market. Thus, their spot wages are not necessarily what the firm perceives as the marginal cost of labor, which is the key concept for most macroeconomic purposes. Barro (1977) used the idea of an implicit contract to criticize the "right to hire" model of wage stickiness, where workers propose a fixed, possibly nominal, wage, and firms choose employment (or hours) along their labor demand curves. He showed that other contracts would increase the payoff to both parties in the bargain, and suggested that an efficient contract would equate workers' marginal rates of substitution between consumption and leisure to firms' marginal products of labor in every period, with the total compensation for labor paid out to workers in "installment payments" over the lifetime of the worker-firm association. This reasoning follows the classic work of Becker (1962), who showed in a neoclassical setting that only the present discounted value of the wages paid by firms to workers over the length their association is allocative for employment. Holding the present value of wage payments constant, the time path of *remitted* wages can have any shape without affecting real outcomes.

Thus, one needs to know how the annuity value of the expected present value of wages, which one might conceptually regard as the "permanent wage," changes in response to changing economic conditions. By analogy to the permanent income hypothesis, the behavior of the permanent wage, not the current wage, is what matters to an optimizing worker or firm. Much of the search literature implicitly ties the behavior of the permanent wage to that of the wage for new hires, but to the extent the two differ, the permanent wage matters more. If workers and firms are in long-term associations and the permanent wage is the correct measure of the cost of labor input, then it is possible for the observed average wage to appear insensitive to cyclical fluctuations (sticky) even if the correct allocative wage is flexible. This conclusion was anticipated by Barro (1977, p. 316), who wrote, "In fact, the principal contribution of the contracting approach to short-run macro-analysis may turn out to be its implication that some frequently discussed aspects of labor markets are a facade with respect to employment fluctuations. In this category one can list sticky wages …."

3. BACKGROUND AND RELATED LITERATURE

We survey the history of research on wage cyclicality, with an eye to distilling and interpreting the evidence on the cyclicality of the marginal cost of an efficiency unit of labor to the average firm. Given the central importance of this subject for macroeconomics and the vast number of papers written about it over several decades, we can only touch on the key ideas that are most closely related to our investigation of the topic. Fortunately, a number of fine surveys of wage cyclicality have been written over the years, and we refer the reader to those for more in-depth discussions of particular issues.[d]

Our survey ranges over estimates of both nominal wage rigidity and real wage cyclicality. Both are important for assessing modern "medium-scale" macroeconomic models, and especially the ability of these models to reproduce the real effects of monetary policy shocks as observed in the data.[e] Ultimately what matters is the behavior of the "shadow" real wage facing firms. The shadow wage is the marginal cost of a unit of labor to the firm, which may or may not be what economists can readily observe in the data. To the extent that the shadow real wage is insensitive to changes in labor demand, it may be due to either real features of the economy (eg, elastic labor supply) or to wage rigidity (even if notional labor supply is inelastic), or both. From the

[d] For example, see Abraham and Haltiwanger (1995).

[e] The "narrative" approach to documenting monetary nonneutrality is exemplified by Friedman and Schwartz (1963) and is developed further by Romer and Romer (1989). The modern VAR literature on estimating the effects of monetary policy shocks originates with Bernanke and Blinder (1992). See Christiano et al. (1999) for a survey of the VAR approach. For an alternative to both the narrative and VAR approaches to identification, see Romer and Romer (2004).

standpoint of a firm, both rationales can explain why total hours fluctuate nearly as much as GDP over the business cycle.[f] However, one generally needs nominal rigidity somewhere in the model to explain why a nominal shock has real effects. But even if nominal wages can adjust freely, acyclical real wages combined with nominal price rigidity can explain why monetary shocks are generally estimated to have sizable and persistent effects on output but little effect on nominal wages.[g]

In keeping with our focus on the cost of labor to a firm, we ignore a number of related and important topics on the behavior of wages. In particular, we do not survey the literature on the *reasons* for wage rigidity, such as efficiency-wage models or insider–outsider models. We also touch only briefly on search models of the labor market, although there is an important literature combining search models with New Keynesian macroeconomics.[h]

3.1 Wage Rigidity in Historical Data

In the *General Theory*, Keynes (1936) made nominal wage rigidity the centerpiece of his theory of aggregate supply. His framework predicted that procyclical changes in prices, combined with money wage inertia, would result in countercyclical real wages. Since money wages and prices should move in the same direction, the *General Theory* predicted that nominal and real wage changes should be negatively correlated. This prediction was tested but not confirmed by Dunlop (1938) and Tarshis (1939), who took their findings as *prima facie* evidence against the hypothesis of nominal wage rigidity.[i]

On the other hand, a variety of papers have examined the historical data and find clear evidence of nominal wage rigidity in the late nineteenth and early twentieth centuries. In a classic paper, Eichengreen and Sachs (1985) used cross-sectional data for 10 countries to show that over the Great Depression period there was a negative relationship between output and real wages. They also show that countries which remained on the gold standard had low output and high real wages, while countries that left gold early experienced high output and low real wages. Bernanke and Carey (1996) extended the Eichengreen–Sachs sample to 22 countries, examined dynamics by using panel data, and performed a number of other econometric and economic robustness tests, all of which supported the basic hypothesis of nominal wage rigidity. As Bernanke and Carey emphasized, they were studying the consequences of a purely nominal shock—the transition from the Gold Standard to a fiat money regime—which took place at different times in different countries. The fact that, when countries left the Gold Standard, real wages systematically fell

[f] For a discussion of the basic statistical regularities of business cycles, see the survey by Stock and Watson (1999). For an interpretation in a neoclassical model, see King and Rebelo (1999).

[g] For a development of this argument, see Ball and Romer (1990), Kimball (1995), and Woodford (2003, chapter 3).

[h] See Walsh (2003), Ravenna and Walsh (2008), and Gertler and Trigari (2009).

[i] Pencavel (2015) discusses the early Keynesian literature on wages. Galí (2013) relates the controversies of the 1930s to modern New Keynesian analysis.

while output rose suggests, first, that purely monetary shocks can have significant real effects, and second, that expansions in nominal aggregate demand raise output by lowering real wages, giving firms an incentive to employ more labor.

Evidence suggests, however, that nominal wage rigidity is not a universal feature of labor markets. Hanes (1993) argues that wages became inflexible around the time of widespread large-scale industrial production and episodes of labor unrest, which he dates to 1890. Hanes argues that nominal wages appear to stay rigid at least through World War I, although there is some weak evidence that they become somewhat more flexible starting in the 1970s. Basu and Taylor (1999) use both time-series and cross-country data to investigate wage cyclicality. They concentrate on real wages, and interpret their results in light of the prediction of countercyclical real wages in the *General Theory*. They find that there is no definite sign of the relation of real wage movements to the business cycle. For their sample of 13 countries, real wages were slightly procyclical in the period before World War I and somewhat countercyclical in the interwar period, before becoming decidedly more procyclical after World War II. Thus, their evidence supports the idea that real wages have become more procyclical over time.

Hanes (1996) and Huang et al. (2004) seek to explain the changing cyclicality of real wages over a long historical period of more than 100 years. Both papers propose an explanation that relies on *prices* becoming more sticky over time—due to a larger number of stages of processing for goods in Hanes's case, and due to a larger output elasticity of intermediate inputs in the work of Huang *et al.* (Both papers also include mechanisms that deliver countercyclical price markups, which are important for the result.) Thus, in these works the change in real wage cyclicality over time emerges from changes that take place in the *product* market rather than the labor market. Whether or not this hypothesis is ultimately adjudged to be plausible, it is a sobering reminder that general-equilibrium effects complicate the interpretation of simple business-cycle correlations, especially in a macroeconomic setting.

3.2 Wage Rigidity in Modern Data

In a benchmark survey of business-cycle facts, Stock and Watson (1999) find an almost zero correlation between detrended real average hourly earnings and detrended GDP in postwar US data.[j] This and similar findings (for example, that labor productivity is also approximately acyclical in US data), has led modelers to emphasize preferences or institutions leading to effective labor supply functions that are nearly infinitely elastic with respect to the wage.[k] Of course, a setting in which both nominal wages and prices are

[j] Stock and Watson detrend both series using the band-pass filter, set to isolate fluctuations lasting between 6 and 32 quarters.

[k] For models in which the wage is insensitive to output fluctuations, see Hansen (1985), Rogerson (1988), Greenwood et al. (1988) and, in nonneoclassical settings, Solow (1979) and Hall (2005).

slow to adjust can also produce a real wage that is approximately acyclical regardless of preferences; this is the path taken by Christiano et al. (2005) and Smets and Wouters (2007), among others.

A near-zero average correlation between output and real wages of course admits another interpretation. It might be the case that real wages fall in response to some shocks (perhaps expansionary monetary shocks) and rise with others (perhaps positive technology shocks). If the two types of shocks are roughly equally important in the data, then on average the real wage may be acyclical. Of course, this small average correlation could hide important conditional correlations that might be far from zero. The "multiple shocks" hypothesis could also explain the instability of the correlation between output and real wages in the historical data discussed earlier. The change in the correlation between the cyclical component of wages and the cyclical component of output might just reflect the changing contributions of the two types of shocks over different subperiods.[l]

Sumner and Silver (1989) present evidence in favor of this hypothesis. They classify periods dominated by "demand shocks" as those in which output and the price level move in the same direction, while periods where the two variables move in opposite directions are classified as being dominated by "supply shocks." They find that wages move countercylically in response to demand shocks but procyclically in response to supply shocks, a finding that is consistent with an augmented version of the "Old Keynesian" model.[m]

Huang et al. (2004) argue against the "multiple shocks" interpretation of the changing correlation between output and real wages over the business cycle. Their main argument is that the observed change in cyclicality applies to conditional correlations and not just simple correlations. For example, they cite the evidence of Eichengreen and Sachs (1985) and Bernanke and Carey (1996) discussed earlier to establish that real wages decline in response to expansionary monetary shocks during the interwar period, but then refer to evidence from structural VARs run on post-war data to show that real wages rise modestly in response to expansionary monetary shocks in the recent period. The empirical results in CEE, for example, show the real wage rising slightly several quarters after a monetary expansion and then declining slightly after 10 quarters, although at no horizon is the real wage response statistically significant in either direction. In their data, one can reject the hypothesis that real wages fall significantly in response to an expansionary monetary policy shock.

[l] Geary and Kennan (1982) present evidence from the manufacturing sectors of 12 OECD countries suggesting that wage cyclicality changes significantly depending on the time period studied. They also find that the choice of deflator (a consumer price or a product price index) can make a noticeable difference. Presumably a product price index is more appropriate for testing the hypothesis that employment and wages move along a stable labor demand curve.

[m] Fleischman (1999) comes to similar conclusions using a structural VAR with long-run restrictions to identify various categories of shocks.

3.3 "The Wage" in Aggregate and Micro Data

Most papers in the historical and macro literatures examining the behavior of wages use aggregate wage data.[n] Unfortunately, aggregate data are subject to a composition bias that makes aggregate (average) wage rates less procyclical than the wages of individual workers. Stockman (1983) conjectured that low-productivity (and hence low-wage) workers would have the most cyclical employment—they would be the most likely to be fired in recessions, but also the most likely to be hired in booms. If true, then the aggregate wage (either approximately or exactly the labor-income-weighted average of the individual wage rates) would be less procyclical than the wages of individual workers, since low-wage workers would earn a larger share of labor income in booms.

Bils (1985) used individual panel data from the National Longitudinal Survey of Young Men covering the period 1966–1980, and found that wages in micro data appear extremely procyclical: a one percentage point decline in the unemployment rate is associated with a rise in real wages of 1.5–2%. While Bils finds a countercyclical composition bias in aggregate wage measures, consistent with Stockman's conjecture, he argues that this bias does not contribute significantly to his finding of a procyclical wage, since aggregate wage data also show a very procyclical real wage over this sample period. Other than the sample period, Bils attributes his finding of a procyclical wage to his inclusion of overtime earnings into his wage measure.

Solon et al. (1994) also investigate Stockman's hypothesis of composition bias in aggregate wage data using longitudinal microdata, in their case from the Panel Study of Income Dynamics (PSID) for the years 1967–87. Unlike Bils, they find that composition bias played a substantial part in reducing the apparent cyclicality of the aggregate real wage over their sample period.[o] Controlling for composition bias, they find that wages are about twice as procyclical as they appear in aggregate data. Solon, Barsky and Parker interpret their finding as consistent with movements of wages and employment along a stable aggregate labor supply curve with a labor supply elasticity between 1.0 and 1.4. They suggest that their finding is more consistent with models that predict procyclical real wages than is the usual stylized "fact" of acyclical wages, and note that both neoclassical and New Keynesian theories of the business cycle tend to predict that wages should be quite procyclical.

It may appear that the finding of strongly procyclical real wages is at odds with the finding, discussed earlier, that US labor productivity is roughly acyclical. In fact, there

[n] Ironically, the historical literature is more likely to use disaggregated data, even though high-quality data are scarce for earlier periods. For example, Hanes (1993) and Hanes and James (2003) use fixed-weight indexes of wages in narrowly-defined occupations, a wage concept akin to the Employment Cost Index (ECI) produced by the Bureau of Labor Statistics.

[o] Solon et al. (1994) argue that the estimates in Bils (1985) apply to composition bias within narrowly defined categories of workers but do not fully reflect compositional changes across groups, and thus understate the aggregate effects of compositional changes.

is no inconsistency. Once one admits that labor is heterogeneous, labor productivity needs to be measured in terms of output per efficiency unit of labor rather than output per raw labor hour. Since the lower-wage workers added in a boom contribute less in efficiency units of labor than their contribution of work hours would suggest, labor productivity correctly measured is also more procyclical than it appears in aggregate data. In fact, when it comes to measuring unit labor cost (the hourly wage divided by output per hour worked), the composition corrections for wages and labor productivity exactly offset. Thus, in the Cobb–Douglas case, the unadjusted unit labor cost measures used in the literature as a straightforward measure of the markup of price over marginal cost are not biased by cyclical changes in composition.[p]

Using data from the CPS, Daly and Hobijn (2016) come to similar conclusions regarding the "intensive" and "extensive" margins of wages. Along the intensive margin—wage changes of continuously employed individuals—wages are clearly procyclical. The extensive margin consists of cyclical changes in employment, which are concentrated among workers with lower-than-average earnings. The extensive margin makes the aggregate wage appear countercyclical. The two effects combine to make the aggregate real wage appear acyclical on average, although Daly and Hobijn note that the relative strength of the two margins varies over time, and so does the cyclicality of the aggregate wage.

Elsby et al. (2016) revisit the issues of wage cyclicality and composition bias, focusing on the experience of the United States and the United Kingdom during the Great Recession of the 2000s. They use longitudinal microdata for both countries, but note that in many respects the UK data are preferable, first because of the larger sample size, and second because the data on earnings and work hours come from the payroll data of employers, which are generally thought to be significantly more accurate than workers' recollections.

Elsby et al. report somewhat nuanced findings. They confirm the earlier microdata-based result for the United States, that men's real wages are significantly procyclical, but find that their wages were less cyclical in the Great Recession than the experience of previous large recessions would suggest. Women's real wages, which had been rising sharply in the period since 1979, stagnated during the Great Recession. However, Elsby et al. find some hints that women's wage growth was declining prior to the last recession, and thus conclude that it is too early to tell whether the lack of wage growth in the most recent recession is due to women's wages being highly procyclical or whether it is due to a shift to a new trend of slow wage growth. In one respect, the findings for the United Kingdom for both men and women are similar to those of the US men: real wages fell significantly in the Great Recession. But the variation in wage cyclicality across recessions is more or less the opposite in the two countries: wages in the United Kingdom were much more procyclical in the Great Recession than in previous recessions, while the

[p] For measures of unit labor costs interpreted as the markup, see Rotemberg and Woodford (1991, 1999), Galí and Gertler (1999), Sbordone (2002) and Nekarda and Ramey (2013).

opposite was true in the United States, at least for men. Another major difference is that composition bias appears to matter much less for measuring wage cyclicality in the United Kingdom than it does in the United States. These differences are important to bear in mind when drawing lessons from the empirical results we report in this chapter, which are based exclusively on US data.

3.4 Downward Nominal Wage Rigidity

A long strand of Keynesian analysis is based on the hypothesis that wages are more rigid downward than upward. For example, Tobin's (1972) Presidential Address suggested that workers care about relative wages, implying that they would tend to resist asynchronized wage cuts but might tolerate a neutral mechanism like inflation that cuts all real wages proportionally. This hypothesis of an asymmetry between wage increases and decreases has been the focus of a substantial literature in labor economics. One of the first researchers to address this question using micro data is McLaughlin (1994), who failed to find much evidence of asymmetry. Later work by Card and Hyslop (1997), Kahn (1997), and Lebow et al. (1999) found evidence of downward nominal wage rigidity (DNWR), including a large spike in the observed wage change distribution at zero (unchanged nominal wages), and a smaller number of small wage declines than small wage increases. Gottschalk (2005) performed an analysis of micro data on wage changes using an econometric procedure to correct for measurement error in self-reported wages, and found substantial downward nominal rigidity. The more recent papers thus suggest significant downward rigidity of nominal wages and, given the low-inflation environment that has prevailed since the mid-1990s, of real wages as well.

Hanes and James (2003) examine historical data on individual wage changes in another low-inflation period, the years 1841–91. Applying the tests for asymmetry in wage changes developed in the literature analyzing modern wage data, they find no evidence of DNWR. They interpret their results as suggesting that an aversion to nominal wage cuts is not a fundamental feature of worker preferences. They note, however, that their results do not contradict the hypothesis that institutions may have changed in such a way as to make DNWR desirable in the modern era, perhaps as a boost to worker morale and hence productivity, as suggested by Bewley (1999). Another cautionary note in interpreting the consequences of DNWR comes from Elsby (2009). Elsby begins by assuming that an aversion to nominal wage cuts is indeed a feature of workers' preferences, but then shows that preferences of this unusual form often have only a small effect on equilibrium outcomes. The reason is that dynamically optimizing firms, when confronted with a workforce that exhibits DNWR preferences, will delay nominal wage increases, thus keeping a cushion that allows real wages to rise without causing substantial employment declines if the constraint on nominal wage declines comes into play. Elsby's model suggests that it is possible to find substantial evidence of DNWR in micro data while observing few macro consequences of such asymmetric behavior. (Indeed, the evidence supporting the macroeconomic implications of DNWR does not seem to be overwhelming: see, for example, Akerlof et al. (1996).)

A number of observers have suggested that DNWR is a good explanation for the recent observation that inflation has been slow to decline during protracted and severe recessions (for example, in Japan starting in the 1990s and the Great Recession in the Untied States and other countries in the 2000s). To our knowledge, no formal evidence of this connection has been offered. However, Schmitt-Grohé and Uribe (2013) suggest that if DNWR exists, then there is a strong case for higher inflation in the Eurozone to lower real wages and stimulate employment.

3.5 Wage Change Frequency in Micro Data

Canonical New Keynesian models, such as CEE and Smets and Wouters (2007), follow Blanchard and Kiyotaki (1987) in assuming that wages for each type of worker are set by a monopoly union. Like the monopolists in the product market, the monopoly unions are subject to the Calvo friction when changing nominal wages. Thus, just as the frequency of price adjustment is important for quantifying the significance of nominal price rigidities, the frequency of wage changes in micro data is important for assessing the plausible degree of inertia in nominal wage rates. However, unlike the large literature on the rigidity of micro-level prices, there are few studies of the frequency of change of individual wages.

Barattieri et al. (2014) provide one such study using micro data from the US Survey of Income and Program Participation (SIPP). One advantage of the SIPP is that participants are surveyed three times a year, unlike participants in the PSID, who are surveyed annually. SIPP data are thus more suitable for high-frequency analysis of individual wages.[q] Because all large surveys of US micro data on individual wages use self-reported wages, a substantial fraction of the paper of Barattieri et al. (2014) is devoted to proposing a method to correct for measurement error in such a way that one can recover a consistent estimate of the frequency of individual wage adjustment. Such studies can be carried out more easily in countries where one can obtain access to administrative data, which presumably have less measurement error. Individual wage change probabilities have been analyzed for France by Le Bihan et al. (2012), for Luxembourg by Lünnemann and Wintr (2009), and for Iceland by Sigurdsson and Sigurdardottir (2016). (Researchers can access confidential administrative data sets for the United States as well, but these generally provide information on total earnings rather than hourly wage rates, which were the focus of Barattieri et al.)

As we shall see in the model of the next section, the frequency of changes in the observed wage at the individual level is an important parameter for calibrating

[q] Other well-known sources of micro wage data, the Current Population Survey (CPS) and the Employment Cost Index (ECI), do not provide sufficiently long time-series data on the wages of individual workers to be useful for this purpose.

Table 1 Wage change frequency in SIPP data

	Hourly workers[a]	Salaried workers
Reported	0.565	0.721
Adjusted	0.120	0.061
Adjusted + correction[b]	0.211	0.209
Number of obs.	17,148	21,947

[a]Based on data and calculations in Barattieri et al. (2014).
[b]Based on calculations from Gottschalk and Huynh (2010).

implicit-contracting models of the labor market. (This is true even if, as in the model we present below, the observed wage need not be the allocative wage.) The estimate reported by Barattieri et al. is not directly applicable to the full US labor market, since these authors restricted their sample to hourly paid workers. Here, we present new estimates for the frequency of wage changes for salaried workers using the methodology of Barattieri et al.[r] The results are in Table 1.

The results for hourly paid workers, the first column, reproduce the first three lines of results for the "Overall" sample in Barattieri et al. (2014, table 6). The new results for salaried workers are in the second column. Earnings per hour change even more frequently for salaried workers in the raw, self-reported data than they do for hourly paid workers. Nearly three-quarters of hourly earnings for salaried workers change each quarter. However, applying the iterative procedure of Gottschalk (2005) to correct for measurement error in wages reduces the estimate of the quarterly probability of actual wage changes for salaried workers to 6.1%. Unfortunately, this is not a consistent estimate of the desired probability due to the presence of Type I and Type II errors. Using the adjustment for the signal-to-noise ratio based on the work of Gottschalk and Huynh (2010), as presented in Barattieri et al. (2014), the final estimate of the quarterly probability of a change in earnings per hour of salaried workers is 20.9%. This figure is remarkably close to the probability of 21.1% for hourly paid workers in Table 1. In our model calibrations below, we generally set the quarterly frequency of an observed change in the remitted wage to 21%.

3.6 Implicit Contracts, Adjustment Costs, and Real Wage Cyclicality

In a classic paper, Becker (1962) showed in a neoclassical setting that only the present discounted value of the wages paid by firms to workers over the length their association is allocative for employment. Holding the present value of wage payments constant, the time path of *remitted* (observed) wages could have any shape without affecting real outcomes. For example, firms and workers might agree to an implicit contract in which remitted wage payments are smoothed relative to changes in the allocative present value

[r] We are greatly indebted to Alessandro Barattieri for these estimates.

of wages, but the fact that the observed wage is smooth would not affect real outcomes. Barro (1977) used the idea of an implicit contract to criticize the "right to hire" model of wage stickiness, where workers propose a fixed wage, and firms choose employment (or hours) along their labor demand curves. He showed that other contracts would increase the payoff to both parties in the bargain, and suggested that an efficient contract would equate workers' marginal rates of substitution between consumption and leisure to firms' marginal products of labor in every period, with the total compensation for labor paid out to workers in "installment payments" over the lifetime of the worker-firm association.

Models where workers and firms have an implicit contract over the present value of wages clearly require some assumptions about the ability of the parties to commit. In some models, such as the one we present later, one simply assumes that commitment is feasible. An alternative is to assume adjustment costs to dissolving the match for one or both parties. Absent such costs, the party that is "ahead" in the installment payments would dissolve the match. The most popular current model of labor adjustment costs is based on search. Hall (2005) addressed Barro's (1977) critique of allocative wage stickiness by showing that the allocative wage could be history dependent and hence sticky within the Diamond–Mortensen–Pissarides model of search in the labor market, as long as the preset wage remains within the Nash bargaining set generated by that model. (This argument addressed the critique of the DMP model due to Shimer (2005), who identified the sharp procyclicality of the wage as the central reason why this canonical model fails to match the volatility of the unemployment and vacancy rates.) Hall and Milgrom (2008) showed that some wage stickiness could emerge from alternative-offer bargaining. Pissarides (2009) and Gertler and Trigari (2009) showed that in the search setting, the key allocative wage is that of new hires. Haefke et al. (2013) examine data from the Current Population Survey and conclude that the wages of newly hired workers are in fact much more procyclical than average hourly earnings of all employed workers.

Relative to the literature on composition bias, the main contribution of the search-based papers is to concentrate attention on a subset of wages, namely the wages of new hires. Thus, for example, Gertler and Trigari (2009) argue that the key statistic is whether new hires receive the same wages as workers currently employed by the firm they are joining, or whether they can be hired at different wages that better reflect current economic conditions.

Assuming that newly-hired workers expect to stay with their current employer for a significant length of time, it is intuitive that their expected labor compensation consists of the expected present value of the wages they will receive over the length of the association. In this case, what matters is actually not even the cyclicality of the spot wage of new hires *per se*, but the cyclicality of the expected present value of wage payments to new hires.

In an important recent paper, Kudlyak (2014) uses such a framework to observe that one way to measure the opportunity cost of hiring a worker this period is, apart from discounting, the cost of hiring the same worker in the next period. If the labor market is a spot market, then this difference is just the current-period wage. But if there are implicit contracts, the difference of present values can differ significantly from the wage. Kudlyak observes that the object of interest, which she terms the "user cost of labor" can be constructed by using panel data on workers to estimate the present discounted value of wages at time t and $t + 1$, correcting for both observed differences in human capital characteristics and for unobserved differences by estimating a worker fixed effect. Using data from the NLSY, she presents such estimates for the period 1978–97. Kudlyak finds that the user cost is significantly more procyclical than average hourly earnings, and more procyclical than even the wage for new hires. In the empirical component of the paper, we also construct the user cost of labor using NLSY data and a procedure much like Kudlyak's, and find very similar results.

Kudlyak shows that her user cost of labor is the right measure of the allocative wage in a large range of search models of the labor market. Thus, she calls into question search models based on sticky allocative wages, as in many of the papers discussed earlier. We embed Kudlyak's insight into a standard New Keynesian model, and find that the user cost is also the allocative wage in that framework.

Kudlyak's empirical finding was foreshadowed in two important earlier papers by Beaudry and DiNardo (1991, 1995), who found that the " permanent" wage might be significantly more procyclical than the wage at a point in time. They found that workers hired when the unemployment rate was high received persistently lower wages, even after the economy recovered. Thus, while the spot wage was cyclical, the present value of the wage fluctuated even more. Beaudry and DiNardo interpreted their finding as support for the Becker–Barro hypothesis of implicit contracts with costly worker mobility. In a sense, Beaudry and DiNardo approached the problem from the workers' side, asking why a worker would take a job in a recession, since the effective (permanent) wage that he or she receives is so low. Our approach (like Kudlyak's) examines the same facts from the firms' side, asking why firms do not hire more in recessions, since the effective cost of hiring a worker in a downturn appears to be low. (Beaudry and DiNardo also argue that the data favor a model where workers cannot commit fully to a time path of future wages since, in addition to the unemployment rate that prevailed when the worker was hired, wages seem to depend positively on the minimum unemployment rate observed since the hiring date.)

Hagedorn and Manovskii (2013) argue that much of the observed history dependence of current wages can be understood by appealing to labor search when workers face a job ladder. In Hagedorn and Manovskii's search model, wages are completely determined by current labor market conditions but because workers gradually "climb" the job ladder, wages appear to be history dependent. To see this, define an employment cycle as the length of time between spells of involuntary unemployment (Wolpin, 1992). *Ceteris paribus*, the longer an employment cycle, the more job offers the worker has received.

Consequently, the current wage must be relatively high to outbid the other competing offers. An individual who enters a period of involuntary employment (thus breaking an employment cycle) falls off the job ladder, and thus his reservation wage falls. Moreover, when a worker begins a new employment cycle, his initial wage offer is determined by current labor market conditions. Workers who start an employment cycle during an expansion, start relatively high up on the job ladder because they receive relatively more offers initially. Workers who start an employment cycle during a recession receive relatively fewer offers and thus accept a lower wage initially.

In Hagedorn and Manovskii's model, the match quality of a job can be proxied by including the cumulative labor market "tightness" during the employment cycle in the wage regression. Labor market tightness is the ratio of vacancies to unemployment. Intuitively, during an employment cycle, a worker gradually climbs up the match-quality ladder. How fast he or she climbs is determined by current aggregate labor-market tightness. Ultimately, how high the person gets is given by *cumulative* labor-market tightness over the employment cycle. Hagedorn and Manovskii use empirical work based on their model to criticize the conclusions of Beaudry and DiNardo (1991, 1995). They find that when they augment wage regressions with empirical proxies for match quality based on the job ladders model, they no longer find a significant role for lagged unemployment in explaining current wages. In our empirical work using NLSY data, we investigate whether Kudlyak's finding of implicit wage contracts is sensitive to Hagedorn and Manovskii's critique.

4. THE BENCHMARK MODEL

We begin by extending a standard business cycle model to allow for several real-world features of wage setting. Our benchmark model is a standard New Keynesian DSGE system built on the basic framework analyzed in CEE. We build on the baseline model by allowing for (i) endogenous variation in the composition of the workforce and (ii) differences between the allocative wage and the measured remitted payments to workers. We will spend more time describing our treatment of labor supply and wage setting and the mapping between the model variables and data because these are the nonstandard features of the model. Many of the other mechanisms in the model are now common in the DSGE literature and the quantitative New Keynesian literature and so we present them with relatively less detailed discussion.

4.1 Households

Consumers get utility from consumption and real money balances and get disutility from working. Let C_t be consumption of a nondurable good, let N_t be labor supplied at date-t and let M_t/P_t be real money balances held at date-t. Households act to maximize

$$E_t\left[\sum_{j=0}^{\infty}\beta^j\left\{\frac{\sigma}{\sigma-1}\left[C_{t+j}-hC_{t+j-1}\right]^{\frac{\sigma-1}{\sigma}}-\phi\frac{\eta}{\eta+1}N_{t+j}^{\frac{\eta+1}{\eta}}+\Lambda\left(\frac{M_{t+j}}{P_{t+j}}\right)\right\}\right] \tag{1}$$

subject to the nominal budget constraint

$$P_t(C_t+I_t+b(u_t)K_t)+S_t+M_t=W_tN_t+R_tK_tu_t+S_{t-1}(1+i_{t-1})+\Pi_t+M_{t-1} \tag{2}$$

and the capital accumulation equation

$$K_{t+1}=K_t(1-\delta)+F(I_t,I_{t-1}) \tag{3}$$

Here, P_t is the nominal prices of the durable and the nondurable, W_t is the nominal wage rate and R_t is the nominal rental price of capital services, which is the product K_tu_t. Π_t denotes profits returned to the household through dividends. M_t is the supply of nominal money balances held at time t, S_t is nominal savings and i_t is the nominal interest rate. σ is the intertemporal elasticity of substitution, η is the Frisch labor supply elasticity, $\Lambda(.)$ expresses the household's valuation for real money balances and $h \geq 0$ is a habit persistence term ($h > 0$ implies habit persistence in utility). The function $F(.)$ is an investment adjustment cost function and $b(u_t)$ gives the resource cost of additional utilization per unit of physical capital. Following CEE, we assume that

$$F(I_t,I_{t-1})=\left[1-f\left(\frac{I_t}{I_{t-1}}\right)\right]I_t$$

with $f(1)=1, f'(1)=0$ and $f''(1)=\kappa$.

Households choose C_t, I_t, M_t, u_t and K_{t+1} to maximize (1) subject to (2) and (3). The determination of labor supply N_t is the key object of interest for this paper and we discuss this in greater detail below.

4.2 Firms and Price Setting

Following much of the New Keynesian literature we model the production and pricing component of the model as a two-stage process. Final goods are produced from a combination of intermediate goods. Final goods producers are competitive and have flexible prices. Intermediate goods firms are monopolistically competitive and change prices infrequently according to the Calvo mechanism.

4.2.1 Final Goods Producers

Final goods are produced from intermediates. Specifically, final output is given by the standard Dixit–Stiglitz aggregator

$$Y_t=\left[\int_0^1 y_t(s)^{\frac{\varepsilon-1}{\varepsilon}}ds\right]^{\frac{\varepsilon}{\varepsilon-1}}, \tag{4}$$

where $\varepsilon > 1$. Final goods producers are perfectly competitive and take the final goods price P_t and intermediate goods prices $p_t(s)$ as given. It is straightforward to show that demand for each intermediate good has the standard isoelastic form

$$y_t(s) = Y_t\left(\frac{p_t(s)}{P_t}\right)^{-\varepsilon}. \tag{5}$$

Competition among final goods producers ensures that the nominal price of the final good is a simple combination of the nominal prices of the intermediate goods used in production. Specifically,

$$P_t = \left[\int_0^1 p_t(s)^{1-\varepsilon} ds\right]^{\frac{1}{1-\varepsilon}}. \tag{6}$$

4.2.2 Intermediate Goods Producers

Intermediate goods are produced by monopolistically competitive firms who take the demand curve (5) as given when they set their prices. Each intermediate goods firm has a constant returns to scale production function

$$y_t(s) = Z_t k_t(s)^{\alpha} l_t(s)^{1-\alpha},$$

where $k_t(s)$, $l_t(s)$ and $y_t(s)$ denote capital, labor and output for intermediate producer s at time t. k_t is the quantity of capital services *inclusive of utilization*, and thus is not the firm-level equivalent of K_t, which is the stock of physical capital. Similarly, l_t is number of standardized units of labor employed by the firm. That is, it is an index of the total labor input the firm derives from the potentially heterogenous workers it employs, expressed in a common numeraire, such as the number of high-school-educated workers. This concept of labor, which is relevant for productivity, should be distinguished from N_t, which is akin to total employment or the total number of hours worked by all persons, and is the object relevant for utility. Here, Z_t is an aggregate productivity shock common to all firms. While the intermediate goods firms have some monopoly power in their output markets, they are competitive in the input markets, and take the nominal input prices W_t and R_t as given when making their decisions. Each period, firms choose their inputs to minimize costs. For any given level of production $\bar{y}$, the firm's cost-minimization problem is $\min_{l,k} Wl + Rk$ subject to $Zk^{\alpha}l^{1-\alpha} \geq \bar{y}$.

Because the production functions have constant returns to scale, and because capital and labor can flow freely across firms, firms choose the same capital-to-labor ratios. That is, for each intermediate producer s,

$$\frac{k_t(s)}{l_t(s)} = \frac{K_t u_t}{L_t},$$

where we have used the market clearing conditions $\int k_t(s)ds = u_t K_t$ and $\int l_t(s)ds = L_t$. The nominal marginal cost of production, MC_t, for the intermediate goods producers is common to all firms (because the firms have constant returns to scale production functions). It can be shown that the date-t nominal marginal cost is

$$MC_t = \left(\frac{1}{\alpha}\right)^{\alpha}\left(\frac{1}{1-\alpha}\right)^{1-\alpha}\frac{W_t^{1-\alpha}R_t^{\alpha}}{Z_t}. \tag{7}$$

Price setting for each intermediate good producer is governed by a Calvo mechanism. Let θ_p be the probability that an intermediate goods producing firm cannot reset its price in a given period. Thus, each period, $1-\theta_p$ firms reset their prices as they see fit. In many DSGE models, the firms that cannot reset their prices (ie, those that don't get the Calvo draw) are assumed to reset their prices according to a backward-looking rule. CEE refer to this modeling device as "lagged inflation indexation" in the DSGE literature. To allow for inflation indexing, we would assume that the remaining θ_p firms set their prices according to the backward-looking rule $p_t(s) = p_{t-1}(s)(1+\pi_{t-1})$ where

$$1+\pi_t = \frac{P_t}{P_{t-1}}$$

is the gross nominal inflation rate. Without inflation indexing, firms that do not get the Calvo draw simply continue to charge the same nominal price they had at the beginning of the period.

Intermediate goods firms maximize the discounted value of profits for their shareholders (the households) and thus discount future nominal profits in period $t+j$ by the stochastic discount factor $\beta^j\lambda_{t+j}$ (technically, λ_t is the Lagrange multiplier associated with the nominal constraint (2)). The optimization problem for an intermediate goods firm is to choose a reset price p_t^* to maximize the objective

$$E_t\left[\sum_{j=0}^{\infty}(\theta_p)^j\left[\beta^j\lambda_{t+j}\left(p_t^*\prod\nolimits_{s=0}^{j-1}(1+\pi_{t+s}) - MC_{t+j}\right)\left(\left[\frac{p_t^*\prod_{s=0}^{j-1}(1+\pi_{t+s})}{P_{t+j}}\right]^{-\varepsilon}Y_{t+j}\right)\right]\right]$$

where is it understood that $\prod_{s=0}^{-1}(1+\pi_{t+s}) \equiv 1$. (The expression above, and those that follow, are written under the assumption that firms index their prices to lagged inflation as discussed earlier. The equations corresponding to the model without inflation indexing are the same except that the terms $(1+\pi_{t+s})$ are all simply 1.)

Given the reset price p_t^*, and using (6), the price of the final good evolves according to

$$P_t = \left[\theta_p([1+\pi_{t-1}]P_{t-1})^{1-\varepsilon} + (1-\theta_p)(p_t^*)^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}.$$

Well-known methods show that the optimal reset price together with the dynamic evolution of the aggregate price level imply that the model satisfies a hybrid New Keynesian Phillips curve of the form

$$\widetilde{\pi}_t - \widetilde{\pi}_{t-1} = \gamma^p \widetilde{mc}_t + \beta E_t\left[\widetilde{\pi}_{t+1} - \widetilde{\pi}_t\right],$$

where $\gamma^p = \dfrac{(1-\theta_p\beta)(1-\theta_p)}{\theta_p}$ is the microeconomic rate of price adjustment.[s] We use the notation $\widetilde{v}_t$ to denote the percent deviation of the variable v_t from its steady-state value $\bar{v}$. That is, $\widetilde{v}_t = dv_t/\bar{v}$.

4.3 Labor Supply and Wage Setting

The supply of labor features several mechanisms that are prominent in the empirical literature on labor supply and the measurement of wages. As in Erceg et al. (2000) and CEE, we allow for nominal wage rigidity in the model. In addition to nominal wage stickiness, we augment the model to include two new features: (i) endogenous composition bias and (ii) a difference between allocative wages and remitted wages. Both mechanisms influence the mapping between model predictions on the one hand and empirical measures of wages and labor supply on the other. To accommodate these mechanisms, we treat the supply of labor as occurring in two separate stages within a period. We refer to these simply as stage 1 and stage 2.

In the first stage, the composition bias mechanism allocates workers with differential productivity to the market. This stage results in a single nominal wage paid for units of productivity-adjusted labor and an average wage for employed workers. We denote the wage for productivity-adjusted labor as W_t^1, the average hourly wage for employed workers as $\bar{W}_t^1$ and the total supply of effective (productivity-adjusted) labor as L_t^1 where the superscript indicates that these variables are determined in stage 1.

In the second stage, an allocative wage is determined. The allocative wage is sticky and evolves according to a Calvo mechanism taking the stage 1 wage W_t^1 as the effective marginal cost of supplying units of effective labor. In addition to the allocative wage, which governs actual employment, the second stage also produces two separate observed wages: a new-hire wage W_t^{New} and a wage for all employed workers that corresponds to average hourly earnings W_t^{AHE}.

[s] We can also allow for the possibility for *partial* inflation indexing $p_t(s) = p_{t-1}(s)(1+\pi_{t-1})^{\omega}$ with $\omega \in [0,1]$. In this case, the implied Phillips curve is

$$\widetilde{\pi}_t - \omega\widetilde{\pi}_{t-1} = \gamma^p \widetilde{mc}_t + \beta E_t\left[\widetilde{\pi}_{t+1} - \omega\widetilde{\pi}_t\right].$$

4.3.1 Composition Bias

It is well understood that the composition of the workforce changes systematically over the course of the business cycle. Typically, the labor force has a higher fraction of low-wage workers in booms than in recessions, making the average wage somewhat more countercyclical than the wage for a representative worker with fixed human capital characteristics, which is the concept of the wage in most macroeconomic models. To the extent that composition fluctuates over the cycle, the changing characteristics of the workforce automatically makes output per person appear more counter-cyclical than otherwise.

To introduce composition bias into the model, we imagine that labor varies by productivity. Specifically, we assume that total actual *hours* of labor (the argument in the utility function (1)) is given by

$$N_t = \int_0^A n_t(a)\varphi(a)da. \tag{8}$$

Here, a is an index of productivity and A is the maximum productivity of any individual in society. $\varphi(a)$ is the measure of the population with labor productivity a and $n_t(a)$ denotes hours worked per person with productivity a. For each type, $n_t(a) \in [0,1]$. The total population is $\bar{N} = \int_0^A \varphi(a)da$. Each type is paid a nominal wage $w_t^1(a)$.

Workers supply labor to labor aggregating firms who in turn sell an effective labor aggregate at a wage W_t^1. The labor aggregating firms' maximization problem is to hire different types of labor to maximize nominal profits.

$$\max_{n_t(a)} \left\{ W_t^1 \int a n_t(a)da - \int w_t^1(a)n_t(a)da \right\}$$

The labor aggregating firms' first order conditions for the choice of $n_t(a)$ requires

$$w_t^1(a) = W_t^1 a$$

for all a. That is, the individual's wage is a direct reflection of the worker's individual productivity.

Consider an increase in $n_t(a)$ from the perspective of the representative household. The utility impact of this increase is

$$\left[-\phi N_t^{\frac{1}{\eta}} + \lambda_t w_t^1(a) \right] \varphi(a) \times dn_t(a)$$

where λ_t is the shadow value of money payments to the representative household (ie, λ_t is the Lagrange multiplier on the nominal budget constraint). If the term in brackets is positive, then it is optimal to set $n_t(a) = 1$. If the term in brackets is negative, then it is optimal to set $n_t(a) = 0$. Using $w_t^1(a) = W_t^1 a$ we can express the critical productivity cutoff $\hat{a}_t$ as

$$\frac{\phi N_t^{\frac{1}{\eta}}}{\lambda_t W_t^1} = \hat{a}_t. \tag{9}$$

For any type $a > \hat{a}_t$ it is optimal to work full-time. Types $a < \hat{a}_t$ are out of the labor force. Total employment is $N_t = \int_{\hat{a}_t}^{A} \varphi(a)da$ and total effective-productivity-adjusted labor is

$$L_t^1 = \int_{\hat{a}_t}^{A} a\varphi(a)da.$$

Except for two important differences, (9) is essentially a standard labor supply condition. First, $\hat{a}_t$ is endogenous and covaries negatively with aggregate employment N_t. Second, there is a difference between effective labor L_t^1 and measured hours of employment N_t.

The *average* wage $\bar{W}_t^1$ of employed workers in the first stage is simply the ratio of total wage payments to total hours of work, that is,

$$\bar{W}_t^1 = \frac{\int_0^A w_t^1(a)n_t(a)\varphi(a)da}{N_t} = \frac{L_t^1 W_t^1}{N_t}.$$

In contrast, the composition-adjusted wage from stage 1 is simply W_t^1. Notice that the ratio of total hours worked to effective labor is equal to the ratio of the composition-adjusted wage to the average wage, $\frac{N}{L} = \frac{W}{\bar{W}}$. Using log-linear expressions for N_t and L_t, one can show that composition bias (the log difference between $\bar{W}_t$ and W_t) satisfies

$$\widetilde{\bar{W}}_t - \tilde{W}_t^1 = -\left[\frac{LN-1}{LN}\right]\tilde{N}_t \tag{10}$$

where we use the notation LN to denote the ratio of effective labor to measured hours worked L/N. Since the average wage exceeds the wage for the marginal worker (ie, since $LN > 1$), composition bias imparts a negative comovement between the average wage and aggregate hours. In US data, the cyclical variation in average real wages is negligible, while the composition-corrected wage is procyclical.

4.3.2 Allocative Wage Rigidity

In addition to the composition-bias mechanism presented above, the model features nominal wage rigidity, as in CEE. The wage block of their model is from Erceg et al. (2000). Like these earlier papers, we assume that wage rigidity applies directly to an "allocative wage," by which we mean, the relevant wage for determining employment and work effort. Unlike these earlier papers, we allow the allocative wage to differ from the remitted wage that is readily observed in data. In addition to this allocative wage, the model produces a measured remitted wage that we discuss later.

We denote the allocative wage by X_t. The allocative wage adjusts sluggishly according to a Calvo mechanism. As we did in our treatment of composition bias, we assume that there is a labor aggregating firm that assembles an aggregate of labor "types." This aggregating firm supplies effective labor to the productive firms at flow allocative wage X_t, but hires labor by type according to the type-specific allocative wages $x_t(s)$. The labor aggregate is given by a CES aggregate of labor types s,

$$L_t^2 = \left[\int_0^1 l_t(s)^{\frac{\psi-1}{\psi}} ds\right]^{\frac{\psi}{\psi-1}},$$

where the superscript 2 refers to the fact that this labor supply is determined in stage 2. (Note, this treatment is essentially identical to our treatment of prices. As we did earlier, we let s be an index of different types though in this context s refers to a type of labor while before s was a type of intermediate good.) If the aggregating firm chooses to supply labor force L_t^2, its demand for type s work is given by the isoelastic function,

$$l_t^2(s) = L_t^2 \left(\frac{x_t(s)}{X_t}\right)^{-\psi}.$$

The allocative wages $x_t(s)$ for each type s of labor are set by a monopolist in that type (similar to a union). The aggregate allocative wage X_t for units of the labor aggregate is a reflection of the type-specific allocative wages $x_{i,t}$

$$X_t = \left[\int_0^1 x_t(s)^{1-\psi} ds\right]^{\frac{1}{1-\psi}}.$$

Note that the labor market clearing condition implies that, up to a first-order approximation, the labor aggregate from stage 1 is equal to the resulting labor aggregate from stage 2 (ie, $\tilde{L}_t^1 \approx \tilde{L}_t^2$).

As we did with the price setters, we assume that the type-specific wages are set according to a Calvo mechanism. The probability of adjusting a type-specific wage is $1 - \theta_w$ and the probability of not adjusting is θ_w. As we did with the price setters, we allow for the possibility of wage inflation indexing. In this case, wage setters who do not get the Calvo draw, instead follow the wage inflation indexing rule $x_t(s) = x_{t-1}(s)(1 + \pi_{t-1})$. Without wage inflation indexing, these wage setters would simply maintain the constant nominal allocative wage they had at the start of the period. The union tries to maximize the present discounted value of wage markups $x_t(s) - W_t^1$. An extra dollar in period $t+j$ is worth $\beta^j \lambda_{t+j}$ to the household. Thus, a monopolist who has the option to set his wage at time t should choose a reset wage w_t^* to maximize

$$\max_{x_t^*} \left\{ E_t \left[\sum_{j=0}^{\infty} (\beta\theta_w)^j \lambda_{t+j} \left(x_t^* \prod_{s=0}^{j-1} (1 + \pi_{t+s}) - W_{t+j}^1 \right) L_{t+j}^2 \left(\frac{x_t^* \prod_{s=0}^{j-1} (1 + \pi_{t+s})}{X_{t+j}} \right)^{-\psi} \right] \right\}$$

where again is it understood that $\prod_{s=0}^{-1}(1+\pi_{t+s})=1$. Given the reset wage x_t^*, the aggregate allocative wage evolves according to

$$X_t=\left[\theta_w\{(1+\pi_{t-1})X_{t-1}\}^{1-\psi}+(1-\theta_w)\left(x_t^*\right)^{1-\psi}\right]^{\frac{1}{1-\psi}}.$$

4.3.3 Remitted Wages

Our discussion highlights the difference between the *allocative* wage—the shadow wage X_t that governs work effort and employment—and the *measured* wage that governs the periodic payments from the employer to the workers. We assume that the remitted wage is a smoothed function of the allocative wage. Specifically, we assume that workers periodically renegotiate their contract terms (or separate from their current jobs and get new jobs with new terms). When wage contracts are renegotiated, the workers are given a new remitted wage. Let PDV_t be the expected present discounted value of future nominal allocative wages for a newly employed worker that resets the remitted wage with probability $s\in(0,1]$. That is,

$$PDV_t=X_t+\beta(1-s)E_t\left[\frac{\lambda_{t+1}}{\lambda_t}PDV_{t+1}\right]=E_t\left[\sum_{j=0}^{\infty}[\beta(1-s)]^j\frac{\lambda_{t+j+1}}{\lambda_t}X_{t+j}\right].$$

Clearly PDV_t depends on the reset rate s (even though the reset rate plays no role in allocations). The measured remitted wage for new hires (or workers who newly renegotiated their contract) at date-t will be a smoothed version of the PDV. Specifically, we assume that the measured wage for new hires will solve

$$PDV_t=W_t^{New}E_t\left[\sum_{j=0}^{\infty}[\beta(1-s)]^j\frac{\lambda_{t+j+1}}{\lambda_t}\right].$$

That is, W_t^{New} is a constant wage that will transfer the same expected amount to the workers given the reset rate s as they would receive by getting the time-varying aggregate allocative wage, X_t. For purposes of comparison with the data, W_t^{New} is the new-hire wage.

We can also track the average outstanding wage for all workers in the model. Let W_t^{AHE} be the average hourly earnings of all employed workers. By construction, the average outstanding wage at time t is the average wage for all workers that did not renegotiate together with the new-hire wage

$$W_t^{AHE}=W_{t-1}^{AHE}(1-s)+H_tW_t^{New}$$

where $H_t=L_t-L_{t-1}(1-s)$ denotes "new hires" which we interpret as all workers who are newly hired plus those who remain employed but receive new contract terms for their remitted wage.

It is worth mentioning some of the difference between the different wage concepts W_t^1, $\bar{W}_t^1$, W_t^{AHE}, W_t^{New} and X_t. One key difference between the wages in the first stage (W_t^1 and $\bar{W}_t^1$) and the wages in the second stage (W_t^{AHE}, W_t^{New} and X_t) is that the wages in the second stage have a (potentially, time varying) wage markup. That is, in the nonstochastic steady state, $W^{AHE} = W^{New} = X = \frac{\psi}{\psi - 1} W^1$. If the allocative wage is flexible, then the markup $\frac{\psi}{\psi - 1}$ is constant even away from the steady state and in this case the dynamic behavior of the allocative wage and the stage 1 wage is the same (ie, $\widetilde{X}_t = \widetilde{W}^1_t$). If there is no composition bias, then the two stage 1 wages are the same, $W^1_t = \bar{W}^1_t$. If the renegotiation rate $s = 1$, then all of the stage 2 wages are identical, $W^{AHE}_t = W^{New}_t = X_t$. In general all of the wages will differ.

4.4 Aggregate Conditions and the Steady State

The goods market clearing condition is

$$Y_t = C_t + I_t + K_t a(u_t).$$

Although in principle there can be many different sources of uncertainty in the model, we focus our attention here on monetary shocks. We assume that monetary policy is described by a Taylor rule

$$\widetilde{\imath}_t = \left(1 - \rho^i\right)\left[\phi_Y \frac{\widetilde{Y}_t}{4} + \phi_\pi \widetilde{\pi}_t\right] + \rho^i \widetilde{\imath}_{t-1} + \varepsilon^i_t$$

Here, ϕ_Y and ϕ_π give the relative reaction of the monetary authority to output and inflation while ρ^i is an interest rate "smoothing" parameter. Here, ε^i_t is a shock to the monetary authorities policy rule. We assume that ε^i_t is mean zero and i.i.d. over time.

4.4.1 Nonstochastic Steady State

We choose parameters to ensure that in the nonstochastic steady state, $L = P = u = 1$. The steady-state markups are $\mu^p = \frac{\varepsilon}{\varepsilon - 1}$ and $\mu^w = \frac{\psi}{\psi - 1}$. We normalize the steady-state productivity cutoff to $\hat{a} = 1$. The steady-state nominal marginal cost is $MC = 1/\mu^p$. Since there is no inflation and no economic growth in the steady state, $1 + r = 1 + i = \frac{1}{\beta}$. It is straightforward to show that the nominal rental price is $R = r + \delta$ and we must have $R = b'(1)$. Steady-state capital is

$$K = \left(\frac{\alpha MC}{R}\right)^{\frac{1}{1-\alpha}}.$$

The remaining details of the steady state are standard and are therefore omitted.

4.4.2 Calibration

To compare the model to the data, as we do in Section 6, we will need to calibrate the parameters in the model. When possible, we adopt calibration settings based on conventional parameter values used for medium scale DSGE models. The discount factor β is set to 0.97 which implies a steady-state annual real interest rate of 3%. We set both the Frisch elasticity η and the intertemporal elasticity of substitution σ to 1.00. Capital's share, α, is set to 0.36. We set the type-specific elasticity of labor demand ψ to 21 which implies a 5% markup of the allocative wage over the base wage. We set the type-specific elasticity of product demand ε to 6 which implies a 20% markup of nominal price over nominal marginal cost. We set the Calvo parameters for wage and price adjustment, θ_w and θ_p, to 0.90 (quarterly). This implies that both wages and prices have an average duration of roughly 10 quarters. These durations are somewhat longer than most studies of microeconomic price adjustment data but are comparable to estimates from DSGE models.

Following Basu and Kimball (1997), we set the utilization elasticity $\frac{b''(1)}{b'(1)} = 1.00$. We set the investment adjustment cost parameter is set at $\kappa = 4.00$ and the habit persistence parameter $h = 0.65$. We allow firms to index their prices to past inflation as in CEE.

In addition to the standard parameters discussed earlier, the model also requires values for the parameters that govern composition bias and the remitted wage. There are three key parameters that govern these mechanisms: the renegotiation hazard s, the steady-state ratio of effective labor to total hours worked LN and the density of types at the steady-state productivity cutoff $\varphi(1)$. For our baseline setting, we assume that neither of these mechanisms is operative and thus we set $s = 1.00$ (so the remitted wage is equal to the allocative wage), $LN = 1.00$ (so there is no difference in average productivity per hour) and $\varphi(1) = \infty$ (so hours can be varied without changing the productivity of the marginal worker). This baseline specification is thus essentially equivalent to a standard medium-scale sticky-price/sticky-wage DSGE model. When we introduce composition bias and infrequent resetting of remitted wages we set $s = 0.21$ following Barattieri et al. (2014), $LN = 2.0$ and $\varphi(1) = 2$.

5. EMPIRICAL MEASURES OF REAL WAGES

Empirically, the cyclicality of the real wage is potentially influenced by several different features. Cyclical variations in the composition of employed workers has been emphasized as an important component of variation in real wages (see, Solon et al. (1994) and Elsby et al. (2016)). Even after correcting for compositional changes, however, it is difficult to speak unambiguously about a single concept of "the real wage." As emphasized by Haefke et al. (2013), wages of newly hired workers appear to be much more cyclical than the wages of workers who are continually employed. Beaudry and DiNardo (1991, 1995) argue that wage payments are shaped by implicit agreements between employers and employees and thus the remitted wage at a given point in time provides at best an incomplete measure of the worker's compensation. Similarly, Kudlyak (2014) finds that wage paths of workers hired at various points of the business cycle exhibit great differences in present value suggesting that the theoretical concerns articulated by Becker (1962) and Barro (1977) regarding implicit long-term contracts have empirical as well as theoretical merit.

In this section, we will examine micro data on real wages to attempt to assess whether available evidence can provide insight into how various measures of real wage payments move over the business cycle and also whether these wage measures react to monetary policy shocks. We empirically quantify the separate contributions of composition bias, variations in the new-hire wage, and variations in the present value of wage commitments.

5.1 Background

In the model in the previous section, there are several objects that map to measured wages, but only one, which we have called X_t, is allocative. Unfortunately this allocative wage is not directly measured in the data. In principle, this allocative wage can be uncovered by differencing measures of the present value at two points in time. This difference is what Kudlyak (2014) calls the *user cost of labor* (UCL).[t] Specifically,

$$UCL_t = PDV_t - \beta(1-s)PDV_{t+1} \approx X_t. \tag{11}$$

The UCL_t is only approximately equal to the allocative wage since the calculation above ignores the expectations operator and the stochastic discount factor. Kudlyak finds that unlike the average wage, the UCL is highly procyclical, even more so than the wage of new hires.

Below we will construct measures of both the UCL and the new-hire wage. Both the UCL and the new-hire wage are difficult objects to measure, since one needs to observe individual workers over time. The two panel data sets for the United States, the NLSY and the PSID, both have relatively small samples and limited sample periods. Furthermore, the data are annual, which is not ideal for business-cycle analysis. A benefit of using individual level panel data is that such data can correct for composition bias. Thus, in their early analyses of composition bias over the business cycle, Bils (1985) uses the NLSY and Solon et al. (1994) use the PSID. The recent papers by Haefke et al. (2013) and Elsby et al. (2016) instead use CPS data. The drawback to using the CPS is that since it is not a true panel, one cannot remove the effects of *unobserved* individual effects from the wages. On the other hand, the CPS has the advantage that it provides a large and nationally-representative sample, and continuous monthly data going through the Great Recession period and its aftermath.

[t] The term is used as an analogy to the "user cost of capital" under adjustment costs, in which case the decision to add an extra unit of capital is a dynamic decision with long-term consequences.

5.2 NLSY

For the purposes of the analysis in this chapter, we focus on wage data from the National Longitudinal Survey of Youth 1979 (NLSY). The NLSY is an unbalanced panel of workers initially interviewed in 1979 and then, if possible, interviewed every subsequent year until 1994 and every second year after 1994 (see below). The initial sample included 12,686 individuals born between 1957 and 1964. The birth years of the individuals in the data are distributed roughly uniformly over the years between 1957 and 1964. At the time of the initial survey in 1979 these individuals were all between 14 and 21 years of age. The initial respondents consisted of 6403 males and 6283 females.

While there are many advantages to using NLSY data, there are some disadvantages as well. Chief among these disadvantages is the fact that, due to the nature of the survey's construction, the sample in the NLSY "ages" systematically with the passage of time. This immediately means that the average wage of employed workers in the NLSY should not be directly compared to the average hourly earnings wage series constructed from NIPA data which presumably reflects wage payments to all individuals employed at any given point in time. Moreover, there is only a small amount of age variation in any single year. Thus, while we can in principle control for age in our wage regressions, the age (or more accurately "experience") coefficients will be difficult to distinguish from the growth of average wages over time.

Our data includes all data from the first interview in 1979 up to 2013. Because the NLSY was modified to a biennial survey starting in 1994, we drop all of the odd years between 1994 and 2012.[u] For our analysis here, we focus exclusively on men. Thus our sample consists of the 6403 men interviewed initially in 1979 and then followed until 2012. Although the NLSY is not a representative sample of the US population, the survey provides a yearly cross-sectional weight variable that can be used to make the sample comparable to that year's population.[v]

[u] While the NLSY does ask the respondent to remember information for the previous years after it made the transition to biennial surveys, the wage series and responses appear to be systematically different for the odd years.

[v] Of the whole initial interview sample, roughly half (6111) comprised what the NLSY refers to as a "representative sample" of the noninstitutionalized working-age population born between 1957 and 1964. In addition to the representative sample, the NLSY also collected data on a "disadvantaged sample" consisting of 5295 individuals who identified as Hispanic, Latino, Black and economically disadvantaged respondents. Finally, the NLSY includes 1280 respondents who are representative of the population serving in the armed forces. This latter sample is referred to as the "military" sample. Both the disadvantaged sample and the military sample were severely cut back or eliminated entirely from the NLSY in 1984 and again in 1990. We keep all males in each of the three subgroups (the representative sample, the military sample and the disadvantaged sample). We then use the cross-sectional weights to convert the NLSY data to an overall representative sample. Note, we do not use the longitudinal weights that are included in the NLSY. The longitudinal weights are intended to produce a representative panel over the entire period.

The NLSY reports wage information for up to five jobs each year. Our sample technically includes data from jobs in 1978 even though the first interviews were done in 1979. (The 1978 data come from interviews that were done early in 1979 and so pertained to jobs in 1978.) We focus on the "hourly rate of pay" variable that is constructed by the NLSY. Respondents are asked for the most convenient way to report their total earnings.[w] They could report pay per hour, per day, per week, per month, or per year. In every case, the reported statistic is then converted to an hourly pay rate based on a measure of the respondents typical hours worked. The resulting hourly rate of pay includes tips, overtime pay, and bonuses but is computed before any deductions. To construct real wages, we deflate the hourly rate of pay with a price index. We considered two separate price deflators in our analysis: the consumer price index and the implicit price deflator for the nonfarm business sector.[x,y] The analysis for the two separate price indices were quite similar overall. Since our focus is on intertemporal labor demand from the firms' perspective (ie, the real product wage), we focus on the real wage measures using the deflator for the nonfarm business sector in the discussion later.

In addition to the information on wages, the NLSY includes information on the industry of the jobs and whether the jobs are covered by a union. We do not include union status in our analysis because the union variables included in the Employer History Roster exhibit an unusual change following the 1994 change from annual to biennial coverage in the NLSY.

5.2.1 Wage Regressions

We begin by describing how we construct the various measures of real wage series from the NLSY data. Given the available data, as described above, we run regressions of the following form:

$$\ln w_{t,\tau}^{i} = c + \alpha^{i} + \zeta t + \Psi X_{t}^{i} + \sum_{d_0=1}^{T} \sum_{d=d_0}^{T} \chi_{d_0,d} D_{d_0,d}^{i} + \varepsilon_{t}^{i}. \tag{12}$$

This is the basic empirical specification considered in Kudlyak (2014). Here, $w_{t,\tau}^{i}$ is the real wage for individual i at time t who was hired at time τ. This regression provides a best linear prediction of the log real wage at time t of a worker i, who started his job in period τ. In its most general form, this wage regression allows for a time trend, demographic and

[w] QES-71A in the 2012 survey.

[x] We used the consumer price index for all urban consumers: All items and the nonfarm business sector implicit price deflator. Both variables are seasonally adjusted and are available from the FRED database as CPIAUCSL and IPDNBS.

[y] We exclude wage rates less than 1 dollar per hour and above 100 dollars per hour measured in 1979 dollars. This restriction led to the elimination of 2894 wage-year observations. This censoring at 1 dollar and 100 dollars is the same censoring used by the BLS when it uses NLSY data.

industry controls (included in X_t^i), individual fixed effects (the α^i coefficients), and time effects that depend on two periods: when the individual began work at his current job and the current date. The additional covariates in the X_t^i matrix are the individual's experience at time t (and experience squared), tenure at time t (and tenure squared), schooling completed, and industry fixed effects. Experience is defined as the maximum of (Age − 6 − years of schooling) and 0. The dummy variables $D_{d_0,d}^i$ take the value 1 if $d_0 = \tau$ and $d = t$ and 0 otherwise.

The χ coefficients are particularly important for interpretation of the new-hire wage series and the user-cost series that we emphasize below. At time t, all workers who began work at their current job at date-τ get an additional adjustment to their predicted wage given by the coefficient $\chi_{\tau,t}$. These adjustments imply that workers who begin at date-τ experience an expected strip of log wage realizations given by $\{\hat{\chi}_{\tau,\tau}, \hat{\chi}_{\tau+1,\tau}, \hat{\chi}_{\tau+2,\tau}, \ldots \hat{\chi}_{\tau+j,\tau}, \ldots \text{etc}\}$. These dummy variables thus adjust for vintages of hired workers, where the vintage is defined by when the worker was hired in addition to the current calendar date. Notice that the variable $\hat{\chi}_{\tau,\tau}$ reflects the wages of a newly hired worker (ie, the date-τ wage of a worker hired at date-τ). In the estimation, we truncate the χ strips at 7 years (including year 0).[z]

This specification can also be used to calculate composition adjusted wages. For instance, if we restrict the $\chi_{\tau+j,\tau}$ coefficients to be zero then the resulting specification gives a predicted wage that adjusts for both observed changes in workforce composition (by including the X_t^i variables) and unobserved workforce composition (by including the individual fixed effects α^i), but does not allow for vintage effects on the wage. Adding the χ dummy variables allows us to recover composition-adjusted wages with vintage effects. For example, the coefficient on $\chi_{\tau,\tau}$ tells us whether a newly hired worker receives a wage increase or reduction relative to workers hired in previous years, controlling for any differences in human capital between new hires and other workers.

5.2.2 Average Hourly Earnings and New-Hire Wages

Before we consider our measures of the user cost of labor, we first examine average hourly earnings. We consider two measures. The first is a measure taken from the BLS. The BLS reports a measure of compensation per hour for the nonfarm business sector. We then deflate this measure by the implicit price deflator.[aa] We refer to this measure as AHE-BLS.

[z] More precisely, we include all of the dummy variables in the estimation of (12); however, following Kudlyak (2014), we use only seven χ estimates when we calculate the user cost of labor.

[aa] The variables used in this calculation are from the FRED database. We use nonfarm business sector: Compensation Per Hour (COMPNFB) and the Nonfarm Business Sector: Implicit Price Deflator (IPBNBS).

Our second measure of average hourly earnings is constructed from the NLSY data. We refer to this measure as AHE-NLSY. This wage series is constructed by first running the simplified version of regression (12)

$$\ln w_t^i = c + \Psi X_t^i + \sum_{d=1}^{T} \omega_d D_d^i + \varepsilon_t^i,$$

where D_d^i is a time dummy variable (D_d^i takes the value 1 if $d = t$ and zero otherwise). For the NLSY measure of average hourly earnings, the controls X_t^i include only experience and experience squared. Because the experience variable is defined as $\max\{\text{age} - 6 - \text{schooling}\}$ this is close to being a control for age and age squared. The estimated time fixed effects $\hat{\omega}_t$ are then an estimated time series of average hourly earnings. Note that, because the NLSY is based on a fixed set of individuals who were entering the workforce in the late 1970s, it is crucially important to include controls for age in this measure. If we did not include experience and experience squared, then the sample would systematically age and this would impart a systematic aging component to the wage measures.

To construct the new-hire wage, we return to the original regression specification (12). As noted, the new-hire wage series corresponds to the estimated coefficients $\hat{\chi}_{t,t}$. We include all of the available demographic controls in X_t^i and we also include individual fixed effects in the regression.[ab]

5.2.3 Calculating the User Cost of Labor

We base our calculation of the user cost of labor (UCL) on equation (11). To calculate the user cost, we need to calculate a forecast of the present value of wage payments for a worker hired at date-t and the present value of wage payments for a worker hired at date-$t + 1$. For an individual hired at date-t and still employed at date-$t + j$, we construct the predicted value of the log real wage $\widehat{\ln w_{t,t+j}}$. We can then calculate the implied present value of compensation as

$$\widehat{PDV}_t = \sum_{j=0}^{\infty} \beta^j (1-s)^j \exp\left\{\widehat{\ln w_{t,t+j}}\right\}.$$

Note that in addition to requiring a sequence of predicted log wages $\{\widehat{\ln w_{t,t+j}}\}_{j=0}^{\infty}$, this calculation requires a separation rate s and a discount factor β.

To construct the projected wage payments $\widehat{\ln w_{t,\tau}}$, we consider the anticipated wage payments for a firm that hires an "average worker" at date-t. As the employment

[ab] Our method for constructing the new-hire wage differs from that in Kudlyak (2014), who simply examines the wages for workers hired in the current year. Our procedure creates a wage series for new hires correcting for composition, in parallel with our construction of the user cost of labor.

relationship continues, our measure of the worker's experience and our measure of the worker's tenure both increase. We assume that the initial experience is fixed at the sample average of 11.72 years and we set the initial tenure variable to 0.5 years (this implicitly assumes that a worker who reports being newly hired at his current job at the time of the interview was hired 6 months earlier). We set the worker's schooling to 12.57 years, again the sample average in the NLSY. Then, based on (12), at date-τ, a worker hired at date-$t \leq \tau$ has a projected log wage

$$\widehat{\ln w_{t,\tau}} = \hat{c} + \hat{\zeta}\tau + \hat{\Psi}\bar{X}_{\tau-t} + \hat{\chi}_{\tau,t} \tag{13}$$

where $\bar{X}_{\tau-t}$ are demographic controls for the "average worker" (ie, schooling = 12.57, experience = 11.72 + $\tau - t$ and tenure = 0.5 + $\tau - t$).

For the separation rate s, we follow Kudlyak (2014) who uses a monthly separation rate of 0.0295. This figure is based on calculations of the average separation rate in the NLSY. We then convert this monthly separation rate into an annual separation rate by setting $s = 1 - (1-0.0295)^{12} = 0.3019$. The NLSY figure might be somewhat low relative to other datasets. The separation rate from the JOLTS dataset is closer to 0.035. The annual discount factor is set to 0.97. Note that our calculation of the present value of wage payments is truncated at 7 years (including the initial year). Given the high observed separation rates in the data, this truncation has a relatively small effect on the present value.

5.3 The Cyclicality of Real Labor Compensation

We are now in a position to examine the cyclical behavior of real wages. Tables 2 and 3 report cyclicality estimates for six different measures of log real wages. For each measure of real wages, we regress the calculated wage series on an indicator of the business cycle (and a time trend and a constant). Table 2 examines the cyclicality of real wages with respect to the HP filtered unemployment rate. We use HP filtered unemployment rather than the unemployment rate in levels because the average unemployment rate changes substantially over the time period for the NLSY.[ac] Thus, the coefficients reported are semielasticities: the percent change in a real wage measure in response to a one percentage-point deviation of unemployment relative to its trend. The sample for columns 1–5 consists of 25 data points from 1979 to 2012, dropping the odd years between 1994 and 2012 (see Section 3.2). To construct the UCL, we need to impute values of wages for the odd years between 1994 and 2012. The final user cost series itself ends

[ac] The HP filtered unemployment rate is taken from monthly data from 1985 to 2016. To avoid the well-known endpoint problem in HP filtering, we add 120 months of predicted unemployment rates taken from an estimated AR(6) to the end of the sample. We then HP filter the padded series using a smoothing parameter of 500,000. The regression uses annual averages of the monthly HP deviations.

Table 2 Real wage cyclicality: Unemployment rate

	AHE–BLS	AHE–NLSY			New hire	UCL
	(1)	Base (2)	Controls (3)	Controls, FEs (4)	(5)	(6)
HP-filtered unemployment rate	−0.507 (0.471)	−0.976 (1.530)	−1.185 (1.507)	−1.328 (1.623)	−0.698 (1.822)	−5.818 (2.079)
Observations	34	25	25	25	25	27

Notes: OLS standard errors are in parentheses. Coefficients are multiplied by 100.

Table 3 Real wage cyclicality: GDP

	AHE–BLS	AHE–NLSY			New hire	UCL
	(1)	Base (2)	Controls (3)	Controls, FEs (4)	(5)	(6)
HP-filtered GDP	0.311 (0.353)	0.984 (1.093)	0.960 (1.082)	1.165 (1.161)	1.325 (1.287)	3.122 (1.351)
Observations	34	25	25	25	25	27

Notes: OLS standard errors are in parentheses.

in 2007 because we require seven subsequent wage observations to calculate the value of the UCL in year *t* (again see Section 3.2 for details).

Columns 1–4 report results for average hourly earnings. For the BLS wage series, AHE-BLS, the coefficient on the unemployment rate is −0.507: real average hourly earnings fall by roughly 0.5% for each percentage point increase in the cyclical component of the unemployment rate. Columns 2–4 report results for our constructed measure of average hourly earnings from the NLSY data, AHE-NLSY. As noted in our discussion earlier, the dependent variables in Columns 2–4 are estimated time fixed effects from regressions of individual wages on the listed set of controls. The columns differ according to the number of controls included in the regression. Column 2 includes only experience and experience squared; column 3 adds industry fixed effects, job tenure and schooling; column 4 includes all of the aforementioned controls and adds individual fixed effects. The NLSY sample exhibits greater cyclicality for all of the measures of average hourly earnings, and the cyclicality rises with the number of controls for worker characteristics. We interpret this finding as being supportive of the basic composition-bias effect emphasized by Bils (1985) and Solon et al. (1994). Typically, as we add more controls for worker heterogeneity, the point estimate of the cyclicality rises (though note, the standard errors are high enough that we cannot say with any certainty that any one of these measures is clearly more or less cyclical than any other).

Column 5 reports results for the new-hire wage. The point estimate for the cyclicality coefficient is − 0.698, so a one percentage point increase in the cyclical component of unemployment corresponds to a 0.7% reduction in the real new-hire wage. By itself, the point estimate seems to be at odds with the findings in Haefke et al. (2013) who reported that in CPS data, the wages of newly hired workers appeared substantially more cyclical than average hourly earnings. We should note that while our point estimates do not indicate greater cyclicality of the new-hire wage, the estimates are quite noisy and admit a range of interpretations.

Column 6 reports results for the user cost of labor (UCL). Our measure of the UCL exhibits *much* greater cyclicality than either the composition-adjusted wage or the new-hire wage series. In Table 2, the cyclicality estimate is − 5.818 indicating that for every one percentage-point increase in the cyclical component of unemployment, the real user cost of labor falls by almost 6% (!).

The estimates in Table 2 are robust to alternate measures of the business cycle. Table 3 reports estimates for the same dependent variables as those in Table 2, but uses HP filtered GDP as the indicator of the business cycle instead of the unemployment rate. Again, average hourly earnings seem to be only moderately cyclical. When HP filtered GDP is above trend by 1%, AHE-BLS is above trend by only 0.311%. By contrast, holding the set of workers fixed in the NLSY and controlling for observed and unobserved heterogeneity increases this estimate to 1.165%. The point estimate of the cyclicality of the new-hire wage is more cyclical. The point estimate is a rise of roughly 1.3% for every 1% change in the cyclical component of GDP. Finally, as before, the UCL is the most cyclical wage measure. For each percent increase in GDP above trend, the UCL rises by approximately 3.1%.

What these results seem to suggest is that both composition bias and implicit contracting play important roles in shaping the wage payments made to workers over the business cycle. Quantitatively, controlling for composition (by including individual fixed effects and controls for observed worker differences in the wage regressions) increases wage cyclicality by perhaps as much as a factor of two relative to a group of workers without such controls. The effects of implicit contracting and wage-smoothing seem to be even greater than the effects of composition bias. According to our calculations, the user cost of labor has a cyclicality that is, in some cases, about six times greater than the log real wages of the base group. Since average payments are less cyclical than the user cost, workers hired in bad times are paid a wage greater than their user cost. In return, the workers expect to receive fewer and smaller wage increases over their employment spell.

Our findings (which are consistent with the results in Kudlyak, 2014) seem to corroborate the results in Beaudry and DiNardo (1991, 1995), who argued that current wage payments seem to be tied to past labor market conditions. In that paper, the authors showed that the maximum unemployment rate during a job spell and the unemployment rate that prevailed when the worker was hired both have a significant influence on current wage payments. The specification above, which we have adapted from Kudlyak's work, is a more general econometric specification than the one in Beaudry and DiNardo, but implicit contracts still appear to play an important role in shaping wage payments.

5.4 Wage Responses to Monetary Shocks

Almost all of the literature on wage cyclicality examines the response of real wages to a cyclical indicator, typically the unemployment rate. However, the monetary business-cycle literature has also emphasized the importance of replicating estimated impulse response functions to identified shocks—most often monetary shocks. The modern literature on estimating the effects of monetary policy shocks using VARs began with Bernanke and Blinder (1992). Here we follow the approach in CEE, since we ultimately want to make comparisons between empirical and theoretical impulse responses to monetary shocks.

To implement the VAR procedure, we first extend our annual real wage measures from the NLSY to a longer quarterly series using the Chow–Lin procedure. The extension to quarterly data is important for the validity of the identifying assumptions commonly used in the VAR literature. The identification assumptions invoked are plausible in quarterly observations but this plausibility becomes strained if the data are sampled at an annual frequency. The Chow–Lin method uses the annual data to estimate the relationship between the annual wage measures constructed above and other variables that are available at a quarterly frequency. The variables used in the Chow–Lin procedure are Real Gross Domestic Product, Real Hourly Compensation in the nonfarm business sector, the Civilian Unemployment Rate, and Total Nonfarm Payrolls for All Employees.[ad] The resulting interpolated series distributes the annual measure to the corresponding quarters (thus, the annual averages of the constructed quarterly series equal the original annual measures). We then extend the series by projecting the missing data to periods outside the years 1979–2012 covered by the NLSY. We first regress the interpolated quarterly wage measures on the variables in the Chow–Lin procedure above. We then use the OLS estimates to form estimates $\hat{w}_t$ for time periods earlier than 1979 and later than 2012. Fig. 1 plots quarterly average hourly earnings, the new-hire wage, and the user cost of labor for the period 1965:1 to 2015:3. For each series, a separate linear trend was removed prior to plotting. Each series is in log points and is plotted so that the mean of each series is centered at 1.00.

The impulse response functions to monetary policy shocks are constructed following the approach recommended by CEE. We include the same variables, in the same

[ad] All variables are in logs except for the unemployment rate which is entered in levels.

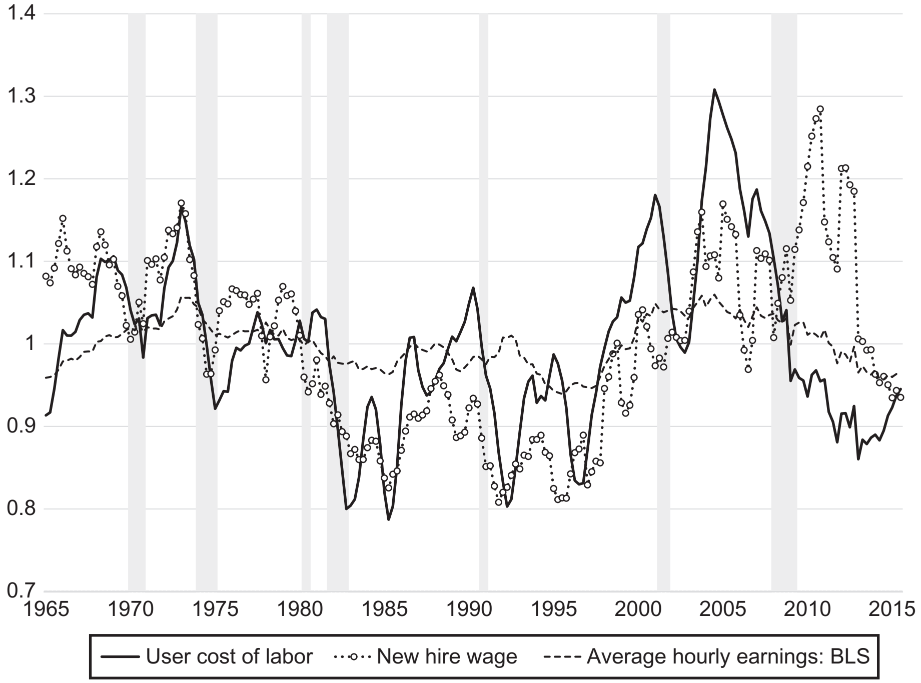


Fig. 1 Measures of the real wage.

Choleski ordering as in the original CEE specification. In order, the variables are real output, real consumption expenditure, the price level, real investment spending, real average hourly compensation, average labor productivity, the federal funds rate, real corporate profits, and the growth rate of the money supply (M2). Following Bernanke and Blinder (1992), the innovation to the funds rate is identified as a structural shock to monetary policy. Notice that by assumption, none of the variables in the first block (output, consumption, the price level, investment, compensation and labor productivity) responds contemporaneously to a shock to monetary policy. In contrast, both corporate profits and the growth rate of M2 respond contemporaneously to monetary shocks. Our approach is to extend the CEE specification by appending a single additional variable—an additional wage measure—to the second block of variables. Thus, our augmented VAR introduces a wage which is allowed to respond contemporaneously to monetary shocks. However, we add the restriction that monetary policy does not respond contemporaneously to shocks to the new wage measure. This restriction is sufficient to identify the impulse response of the wage measure to a monetary policy shock.

We do not want to allow the new wage measures to influence the identified monetary policy shocks. That is, we wish to ensure that the identified shocks remain the same as we change our measure of wages in the VAR. This first consideration implies that the new wage measures should be excluded from the dynamic equations governing the variables originally included in the CEE specification. It also suggests that we should order the new wage series last so that these new measures will respond to the other variables but the other variables—in particular the federal funds rate—will not respond to the alternate wage measures.[ae] One consequence of ordering the new wage measures last is that they respond contemporaneously to a monetary policy shock. In CEE's original specification their measure of the wage, average hourly earnings, comes before the federal funds rate, and may respond to monetary policy shocks only with a one-quarter lag. Thus, our procedure treats the new wage measures differently from average hourly earnings, but only for the first quarter after a monetary policy shock.[af]

As in CEE, consumption, investment and corporate profits come from the BEA's NIPA tables. Unlike CEE, our measures of output, the price level, employee compensation and labor productivity are only for the nonfarm business sector. Our decision to use the nonfarm business sector is motivated primarily by our belief that the nonfarm business sector is a better match to models of infrequent price adjustment by firms that are trying to maximize profits. Excluded industries (such as utilities and government production) likely do not set prices optimally the way most macroeconomic models posit. All variables are in log levels, except for the federal funds rate which is in levels and M2 which is in log differences. All variables were downloaded from the Federal Reserve Bank of St. Louis FRED web database.

To estimate the VAR, we use the same sample used by CEE, namely 1965:3 to 1995:3. We do so in order to make it easy to compare our results to the ones in this benchmark paper.[ag] We experimented with other sample periods, including extending the sample forward to 2007. The extended sample would allow us to estimate the VAR using a larger data set, while stopping short of the zero-lower-bound period in which the identifying assumptions do not apply. Unfortunately, we found that the impulse responses reported by CEE change significantly when the later data are added, in ways that are difficult to interpret in light of the underlying theory. (For example, the well-known "price puzzle" is clearly apparent in the extended sample.) Since our main objective is to explore how our novel real wage measures responded to a well-known identified shock, we chose to estimate the VAR over the original CEE sample. However, we believe that the instability of the impulse responses to monetary shocks over different

[ae] As in CEE, the federal funds rate is assumed to respond to average hourly compensation. Thus, we *always* include real hourly compensation in the VAR. This ensures that our identification assumptions match those in CEE.

[af] We experimented with other identification schemes, including ones that constrain the contemporaneous responses of the new wage measures to zero, symmetrically with average hourly earnings. The results we report below are qualitatively robust to all the variants we tried.

[ag] Even though our sample is identical with the sample in CEE, differences will arise because of data revisions and also because we use NFB output and prices rather than GDP and the GDP deflator.

sample periods—reminiscent of the results in Hanson (2004)—is worthy of investigation in its own right.[ah]

Our focus in this chapter is on the responses of the various wage measures constructed above. For each "new" measure of the real wage, we estimate a different VAR system to recover the response of the wage to the monetary shock. For example, we estimate the system separately for the new-hire wage and for the user cost. Given the VAR structure discussed later, the new variable does not affect the responses of the original CEE variables in any way. As a result, we report the impulse responses of the baseline set of variables to a monetary policy shock in one block, since these do not change as we change the additional wage variable added to the CEE specification.

Fig. 2 shows the reaction of the standard macroeconomic variables included in the CEE system to an identified increase in the federal funds rate of 50 basis points. (To save space, the figure omits the responses for corporate profits and for the growth rate of M2.) Each panel reports the impulse response of a single variable. The units of all variables are reported in percentage points (ie, 1.00 corresponds to 1%). The dotted lines correspond to 1 standard deviation error bands. The shock leads to a reduction in output of roughly 0.25%, a reduction in nondurable consumption of slightly less than 0.2% and a reduction

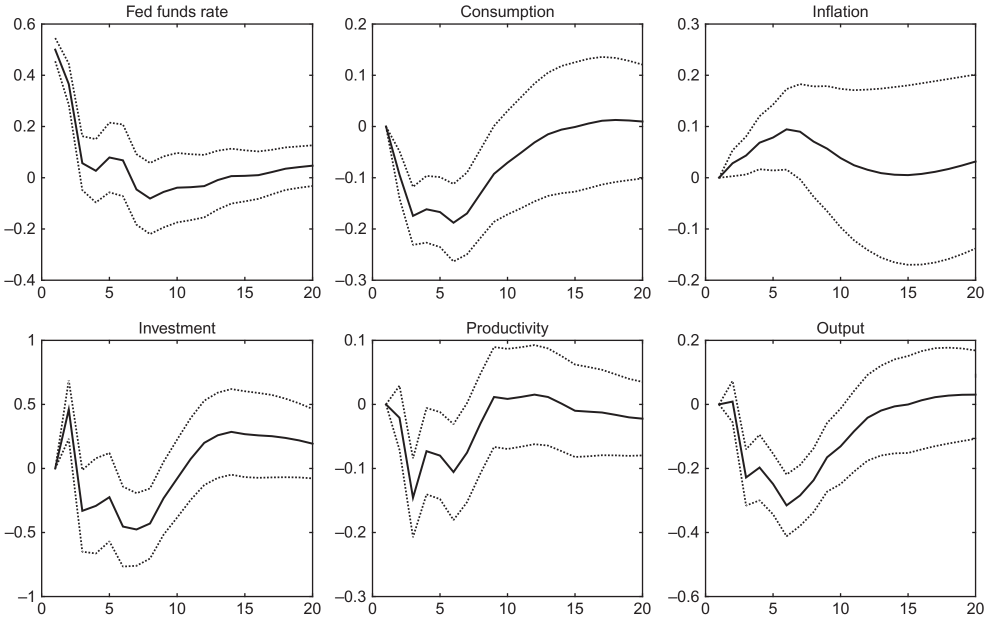


Fig. 2 Impulse responses to an identified monetary contraction: Standard variables.

[ah] In line with our findings, in her chapter in this *Handbook*, Ramey (2016) also emphasizes that the original CEE estimates "do not hold up well in later samples." She concludes that "the most likely reason for the breakdown [...] in the later sample is simply that we can no longer identify monetary policy shocks well."

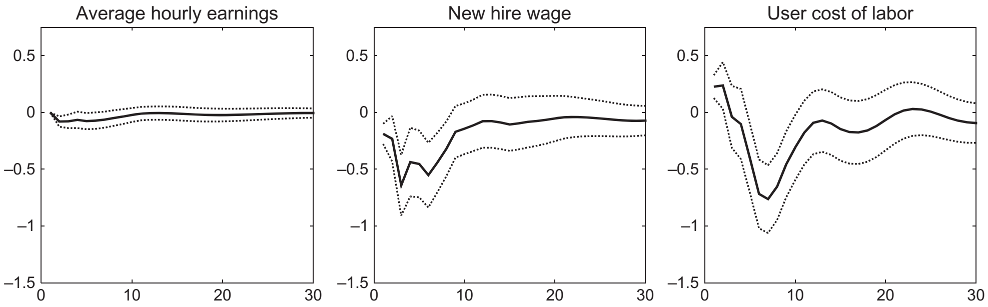


Fig. 3 Impulse responses to an identified monetary contraction: Real wage measures.

in investment of nearly 0.5%. Note that measured productivity also declines, suggesting that unobserved factor utilization contributes to the decline in production.

Fig. 3 reports the impulse responses for the three wage measures we constructed above. The left side panel reports the reaction of average hourly earnings (AHE-BLS). This wage measure barely reacts to the shock. In the center panel, the new-hire wage falls by substantially more. After roughly a year and a half, the new-hire wage has fallen by more than 0.5%. The right side panel shows the user cost of labor. The UCL falls even more than the new-hire wage and remains relatively low even more than 2 years after the shock.

We found similar results when estimating impulse response functions over the time period 1979:4–2007:4. The beginning of this alternative sample corresponds to the beginning of Paul Volcker's chairmanship of the Federal Reserve, but also has the benefit of excluding any backward projection of the user cost series. In general, the impulse responses from the main block of variables are more muted but take longer to return to trend. Despite this difference, the user cost series still has a peak response near 0.75%. The new-hire wages oscillate rapidly, but reach a similar peak response after a similar lag. We conclude that the results are qualitatively unchanged over this shorter sample period that overlaps significantly with Kudlyak's data sample and is also the period when the "modern" era of US monetary policy may be said to have begun.

5.5 Extension: Controlling for Match Quality

As discussed in Section 3, Hagedorn and Manovskii (2013) argue that much of the observed history dependence of current wages can be understood by appealing to labor search when workers face a job ladder. In Hagedorn and Manovskii's model, the match quality of a job can be proxied by including the cumulative labor market "tightness" during the employment cycle in the wage regression. Labor market tightness is the ratio of vacancies to unemployment. Intuitively, during an employment cycle, a worker gradually climbs up the match-quality ladder. How fast he or she climbs is determined by current aggregate labor-market tightness. Ultimately, how high the person gets is given by *cumulative* labor-market tightness over the employment cycle. In this section, we extend the results above to include Hagedorn and Manovskii's proposed measure of labor market tightness, to see whether the results we have reported are robust to the inclusion of this variable.

To implement Hagedorn and Manovskii's proposed correction, we use the NLSY's weekly arrays to classify each respondent's work history into employment cycles. An employment cycle begins when a person finds a job and exits involuntary unemployment. The employment cycle spans the full length of time employed, even if a worker switches employers, as well as *voluntary* spells of unemployment. The employment cycle ends through involuntary unemployment or voluntary unemployment that turns involuntary if the person cannot find a job within 2 months of voluntarily entering unemployment. The NLSY survey asks individuals why they left their last job, and we use this information to determine whether unemployment is voluntary or involuntary.[ai]

We then calculate the sum of labor market tightness for each job cycle for each individual and include the resulting variable in the individual wage regression (12) as an additional control.[aj,ak] Formally, let $\xi_t = \frac{v_t}{u_t}$ denote labor market tightness at date t. Then, for an individual i, currently in an employment spell that began at date $J(i)$, we calculate the sum of the individual's labor market tightness as $\omega_t^i = \sum_{s=t-J(i)}^{t} \xi_s$. We then reestimate (12) and include ω_t^i in the vector of controls X_t^i.

Finally, we modify the prediction equation by assuming that firms hire an individual with average characteristics (as before) and with a fixed average duration of an employment cycle in the NLSY, $\bar{J} = 3.24$ years. That is, we form the projections $\widehat{\ln w_{t,\tau}}$ by including the variable $\bar{x}_t = \sum_{s=t-\bar{J}}^{t} \xi_s$ in the estimated equation (13). Unlike the variable x_t^i, which varies across workers depending on when their employment spell began, the variable $\bar{x}_t$ exhibits no cross-sectional variation. However, since aggregate labor market tightness ξ_t changes over time, $\bar{x}_t$ has time-series variation which is included in the forecasts of $\widehat{\ln w_{t,\tau}}$.

[ai] A person leaves a job involuntarily if he or she is fired, laid off or if the plant closes. If the person voluntarily quits to look for a new job (etc.) and finds a new job within 2 months, the employment cycle is assumed to continue. If the person voluntarily leaves but it takes longer than 2 months to find a job, the employment cycle ends and the person falls off the job ladder.

[aj] To construct our measure of labor market tightness, we use the help wanted index calculated by Barnichon (2010).

[ak] Hagedorn and Manovskii actually use two separate controls. They control for cumulative labor market tightness both during a job spell (the variable q^{HM} in their 2013 paper) and also prior to starting a job if the worker either makes a job-to-job transition or starts from a period of voluntary unemployment (the variable q^{EH}). Because (12) includes an arbitrary set of fixed effects $\chi_{\tau,t}$ for the current job spell, the first adjustment (the q^{HM} variable) is already included in our baseline specification. Thus in our analysis we confine our attention to the second adjustment, which controls only for cumulative labor market tightness *prior* to the start of a job.

Table 4 Real wage cyclicality controlling for match quality: Unemployment rate

	AHE–BLS	AHE–NLSY			New hire	UCL
	(1)	Base (2)	Controls (3)	Controls, FEs (4)	(5)	(6)
HP-filtered unemployment rate	−0.507 (0.471)	−1.039 (1.833)	−1.092 (1.729)	−1.294 (1.764)	−0.691 (1.851)	−4.773 (2.049)
Observations	34	25	25	25	25	27

Notes: OLS standard errors are in parentheses. Coefficients are multiplied by 100.

Table 5 Real wage cyclicality controlling for match quality: GDP

	AHE–BLS	AHE–NLSY			New hire	UCL
	(1)	Base (2)	Controls (3)	Controls, FEs (4)	(5)	(6)
HP-filtered GDP	0.311 (0.353)	1.000 (1.682)	0.844 (1.568)	1.069 (1.631)	1.244 (1.311)	2.284 (1.336)
Observations	34	25	25	25	25	27

Notes: OLS standard errors are in parentheses.

Tables 4 and 5 report results for the cyclicality of the various wage measures after we control for unobserved idiosyncratic match quality. Notice that the cyclicality estimates do decline somewhat, though the overall cyclicality of the wage series is only moderately changed. In particular, the high cyclicality of the new-hire wage and the user cost of labor remain. Similar results are found in response to monetary shocks. Repeating the steps above for the wage measures including the control for match quality gives impulse response functions that are close to the response functions we saw earlier. Fig. 4 compares the impulse response of the new-hire wage and the user cost of labor with and without the control for match quality. As shown in the figure, the impulse responses are almost indistinguishable.

6. COMPARING THE MODEL AND THE DATA

We now relate the empirical evidence we have presented regarding the cyclicality of real wages to the predictions of New Keynesian models of the class developed in Section 4. To build intuition, we begin by examining the models without either composition bias or infrequent renegotiation of wage remittances. This requires that we set the renegotiation rate s and the ratio of effective labor to total hours L/N both to 1.00, and the inverse

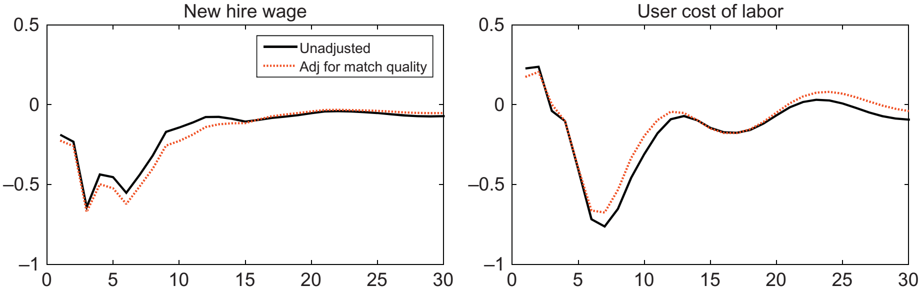


Fig. 4 Impulse responses controlling for match quality.

density of types $\varphi(1)^{-1} = 0$. Since s is the renegotiation rate of long-term wage contracts, $s = 1$ implies that remitted wages are changed in every period and thus move in lock step with changes in the (sticky) allocative wage. Since the remitted wage is equal to the allocative wage in this specification, there are no implicit wage contracts in the economy. The assumption of no composition bias implies that the marginal and average worker supplies the same number of efficiency units of labor per observed hour of work. We examine the responses of this baseline New Keynesian model to a monetary shock under the assumption of sticky prices only, sticky wages only, and both sticky prices and wages. The results will guide us regarding the features to add to the model in order to bring the model's predictions closer to the observed wage data.

We set the Calvo parameter for price adjustment to $\theta_p = 0.90$ (quarterly) implying that prices adjust on average once every 10 quarters, or once every two and a half years. We do the same for the initial sticky wage specification, so $\theta_w = 0.90$. While these calibrations are somewhat high relative to the micro data on the average frequency of price and wage changes, they are in line with many DSGE estimates and the implied model impulse response functions have enough persistence for their computed trajectories to be comparable with the empirical impulse responses. The DSGE model also features traditional mechanisms considered by business cycle analysts to better match the dynamic effects of monetary shocks on output. Specifically, the model features investment adjustment costs, habit persistence in consumption, variable capital utilization, price and wage indexation, and increasing returns to scale in production. The parameter values used are reported in Table 6, and are roughly in line with prevailing estimates in the literature.

Fig. 5 shows impulse responses of this baseline model to a 25 basis point shock to the central bank's policy rate (a shock to the Taylor rule). Since our main interest is in comparing the model responses of different wage measures to the corresponding empirical responses, we show the model and data responses of average hourly earnings (AHE), the new-hire wage (NHW) and the user cost of labor (UCL), as well as output. We reproduce the data responses that appeared in Fig. 3 together with responses from the baseline model with only price rigidity, only wage rigidity, and with both types of nominal inertia. We find that, as one might expect, following the increase in the interest rate,

Table 6 Parameters for New Keynesian DSGE model

Parameter	**Value**
Discount factor, annual (β)	0.97
Intertemporal elasticity of substitution (σ)	1.00
Frisch labor supply elasticity (η)	1.00
Depreciation rate, annual (δ)	0.10
Capital's share (α)	0.36
Type-specific labor elasticity (ψ)	21.00
Type-specific product elasticity (ε)	6.00
Average duration of prices, quarterly ($(1-\theta_p)^{-1}$)	10.00
Average duration of wages, quarterly ($(1-\theta_w)^{-1}$)	10.00
Inflation indexing	Yes
Marginal cost of capital utilization ($b''(1)/b'(1)$)	1.00
Investment adjustment cost (κ)	4.00
Habit weight (h)	0.65
Ratio of effective labor to total hours (LN)	2.00
Inverse density at unit productivity ($\varphi(1)^{-1}$)	2.00
Renegotiation rate, quarterly (s)	0.21

all three types of nominal rigidity cause output to decline. Output falls significantly more in the model with only sticky prices than in the model with just sticky wages; of course, it falls further in the model with both nominal fractions. Although the baseline models reproduce the hump-shaped output response observed in the data, the trough of output comes 2–3 quarters earlier in the models than in the data. That is, the models need additional persistence mechanisms or stronger persistence mechanisms to match the estimates.

One of the main findings in the empirical section is that different wage measures behave differently over the business cycle and in response to monetary shocks. In the baseline model, since there is just a single wage (or more precisely, since the average wage, the new-hire wage, and the allocative wage are all identical), the model is completely incapable of matching the differential patterns of wages in the data. We see that the impulse responses for the three concepts of the wage are identical in each model. As usual, the UCL in the model is the allocative wage, but with constant wage negotiation and no composition bias, AHE and the NHW are identical to the UCL. Thus, in this set of models there is a single unambiguous wage response to a monetary shock.

In the sticky-price model with flexible wages, the wage declines sharply (thus, it is "procyclical" in the sense that it moves in the same direction as output does following the monetary shock). The wage decline qualitatively matches the responses of the NHW and UCL, but does not match the fact that AHE responds much less. On the other hand, the model with sticky wages and flexible prices shows a mild *countercyclical* response of the wage to a monetary shock, for the same reasons that Keynes's (1936) model in the *General Theory* predicted high real wages in recessions. Finally, note that the model with equal

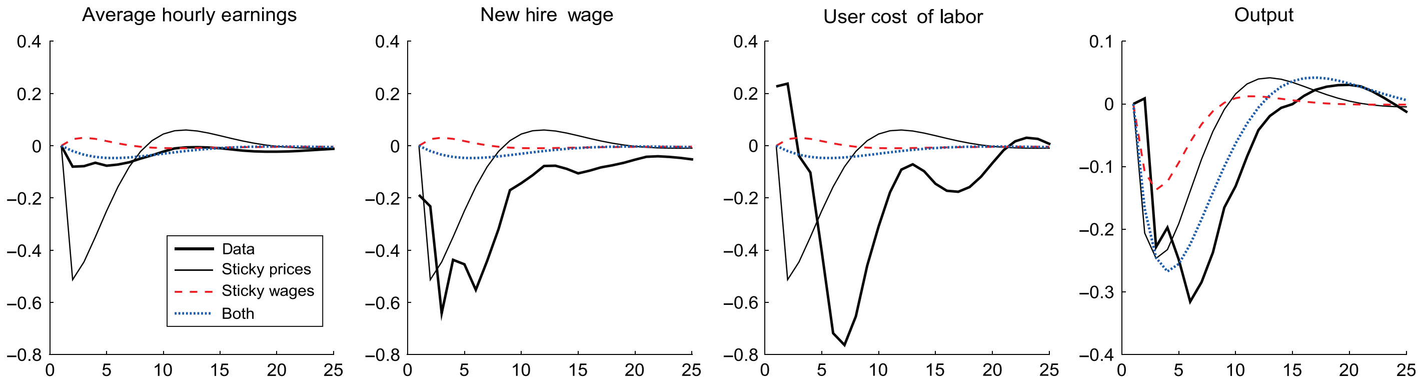


Fig. 5 Wage dynamics in baseline New Keynesian models. *Notes:* Each panel reports the estimated impulse responses (heavy line) and model impulse responses to a 25 basis point shock to the federal funds rate. The thin solid line displays the response of the baseline New Keynesian model with sticky prices but flexible wages. The dashed line displays the response of the model with sticky wages but flexible prices, and the dotted line displays the responses of the model with both sticky prices and sticky wages.

(and high) price and wage rigidity shows that the real wage is basically acyclical in the wake of a monetary policy shock. One might summarize the three models by saying that in the model with price rigidity firms are off their labor demand curves but workers are on their labor supply curves, so real wages are procyclical. The situation is reversed in the model with sticky wages, so real wages are countercyclical. (However, the assumption of variable capital utilization flattens the labor demand curve significantly, so the degree of countercyclicality is modest.) Finally, with both sticky wages and prices, both workers and firms are off their notional supply and demand curves in the labor market, and the real wage has no clear cyclical pattern. Note that the wage response in this variant of the model is qualitatively consistent with the empirical response of AHE. Hence, it is clear why modelers who interpret AHE as the allocative wage have been led towards models with both wage and price rigidity, as in CEE.

Starting from the baseline model above, we now consider the effects of implicit contracting and composition bias on model predictions for our three wage measures. In the following discussion, we consider models with sticky prices and flexible wages only. We do this both to conserve space and also because the sticky-wage models typically have simulated wage paths that are either sharply counterfactual (ie, there are sharp increases in wage payments following a negative monetary shock) or wage paths that are acyclical which, while matching the observed behavior of average hourly earnings, fail to match the responses of new-hire wages and the UCL.

We begin by examining the role of implicit contracting. Starting with the baseline model above, we consider the effects of gradually reducing the parameter s from its initial value of 1.00. When $s < 1$, the *remitted* wage is changed infrequently even though the allocative wage (the UCL) is fully flexible, since in this model we assume no wage rigidity. While the UCL is free to react to changing economic conditions, other measures of the wage—AHE and the NHW—change by substantially less than the UCL. The results are shown in Fig. 6. Note that the results for $s = 1$ reproduce the sticky-price impulse responses of the previous figure. As discussed in Section 3, Barattieri et al. (2014) find that $s = 0.21$ is the approximate frequency of changes in remitted wages observed in micro wage data. We include $s = 0.50$ as an intermediate case. Note that the three impulse responses are identical for the UCL—the allocative wage is unaffected by the value of s. However, $s < 0$ implies the existence of implicit contracts, which makes the three wage measures differ in their responses to monetary policy shocks. Particularly interesting is the result for the measured value of $s = 0.21$. For this value of s, the allocative wage falls sharply, the wage for new hires falls less, and average hourly earnings fall only slightly. The pattern of wage responses for the three wage measures relative to the output response bears a strong qualitative resemblance to the empirical impulse responses. This observation leads us to conclude tentatively that the evidence suggests that a model with sticky prices, flexible wages, and a significant role for implicit wage contracts has the best chance of matching the data.

Fig. 7 shows the impulse response function for the model with composition bias effects. Endogenous composition adjustment has two separate effects on the responses

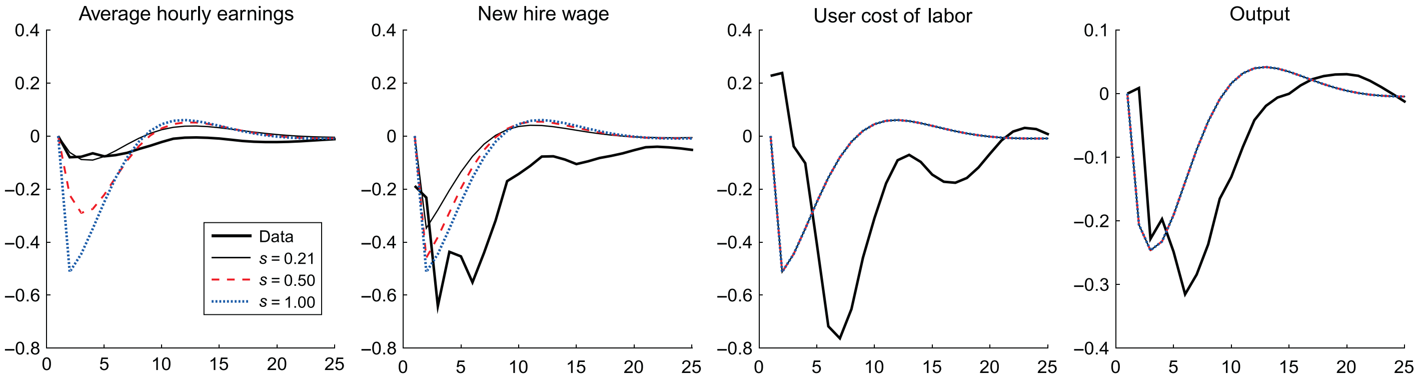


Fig. 6 Implicit contracts in sticky price models. *Notes:* Each panel reports the estimated impulse responses and model impulse responses to a 25 basis point shock to the federal funds rate. The different lines correspond to different parameter values. In all cases, prices are sticky but wages are flexible. The thin solid line displays the responses of the baseline model when remitted wages are reset with a quarterly probability (s) of 0.21 (roughly every 15 months). The dashed line displays results for the model with $s = 0.50$ and the dotted line displays results when $s = 1.00$ (continual adjustment of the remitted wage).

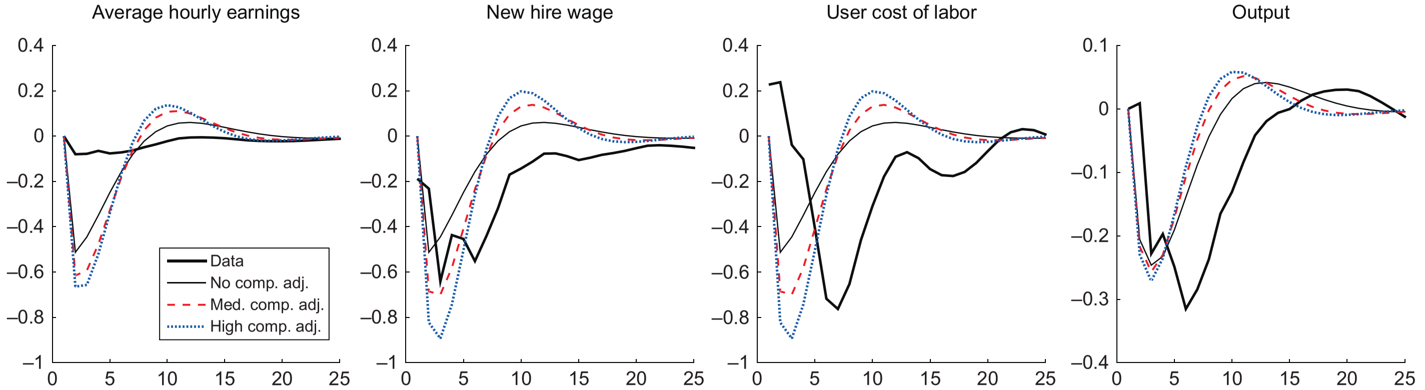


Fig. 7 Wage dynamics and composition bias in sticky price models. *Notes:* Each panel reports the estimated impulse responses and model impulse responses to a 25 basis point shock to the federal funds rate. The different lines correspond to different parameter values. In all cases, prices are sticky but wages are flexible. The thin solid line displays the responses of the model with no composition bias ($LN = 1.00$ and $\varphi^{-1}(1) = 0.0001$). The dashed line displays results for an intermediate level composition bias ($LN = 2.00$ and $\varphi^{-1}(1) = 2.00$). The dotted line displays results for high composition bias ($LN = 4.00$ and $\varphi^{-1}(1) = 4.00$).

of wages to monetary (and nonmonetary) shocks. First, by introducing a difference between the average labor compensation of employed workers and the labor compensation for the "marginal" workers, composition adjustments cause average hourly earnings to be less responsive than the user cost of labor which holds labor force composition fixed. The magnitude of this differential is given by the ratio $(LN-1)/LN$ in equation (10) where $LN \geq 1$ is the steady-state ratio of the effective labor supply to hours worked or equivalently, LN is the ratio of average labor compensation to wages paid to marginal workers. In a model without endogenous composition adjustment, this ratio is 1 and there are no effects of composition bias on measured wages. If $LN > 1$ then average wages move by less than the user cost.

The second effect of compositional changes is that the effective labor supply elasticity in such an environment is strictly less than the individual labor supply elasticities. The reason for this is that expanding employment means hiring workers who are increasingly less productive. The magnitude of this effect is governed by the inverse density of types at the productivity cutoff $\varphi(1)^{-1}$. In a typical model in which all workers are the same, $\varphi(1)=\infty$ (ie, there is a mass point at the common productivity 1) and thus $\varphi(1)^{-1}=0$. If the density of types at the cutoff is smooth, however, $\varphi(1)^{-1}$ is greater than 0 indicating that hiring more workers requires lowering the marginal productivity. If $\varphi(1)^{-1}$ is large then expanding the workforce requires tolerating much lower productivity workers and thus the effective labor supply elasticity is substantially lower.

We consider three different model specifications in Fig. 7. First we report the response for the standard model (the thin solid line) without composition bias. The dashed line reports the impulse responses for a model with an "intermediate" degree of composition bias. For this specification we consider a case with $LN=2.0$ (so the average worker is paid twice as much as the marginal worker) and $\varphi^{-1}(1)=2.00$. The dotted line reports the responses for a "high" degree of composition bias in which $LN=\varphi^{-1}(1)=4.00$. The figure displays both of the effects mentioned above. Notice in particular that the specifications with composition bias feature notably sharper reductions in wages. This is because, in the sticky-price environment, output, and thus labor, is effectively demand determined. Given demand, the firms simply hire or fire as many workers as necessary to increase or decrease production. Since the effective labor supply elasticity is reduced by the compositional adjustments, the wages must fall by more. Also, notice that the reduction in average hourly earnings is less than the decline in the user cost. This is a direct consequence of $LN > 1$: exactly the effect highlighted in Solon et al. (1994).

Fig. 8 considers the baseline model with both implicit wage smoothing contracts and a modest amount of composition bias. For this simulation, we set $s=0.21$ as suggested by Barattieri et al. (2014) and we adopt the intermediate composition bias specification, $LN=\varphi(1)^{-1}=2.0$.

As we did in Fig. 5, Fig. 8 shows the impulse responses under the assumption of pure sticky prices (solid line), pure sticky wages (dashed line) and a specification with both

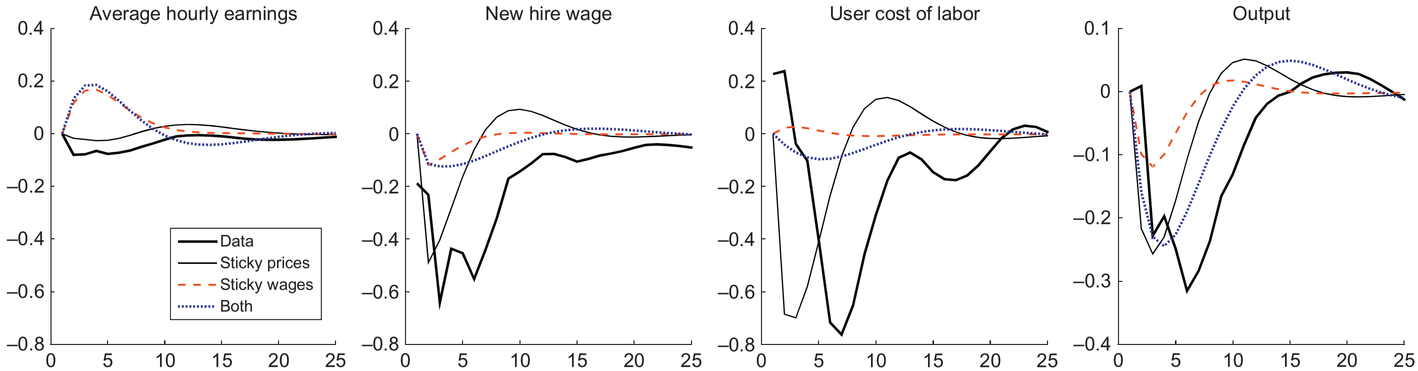


Fig. 8 Real wage dynamics in the modified model. *Notes:* Each panel reports the estimated impulse responses (heavy line) and model impulse responses to a 25 basis point shock to the federal funds rate. The parameter values for the model are set according to Table 6. We use $s = 0.21$ and the intermediate level of composition adjustment ($LN = 2.00$ and $\varphi^{-1}(1) = 2.00$). The thin solid line displays the response of a model with sticky prices but flexible wages. The dashed line displays the response of a model with sticky wages but flexible prices, and the dotted line displays the responses of a model with both sticky prices and sticky wages.

sticky prices and sticky wages (dotted line). Based on this simulation, the sticky price specification seems to outperform both of the other cases. In the sticky-wage case and the specification with both types of nominal rigidities, average hourly earnings rise noticeably following the monetary contraction, while the sticky price model is surprisingly close to the actual point estimates from the SVAR impulse response, which show a slight decline. While still not matching the shape of the dynamic responses to the new-hire wage or the user cost of labor, the model responses are quantitatively close. The model predicts a maximum decline in the new-hire wage of roughly 0.50% and a maximum decline in the user cost of labor of 0.70%. In comparison, the SVAR estimates a decline in the new-hire wage of roughly 0.70% and a decline in the user cost of almost 0.80%. The simulated responses for the other two model specifications display much smaller movements of either the new-hire wage or the user cost (in all cases, the peak declines are less than 0.10%).

The main problem with the impulse responses of the sticky-price model relative to the data is the lack of persistence of output following a monetary contraction. Output in the model attains its maximum response three quarters after the shock, as opposed to six quarters in the data. The new hire wage and the user cost of labor hit their troughs at the same time as output, meaning that they display a phase shift relative to the empirical impulse responses. Thus, we conjecture that a more persistent behavior of real variables after a monetary shock would bring the model and the data into closer alignment.

At the same time, it is easy to see why the model lacks persistence. Models such as CEE use sticky wages as an important persistence mechanism, but as Fig. 8 shows, allocative wage stickiness is inconsistent with the observed behavior of the UCL following a monetary contraction. Thus, our preferred model specification with sticky prices and flexible wages is missing one of the key propagation mechanisms featured in many standard medium-scale DSGE models in the literature. An important agenda for future research is to find new propagation mechanisms for sticky-price models to replace the assumption of sticky allocative wages, which seems to be fundamentally inconsistent with the data.

7. CONCLUSION

Recent empirical studies suggest that the cyclicality of real wages is greater than conventional wisdom would suggest. The literature emphasizes two reasons for this enhanced cyclicality. First, endogenous changes in the composition of the workforce mechanically causes average hourly wage payments to understate the change in wages relative to wage changes holding workforce composition fixed. Second, there are indications that the allocative wage—the wage that governs hours worked and that firms internalize when making production and pricing decisions—may not equal the contemporaneous remitted wage. In particular, firms and workers may well have an implicit understanding that the remitted wage will be a smoothed version of the expected allocative wage. By estimating the expected present value of wage payments, one can construct a "user cost of labor," which should measure the underlying allocative wage.

In this chapter, we have reproduced and extended the key empirical results in Kudlyak's (2014) work. Our empirical analysis confirms her calculations for the cyclicality of the allocative wage. In addition, the NLSY also allows us to decompose the cyclical response of wages by controlling separately for compositional changes and at the same time controlling for the wage smoothing effects of implicit contracts. The data suggest that while compositional changes contribute significantly to the dynamics of average hourly earnings, the effects of implicit contracts and wage smoothing are even greater.

Using the annual estimates of the user cost of labor and the new-hire wages from the NLSY as a starting point, we extend the estimated series to a quarterly series and include the extended data in a structural VAR for the purposes of studying the reactions of real wages to monetary shocks. The estimated structural VAR suggests that the user cost of labor and new-hire wages both decline sharply following a contractionary monetary shock. In contrast, average hourly earnings—the usual measure of the wage in macroeconomic research—barely respond to such shocks. Our model, if extended to allow for wage smoothing within long-term worker-firm associations, can match the fact that average hourly earnings and even the wage paid to new hires are significantly less cyclical than the allocative user cost of labor.

The differential reactions of these wage measures present two key challenges for prevailing New Keynesian models of the monetary transmission mechanism. First, in most New Keynesian models, there is no conceptual difference between the allocative wage, the remitted wage, and average hourly earnings. Thus, at a basic level, these models will not be able to match the empirical findings we study. Second, to the extent that New Keynesian models include a prominent role for sticky nominal wages, the models typically predict that either the allocative wage will counterfactually rise in the wake of an increase in the policy rate or that wages will not respond at all. Neither of these predictions would match our basic finding that allocative wages appear to fall sharply after a monetary tightening.

Analysis of a medium-scale DSGE model suggests that successful models will emphasize price rigidity rather than wage rigidity. In addition, to match the estimates in this chapter, such models will likely allow for a relatively smooth remitted wage which is not allocative. The wage data, therefore, favor the "old New Keynesian economics"—the early models of Rotemberg (1982), Ball and Romer (1989), and Kimball (1995), which all assumed competitive labor markets and flexible allocative wages—rather than the "new New Keynesian economics" with allocative prices and wages both sticky, as in Blanchard and Kiyotaki (1987), Erceg et al. (2000), Smets and Wouters (2007), and CEE.

We conclude by suggesting that research is needed on two fronts, one theoretical and one empirical. The theoretical challenge arising from these results is the need to explain the observed persistence of the real effects of monetary shocks without being able to rely on wage stickiness to temper the response of marginal cost to a monetary shock. Our estimates suggest that real marginal cost, properly computed, is strongly procyclical.[al] This conclusion, if correct, casts doubt on the main persistence mechanism of medium-scale New Keynesian models, such as CEE and Smets and Wouters (2007), which generally rely on assumptions that make marginal cost acyclical. Thus, the problem facing monetary economics can be restated as: Why do prices behave sluggishly, even though wages and hence marginal costs are strongly procyclical?

This question was the focus of a major research program several decades earlier. Relative to the research that took place in the 1980s, new observations on firm-level prices and quantities, and the desire to have persistence mechanisms that are consistent with both time-dependent and state-dependent pricing models, impose additional constraints on the proposed solutions.[am] To use the language of Ball and Romer (1990), the search for "real rigidities" in price setting must arrive at a satisfactory conclusion in order to make models of the monetary transmission mechanism consistent with recent observations of the data.

The empirical challenge is to extend the measurement of real wage cyclicality to other data sets and other countries. It would be particularly interesting to compare the results reported here with similar calculations for the major continental European economies, or for the Euro area as a whole. The labor market is one area where economists have argued that the differences between the United States and Europe are most pronounced. Galí (2016) follows in this tradition by incorporating "hysteresis" effects into a standard New Keynesian framework, arguing that this modification is necessary to match the greater persistence of the unemployment rate in Europe. Yet this change, where the unemployed exert little downward pressure on wages, makes real wages even less sensitive to the business cycle than they are in the standard New Keynesian models developed to explain US macro data. Do European micro data indicate that there is enormously more allocative wage rigidity in Europe than in the United States? Finding out whether the answer is yes or no is clearly of first-order importance for understanding cyclical fluctuations in these two major economies.

[al] This agrees with some, although not all, of the literature on the countercyclicality of the price markup, which is of course just the inverse of real marginal cost. See for example, Rotemberg and Woodford (1999) and Bils et al. (2014).

[am] See, for example, Klenow and Willis (2016), Dotsey and King (2005), and Nakamura and Steinsson (2010). Leahy (2011) and Nakamura and Steinsson (2013, section 12) provide insightful discussions based on Ball and Romer (1990).

ACKNOWLEDGMENTS

We thank Tyler Atkinson and, especially, Rich Ryan for outstanding research assistance. We thank the editors, John Taylor and Harald Uhlig. We thank our formal discussants, Eric Sims and Andreas Westermark, for their insights. We received many useful comments and suggestions from seminar participants at the University of Michigan, the Handbook of Macroeconomics conference at the Becker-Friedman Institute, University of Chicago, Boston College, and the conference on New Developments in the Macroeconomics of Labor Markets at CREI, Pompeu Fabra University.

REFERENCES

Abraham, K.G., Haltiwanger, J.C., 1995. Real wages and the business cycle. J. Econ. Lit. 33 (3), 1215–1264.

Akerlof, G., Dickens, W.R., Perry, G., 1996. The macroeconomics of low inflation. Brook. Pap. Econ. Act. 27 (1), 1–76.

Ball, L., Romer, D., 1989. The equilibrium and optimal timing of price changes. Rev. Econ. Stud. 56 (2), 179–198.

Ball, L., Romer, D., 1990. Real rigidities and the non-neutrality of money. Rev. Econ. Stud. 57 (2), 183–203.

Barattieri, A., Basu, S., Gottschalk, P., 2014. Some evidence on the importance of sticky wages. Am. Econ. J. Macroecon. 6 (1), 70–101.

Barnichon, R., 2010. Building a composite help-wanted index. Econ. Lett. 109 (3), 175–178.

Barro, R.J., 1977. Long-term contracting, sticky prices, and monetary policy. J. Monet. Econ. 3 (3), 305–316.

Barro, R.J., King, R.G., 1984. Time-separable preferences and intertemporal-substitution models of business cycles. Q. J. Econ. 99 (4), 817–839.

Basu, S., Bundick, B., 2012. Uncertainty shocks in a model of effective demand. Working Paper 18420. National Bureau of Economic Research.

Basu, S., Kimball, M.S., 1997. Cyclical productivity with unobserved input variation. Working Paper 5915. National Bureau of Economic Research.

Basu, S., Taylor, A.M., 1999. Business cycles in international historical perspective. J. Econ. Perspect. 13 (2), 45–68.

Beaudry, P., DiNardo, J., 1991. The effect of implicit contracts on the movement of wages over the business cycle: evidence from micro data. J. Polit. Econ. 99 (4), 665–688.

Beaudry, P., DiNardo, J., 1995. Is the behavior of hours worked consistent with implicit contract theory? Q. J. Econ. 110 (3), 743–768.

Becker, G.S., 1962. Investment in human capital: a theoretical analysis. J. Polit. Econ. 70 (5), 9–49.

Bernanke, B.S., Blinder, A.S., 1992. The Federal funds rate and the channels of monetary transmission. Am. Econ. Rev. 82 (4), 901–921.

Bernanke, B.S., Carey, K., 1996. Nominal wage stickiness and aggregate supply in the great depression. Q. J. Econ. 111 (3), 853–883.

Bewley, T.F., 1999. Why Wages Don't Fall During a Recession. Harvard University Press, Cambridge, MA.

Bils, M., 1985. Real wages over the business cycle: evidence from panel data. J. Polit. Econ. 93 (4), 666–689.

Bils, M., Klenow, P.J., 2004. Some evidence on the importance of sticky prices. J. Polit. Econ. 112 (5), 947–985.

Bils, M., Klenow, P.J., Malin, B.A., 2014. Resurrecting the role of the product market wedge in recessions. Working Paper 20555. National Bureau of Economic Research.

Blanchard, O.J., Kiyotaki, N., 1987. Monopolistic competition and the effects of aggregate demand. Am. Econ. Rev. 77 (4), 647–666.

Card, D., Hyslop, D., 1997. Does inflation "grease the wheels of the labor market"? In: Romer, C.D., Romer, D.H. (Eds.), Reducing Inflation: Motivation and Strategy. University of Chicago Press, Chicago.

Christiano, L.J., Eichenbaum, M., Evans, C.L., 1999. Monetary policy shocks: what have we learned and to what end? In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics, vol. 1, (Chapter 2). Elsevier, Amsterdam, Netherlands, pp. 65–148.

Christiano, L.J., Eichenbaum, M., Evans, C.L., 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. J. Polit. Econ. 113 (1), 1–45.

Daly, M.C., Hobijn, B., 2016. The intensive and extensive margins of real wage adjustment. Federal Reserve Bank of San Francisco Working Paper 2016-04. Available from http://www.frbsf.org/economic-research/files/wp2016-04.pdf.

Dotsey, M., King, R.G., 2005. Implications of state-dependent pricing for dynamic macroeconomic models. J. Monet. Econ. 52 (1), 213–242.

Dunlop, J.T., 1938. The movement of real and money wage rates. Econ. J. 48 (191), 413–434.

Eichengreen, B., Sachs, J., 1985. Exchange rates and economic recovery in the 1930s. J. Econ. Hist. 45 (4), 925–946.

Elsby, M.W.L., 2009. Evaluating the economic significance of downward nominal wage rigidity. J. Monet. Econ. 56 (2), 154–169.

Elsby, M.W.L., Shin, D., Solon, G., 2016. Wage adjustment in the great recession and other downturns: evidence from the United States and Great Britain. J. Labor Econ. 34 (S1), S249–S291.

Erceg, C.J., Henderson, D.W., Levin, A.T., 2000. Optimal monetary policy with staggered wage and price contracts. J. Monet. Econ. 46 (2), 281–313.

Fleischman, C.A., 1999. The Causes of Business Cycles and the Cyclicality of Real Wages. Finance and Economics Discussion Series 1999–53. Board of Governors of the Federal Reserve System.

Friedman, M., Schwartz, A.J., 1963. A Monetary History of the United States. Princeton University Press, Princeton.

Galí, J., 2013. Notes for a new guide to Keynes (I): wages, aggregate demand, and employment. J. Eur. Econ. Assoc. 11 (5), 973–1003.

Galí, J., 2016. Insider-outsider labor markets, hysteresis and monetary policy. Working paper. Available from http://crei.cat/people/gali/mp_hysteresis_jan2016.pdf.

Galí, J., Gertler, M., 1999. Inflation dynamics: a structural econometric analysis. J. Monet. Econ. 44 (2), 195–222.

Galí, J., Gertler, M., López-Salido, J.D., 2007. Markups, gaps, and the welfare costs of business fluctuations. Rev. Econ. Stat. 89 (1), 44–59.

Geary, P.T., Kennan, J., 1982. The employment-real wage relationship: an international study. J. Polit. Econ. 90 (4), 854–871.

Gertler, M., Trigari, A., 2009. Unemployment fluctuations with staggered nash wage bargaining. J. Polit. Econ. 117 (1), 38–86.

Gottschalk, P., 2005. Downward nominal-wage flexibility: real or measurement error? Rev. Econ. Stat. 87 (3), 556–568.

Gottschalk, P., Huynh, M., 2010. Are earnings inequality and mobility overstated? The impact of nonclassical measurement error. Rev. Econ. Stat. 92 (2), 302–315.

Greenwood, J., Hercowitz, Z., Huffman, G.W., 1988. Investment, capacity utilization, and the real business cycle. Am. Econ. Rev. 78 (3), 402–417.

Haefke, C., Sonntag, M., van Rens, T., 2013. Wage rigidity and job creation. J. Monet. Econ. 60 (8), 887–899.

Hagedorn, M., Manovskii, I., 2013. Job selection and wages over the business cycle. Am. Econ. Rev. 103 (2), 771–803.

Hall, R.E., 2005. Employment fluctuations with equilibrium wage stickiness. Am. Econ. Rev. 95 (1), 50–65.

Hall, R.E., Milgrom, P.R., 2008. The limited influence of unemployment on the wage bargain. Am. Econ. Rev. 98 (4), 1653–1674.

Hanes, C., 1993. The development of nominal wage rigidity in the late 19th century. Am. Econ. Rev. 83 (4), 732–756.

Hanes, C., 1996. Changes in the cyclical behavior of real wage rates, 1870–1990. J. Econ. Hist. 56 (4), 837–861.

Hanes, C., James, J.A., 2003. Wage adjustment under low inflation: evidence from U.S. history. Am. Econ. Rev. 93 (4), 1414–1424.

Hansen, G.D., 1985. Indivisible labor and the business cycle. J. Monet. Econ. 16 (3), 309–327.

Hanson, M.S., 2004. The "price puzzle" reconsidered. J. Monet. Econ. 51 (7), 1385–1413.

Huang, K.X.D., Liu, Z., Phaneuf, L., 2004. Why does the cyclical behavior of real wages change over time? Am. Econ. Rev. 94 (4), 836–856.

Hume, D., (1742) 1987. Essays, Moral, Political, and Literary. Library of Economics and Liberty, Indianapolis, IN.

Kahn, S., 1997. Evidence of nominal wage stickiness from microdata. Am. Econ. Rev. 87 (5), 993–1008.

Keynes, J.M., 1936. The General Theory of Employment, Interest, and Money. Macmillan, London.

Kimball, M.S., 1995. The quantitative analytics of the basic neomonetarist model. J. Money Credit Bank. 27 (4), 1241–1277.

King, R., Rebelo, S., 1999. Resuscitating real business cycles. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics, first ed., vol. 1, Part B, (Chapter 14). Elsevier, Amsterdam, Netherlands, pp. 927–1007.

Klenow, P.J., Willis, J.L., 2016. Real rigidities and nominal price changes. Working Paper. Available from http://klenow.com/RealRigidities.pdf.

Kudlyak, M., 2014. The cyclicality of the user cost of labor. J. Monet. Econ. 68, 53–67.

Le Bihan, H., Montornès, J., Heckel, T., 2012. Sticky wages: evidence from quarterly microeconomic data. Am. Econ. J. Macroecon. 4 (3), 1–32.

Leahy, J., 2011. A survey of new Keynesian theories of aggregate supply and their relation to industrial organization. J. Money Credit Bank. 43, 87–110.

Lebow, D.E., Saks, R.E., Wilson, B.A., 1999. Downward nominal wage rigidity: evidence from the employment cost index. Finance and Economic Discussion Series Working Paper 1999-31. Available from http://www.federalreserve.gov/pubs/feds/1999/199931/199931pap.pdf.

Lünnemann, P., Wintr, L., 2009. Wages are flexible, aren't they? Evidence from monthly micro wage data. European Central Bank (ECB) Working Paper 1074.

McLaughlin, K.J., 1994. Rigid wages? J. Monet. Econ. 34 (3), 383–414.

Nakamura, E., Steinsson, J., 2010. Monetary non-neutrality in a multisector menu cost model. Q. J. Econ. 125 (3), 961–1013.

Nakamura, E., Steinsson, J., 2013. Price rigidity: microeconomic evidence and macroeconomic implications. Annu. Rev. Econ. 5 (1), 133–163.

Nekarda, C.J., Ramey, V.A., 2013. The cyclical behavior of the price-cost markup. Working Paper 19099. National Bureau of Economic Research.

Pencavel, J., 2015. Keynesian controversies on wages. Econ. J. 125 (583), 295–349.

Pissarides, C.A., 2009. The unemployment volatility puzzle: is wage stickiness the answer? Econometrica 77 (5), 1339–1369.

Ramey, V.A., 2016. Macroeconomic shocks and their propagation. In: Taylor, J.B., Uhlig, H. (Eds.), Handbook of Macroeconomics, vol. 2A. Elsevier, Amsterdam, Netherlands, pp. 71–162.

Ravenna, F., Walsh, C.E., 2008. Vacancies, unemployment, and the phillips curve. Eur. Econ. Rev. 52 (8), 1494–1521.

Rogerson, R., 1988. Indivisible labor, lotteries and equilibrium. J. Monet. Econ. 21 (1), 3–16.

Romer, C.D., Romer, D.H., 1989. Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz, NBER Chapters. In: NBER Macroeconomics Annual 1989. National Bureau of Economic Research, Inc, Volume 4, pp. 121–184.

Romer, C.D., Romer, D.H., 2004. A new measure of monetary shocks: derivation and implications. Am. Econ. Rev. 94 (4), 1055–1084.

Rotemberg, J.J., 1982. Monopolistic price adjustment and aggregate output. Rev. Econ. Stud. 49 (4), 517–531.

Rotemberg, J.J., Woodford, M., 1991. Markups and the business cycle. In: Blanchard, O.J., Fischer, S. (Eds.), NBER Macroeconomics Annual 1991, vol. 6. MIT Press, Cambridge, MA, pp. 63–140.

Rotemberg, J.J., Woodford, M., 1999. The cyclical behavior of prices and costs. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics, vol. 1, Part B, (Chapter 16). Elsevier, Amsterdam, Netherlands, pp. 1051–1135.

Sbordone, A.M., 2002. Prices and unit labor costs: a new test of price stickiness. J. Monet. Econ. 49 (2), 265–292.

Schmitt-Grohé, S., Uribe, M., 2013. Downward nominal wage rigidity and the case for temporary inflation in the eurozone. J. Econ. Perspect. 27 (3), 193–212.

Shimer, R., 2005. The cyclical behavior of equilibrium unemployment and vacancies. Am. Econ. Rev. 95 (1), 25–49.

Sigurdsson, J., Sigurdardottir, R., 2016. Time-dependent or state-dependent wage-setting? Evidence from periods of macroeconomic instability. J. Monet. Econ. 78, 50–66.

Smets, F., Wouters, R., 2007. Shocks and frictions in US business cycles: a Bayesian DSGE approach. Am. Econ. Rev. 97 (3), 586–606.

Solon, G., Barsky, R., Parker, J.A., 1994. Measuring the cyclicality of real wages: how important is composition bias. Q. J. Econ. 109 (1), 1–25.

Solow, R.M., 1979. Another possible source of wage stickiness. J. Macroecon. 1 (1), 79–82.

Stock, J.H., Watson, M.W., 1999. Business cycle fluctuations in US macroeconomic time series. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics, vol. 1, Part A, (Chapter 1). Elsevier, Amsterdam, Netherlands, pp. 3–64.

Stockman, A., 1983. Aggregation bias and the cyclical behavior of real wages (Mimeo.). University of Rochester, Economics Department.

Sumner, S., Silver, S., 1989. Real wages, employment, and the Phillips curve. J. Polit. Econ. 97 (3), 706–720.

Tarshis, L., 1939. Changes in real and money wages. Econ. J. 49 (193), 150–154.

Taylor, J.B., 2016. The staying power of staggered wage and price setting models in macroeconomics. In: Taylor, J.B., Uhlig, H. (Eds.), Handbook of Macroeconomics. vol. 2B. Elsevier, Amsterdam, Netherlands, pp. 2009–2042.

Tobin, J., 1972. Inflation and unemployment. Am. Econ. Rev. 62 (1/2), 1–18.

Walsh, C.E., 2003. Labor market search and monetary shocks. In: Altug, S., Chadha, J.S., Nolan, C. (Eds.), Dynamic Macroeconomic Analysis. Cambridge University Press, Cambridge, UK, pp. 451–486.

Wolpin, K.I., 1992. The determinants of black-white differences in early employment careers: search, layoffs, quits, and endogenous wage growth. J. Polit. Econ. 100 (3), 535–560.

Woodford, M., 2003. Interest and Prices: Foundations of a Theory of Monetary Policy. Princeton University Press, Princeton.

CHAPTER 7

Fiscal and Financial Crises

M.D. Bordo[*,†], **C.M. Meissner**[†,‡]
[*]Rutgers University, New Brunswick, NJ, United States
[†]NBER, Cambridge, MA, United States
[‡]University of California, Davis, CA, United States

Contents

Abstract

Interconnections between banking crises and fiscal crises have a long history. We document the long-run evolution from classic banking panics toward modern banking crises where financial guarantees are associated with crisis resolution. Recent crises feature a feedback loop between bank guarantees and bank holdings of local sovereign debt thereby linking financial to fiscal crises. Earlier examples

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.04.001

include the crises in Chile (early 1980s), Japan (1990), Sweden and Finland (1991), and the Asian crisis (1997). We discuss the evolution in economic theorizing on crises since the 1950s, and then provide an overview of the long-run evolution of connections between different types of crises. Next we explore the empirics of financial crises. We discuss the methodological issue of crisis measurement encompassing the definition, dating, and incidence of financial crises. Leading datasets differ markedly in terms of their historical frequency of crises leading to *classification uncertainty*. There is a range of estimates of output losses from financial crises in the literature, and these are also dependent upon definitions. We find economically significant output losses from various types of crises using a consistent methodology across time and datasets. Predicting crises also remains a challenge. We survey the *Early Warnings Indicators* literature finding that a broad range of variables are potential predictors. Credit booms have been emphasized recently, but other factors still matter. Finally, we identify a new policy trilemma. Countries can have two of the following three choices: a large financial sector, fiscal bailouts devoted to financial crises, and discretionary fiscal policy aimed at raising demand during the recessions induced by financial crises.

Keywords

Banking crises, Currency crises, Fiscal crises, Fiscal resolution, Output losses, Crisis chronologies, Fiscal trilemma, Early warning indicators, Credit boom, Capital flows

JEL Classification Codes

E62, F34, G01, N1

1. INTRODUCTION

The recent financial crisis in the Eurozone involved both sovereign debt and the banking system. The circumstances of this crisis were unique as were the country experiences, but the combined incidence of fiscal and financial crises is actually not new. In fact, these connections have changed progressively over the long run. Recurrent and systemic financial crises emerged as a side effect of the modern process of financial development, globalization, and economic growth which got underway in the early 19th century. Over time, economic theory, economic data, and changes in the objectives of policy makers have shaped the reactions to crises and their subsequent contours. Interconnections between types of financial crises indeed have a long history.

From the mid-19th century, financial crises in the banking sector moved from being the responsibility of markets alone to receiving aid from central banks in a lender of last resort capacity. In the post-World War II period, especially since the 1970s, banking, currency, and debt crises became linked because governments became more willing to guarantee significant fractions of the liabilities of the banking system. The seminal paper by Diaz-Alejandro (1985) generated an enormous literature to explain the Latin American crises of the early 1980s. The Nordic crisis of 1991–92 and the Japanese Banking crisis of 1990 involved many of these elements. The Asian crisis of 1997–98 led to new theories which explained "triple crises" based on guarantees and foreign currency denominated debt. Finally, the recent Eurozone crisis has led to new work which emphasizes the feedback loop between bank guarantees and banks' holding of member states' sovereign debt which links financial to debt crises.

In this chapter, we examine the interconnections between financial and fiscal crises based on history, theory, and empirics. Section 2 presents a brief historical overview of financial crises. Banking crises can be traced back hundreds of years. Before the advent of deposit insurance and effective use of the lender of last resort, banking crises were banking panics. In the depression of the 1930s, governments instituted numerous interventions and guarantees effectively laying a strong precedent for subsequent fiscal resolutions. Since the breakdown of the Bretton Woods system in the 1970s and the advent of liberalized domestic and international financial markets, banking panics have increasingly evolved into fiscally resolved banking crises. Banking crises have often been global or regional events as countries have been linked together by fixed exchange rates, capital flows, and other sources of contagion. Debt crises—sovereign debt defaults—have also been around for centuries, associated with overborrowing and have been triggered by international and domestic shocks. Today they occur primarily in emerging countries, but again, several advanced countries in the Eurozone faced a tough test after 2008 (with a sovereign default in Greece). Currency crises—speculative attacks on pegged exchange rates—often accompanied banking crises and sometimes debt crises because of linkages between monetary policy and crisis resolution.

Section 3 surveys theoretical perspectives on financial crises. Banking crises traditionally were analyzed using three approaches: the monetarist approach, the financial fragility approach, and the business cycle approach. Modern perspectives build upon these earlier theories. The key approach is based on the Diamond and Dybvig (1983) notion of the inherent instability of banking because of a maturity mismatch. Also seminal are theories based on asymmetric information. In the recent decade, the financial frictions studied in partial equilibrium models have successfully been added to dynamic general equilibrium models. The pioneering modern work to explain why countries issue sovereign debt and try to avoid debt crises traces back to Eaton and Gersovitz (1981) who emphasize reputation. By contrast Bulow and Rogoff (1989a) focus on the deterrence effect of sanctions. Reinhart and Rogoff (2009) emphasize serial defaults, debt intolerance, and the distinction between domestic and foreign debt. New research in dynamic general equilibrium models also incorporates connections between the fiscal and financial side of the economy.

Section 4 provides empirical perspectives on financial crises. We discuss the methodological issue of crisis measurement which encompasses the definition, dating, and incidence of financial crises. Different approaches to definition and dating which are taken in the literature lead to very different patterns of recorded incidence and hence very different interpretations of the historical record. These classification problems must be acknowledged before any definitive general statements can be made. We also discuss the many and varied causes or determinants of financial crises, including bank credit-driven asset booms which have resonance for the recent crisis. A number of approaches have been taken to identify the key determinants of crises and to assess the predictive power of empirical models. This early warning indicators (EWIs) literature has made significant advances in the past two decades. However, our reading of the literature is that it remains very difficult to predict crises with a high level of accuracy both because of Goodhart's law as well as because of the complex economic ecosystem represented by the financial sector and the high dimensionality of the potential causes.[a] We then review measures of the output costs of financial crises and provide some measures of these losses using a comparable methodology across datasets. Again, different approaches in the literature and different classification systems lead to significantly different conclusions and hence different perspectives on the economic costs of crises.

Section 5 contains a preliminary examination of the empirical connection between financial and fiscal crises and identifies a potential new policy "trilemma." In the future, countries will be able to have two of the following three: a large financial sector, fiscal bailouts devoted to the inevitable crises that accompany leverage and financial deepening, and discretionary fiscal policy aimed at raising demand in the recessions occasioned by financial crises. This story is different from the older argument in the literature that fiscal policy is procyclical in less-developed countries. Moreover, as the recent crisis suggests, this trilemma may become more binding at higher initial levels of debt-to-GDP.

Section 6 concludes. Here we discuss the strengths and weaknesses of the literature and we consider some issues for further research.

2. HISTORICAL OVERVIEW

Financial crises can be traced back hundreds of years (Kindleberger, 1978). Historical narratives identify separate banking, currency, and debt crises and combinations of them (Bordo and Eichengreen, 1999; Bordo and Meissner, 2006; Reinhart and Rogoff, 2009). While financial crises cum fiscal crises are certainly not a new phenomenon, it would be incorrect to say that the recent global financial crisis and the subsequent Eurozone crisis were no different than all of those that have come before. The nature and origins of fiscal crises and their relationship to financial crises have in fact changed dramatically over the long run in important ways.

Banking crises before the advent of deposit insurance (and other components of the financial sector safety net) were banking panics—attempts by the public to convert their

[a] Goodhart's law proposed that "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes." So for instance, if policy makers were to regulate financial variables previously associated with crises, previous relationships will break down, but crises will still occur. See Arnold et al. (2012) on macroprudential policy and financial stability.

deposits into currency en masse. Unless resolved by lender of last resort actions, banking panics could seriously impact the real economy by reducing the money supply (Friedman and Schwartz, 1963) and by reducing financial intermediation (Bernanke, 1983). Banking panics would propagate through asset markets as banks under threat dumped assets in fire sales. They could also propagate via interbank connections and other institutional arrangements to create a systemic collapse (Mitchener and Richardson, 2014).

Banking panics could also be caused by shocks leading to the failure of important financial firms outside the traditional banking sector like shadow banks (Rockoff, 2014). They could occur as a consequence of a bank credit-driven asset price boom–bust cycle. Schularick and Taylor (2012), Brunnermeir and Oehmke (2013) and many others recently have argued that systemic banking crises are very likely to follow bank credit-driven asset price booms.

Finally banking crises can also have an international dimension as for example during the Baring crisis of 1890–91, the global instability of 1907, the Credit Anstalt crisis of 1931, the Asian financial crisis of 1997–98, and the subprime mortgage crisis of 2007–09. Bordo and Landon-Lane (2012) identify five global financial crises (1890–91, 1914, 1929–30, 1980–81, and 2007–08) where the incidence of banking crises affected banks in multiple countries and in several continents in the same year. In all of these cases of "contagion," cross-border claims and faltering foreign banks or counterparties led to insolvency or liquidity problems at home. In addition, interest rate shocks emanating from leading financial centers (eg, by the Bank of England in 1890, the Federal Reserve in 1929 and in 1980–81) could contribute directly or indirectly to starting or exacerbating financial stress especially in emerging countries (Kaminsky and Vega-García, 2016).

The incidence of banking panics was high in many advanced countries in the 19th century before monetary authorities learned to act as lenders of last resort. In the United Kingdom, the last depositor-led banking panic was in 1866. In France it was in 1882, and in Germany it was in 1873. In the United States, it took until 1933 and the advent of deposit insurance before banking panics ceased (Schwartz, 1987).

With the advent of deposit insurance and other forms of government guarantees during the Great Depression, and progressively in some countries even earlier, the nature of banking crises changed from panics to crises which were increasingly resolved by a fiscal rescue.[b] This created a direct link between the banking system and the government's

[b] Banking crises which were resolved by a fiscal bailout were quite common in emerging countries before World War I (Grossman, 2010). Even some prominent advanced countries like France in 1889 and Britain in 1890 violated Bagehot's (1873) stricture for a central bank to lend only to illiquid and not insolvent institutions and arranged a government-led, fiscally backed lifeboat operation rescue (White, 2015). Bordo and Flandreau (2003) show that in emerging countries, the bailouts of the late 19th century on several occasions led to a big run up in the debt-to-GDP ratio and serious fiscal crises (eg, Portugal, Greece, and Russia). However, there were no cases in advanced countries where banking crises led to fiscal crises before the 1930s (Schularick, 2012).

balance sheet. Once this precedent was set, a costly bailout now had the potential to create significant fiscal imbalance and even lead to a default. Moreover, guarantees could lead to moral hazard (ie, protected banks would increase their balance sheets and take on more risk knowing that they would be bailed out).[c] This would in turn increase the cost of bailouts ex post and increase the strain on the government's finances. In turn, if the deficits were money financed with an expansion in the monetary base, this would increase the likelihood of inflation, currency crisis, or a sovereign default.

Before the 1930s, sovereign defaults had long been a fact of life reflecting the precarious nature of borrowing (often in foreign currencies) to finance wars, cover revenue gaps, or build infrastructure none of which had immediate growth or revenue payoffs leaving a maturity mismatch. Sudden stops of capital flows often led to sovereign defaults for this reason (Bordo, 2006; Bordo et al., 2010). Banking crises, even in the absence of guarantees, could lead to fiscal distress by reducing real income and government revenues.

A wave of sovereign defaults tied to international capital flows occurred in the 1820s in many Latin American Republics as overoptimistic investors from Europe lent these fledgling republics more than their weak public finances could handle. It took four decades before these countries paid into arrears and could access international capital markets again. In the next two centuries, Latin America had three more waves of default (Marichal, 1989).[d] Most countries, with the principal exception of a few advanced countries, had sovereign debt defaults in the 19th and 20th centuries (Reinhart and Rogoff, 2009). Many of them were serial defaulters (Reinhart et al., 2003).

Currency crises—a speculative attack on a pegged exchange rate reflecting an inconsistency between domestic fundamentals and the peg—also were a frequent occurrence for emerging countries throughout the 19th and 20th centuries (Bordo and Schwartz, 1999). Advanced countries generally avoided them under the pre-1914 gold standard, but they became a bigger problem for them in the interwar and during the Bretton Woods system (Bordo et al., 2001).

Currency crises often occurred simultaneously with banking crises, referred to as twin crises (Kaminsky and Reinhart, 1999). Causality between them was often two way. A banking crisis could lead to capital flight by foreign depositors as occurred in 1931 in Germany (Eichengreen, 1992). Per contra, a currency crisis could lead to insolvency for banks with extensive foreign currency denominated liabilities and domestic currency denominated assets as occurred in a number of emerging countries in both the pre-1914

[c] According to Akerlof and Romer (1993) and White (2000) in the case of the US Savings and Loan crisis of the 1980s, guarantees led directly to regulatory forbearance which engendered moral hazard leading to a crisis.

[d] Kaminsky and Vega Garcia (2016) show that most of these defaults followed systemic financial crises in the core countries of Europe.

and post-1973 eras of financial globalization (Bordo and Meissner, 2006; Reinhart and Rogoff, 2009).

Currency crises became linked to debt crises for emerging countries who had borrowed abroad in foreign currencies in the 1890s (Bordo and Flandreau, 2003). With the advent of government guarantees on top of foreign currency denominated debt, currency, banking, and debt crises became interlinked in the emerging market crises of the late 1990s and early 2000s.

Thus, the recent Eurozone crisis was the culmination of a long history of different types of crises and their growing interconnections which evolved along with the deep seated forces of financial globalization and a belief in the necessity for government to socialize the income losses of financial crises.

3. FINANCIAL AND FISCAL CRISES: A LONG-RUN REVIEW OF THEORETICAL DEVELOPMENTS

In this section, we survey the theoretical literature on financial and fiscal crises. We first survey traditional approaches. Most of the literature treats the two types of crises, along with currency crises separately. We then examine more recent approaches that often combine banking and fiscal crises along with currency crises.

3.1 Banking Crises

The traditional view of a banking crisis was a banking panic or a liquidity crisis. It involved a scramble by the public for means of payment. Two frequent scenarios in which it occurred were: contagious banking panics when the public fearful that banks will not be able to convert their deposits into currency attempts en masse to do so; the second is a stock market crash that leads to fears that loans will become unavailable at any price. Without intervention by the monetary authorities or lender of last resort—through open market operations or liberal discount window lending—the real economy will be impacted by a decline in the money supply, by impairment of the payment system, and by the interruption of bank lending.

In the post-World War II period, with the widespread adoption of deposit insurance (both explicit and implicit), and with a generalized understanding of the role of the lender of last resort, old fashioned banking panics have become rare events. Instead, banking crises largely involve the insolvency of significant parts of the banking system. They have occurred when asset prices have plunged, whether prices of equities, real estate or commodities; when the exchange value of a national currency experiences substantial depreciation; when a large financial firm or nonfinancial firm faces bankruptcy; or a sovereign debtor defaults. Unlike banking panics which are brief episodes resolved by the central bank, a banking crisis is a prolonged disturbance that is resolved by means other than the lender of last resort, although at some stage it may supply liquidity through the discount window or open market operations.

Three traditional approaches to conceptualizing banking crises are: the *monetarist approach*, the *financial fragility approach*, and the *business cycles approach*. The contemporary literature based on rational expectations and game theory follows from these.

3.1.1 The Monetarist Approach

The monetarist approach of Friedman and Schwartz (1963) identifies financial crises with banking panics that either produce or aggravate the effects of monetary contractions. In a *Monetary History of the United States 1867–1960*, Friedman and Schwartz devote considerable attention to the role of banking panics in producing monetary instability in the United States. For Friedman and Schwartz, banking panics are important because of their effects on the money supply, and hence on economic activity.

According to them, banking panics occur because the public loses confidence in the ability of banks to convert deposits into currency. A loss of confidence is typically associated with the failure of some important financial institution (as happened in 1873, 1893, and 1907). Attempts by the public in a fractional reserve banking system to increase currency as a fraction of its money holdings, if not offset, can only be met by a multiple contraction of deposits. A banking panic, in turn, if not short-circuited by the monetary authorities, will lead to massive bank failures of otherwise sound banks.[e] They are forced into insolvency by a fall in the value of their assets in a vain attempt to satisfy a mass scramble for liquidity. Banking panics, such as occurred in 1930–33, have deleterious effects on economic activity primarily by reducing the money stock through a decline in both the deposit–currency and deposit–reserve ratios.

An extensive literature in economic history has been devoted to reexamining the banking panics of the 1930s. The debate swirled over the issue of whether the banking crises were really liquidity panics driven by "a contagion of fear" or whether they reflected bank insolvency as an endogenous response to the recession. Temin (1976) and most recently Calomiris and Mason (2003) provided evidence that cast doubt on the Friedman and Schwartz liquidity panic story. Richardson (2007) and Bordo and Landon-Lane (2010) provide evidence in its favor.

3.1.2 The Financial Fragility Approach

A tradition going back to the 19th century regards financial crises as an essential part of the upper turning point of the business cycle and as a necessary consequence of the "excesses" of the previous boom. Its 20th century proponents, Minsky (1977) and Kaufman (1986), basically extend the views Irving Fisher expressed in *Booms and Depressions* (Fisher, 1932) and in the "Debt deflation theory of Great Depressions" (Fisher, 1933).

[e] Carlson et al. (2011) and Richardson and Troost (2009) provide historical evidence on these issues.

According to Fisher, the business cycle is explained by two key factors: overindebtedness and deflation. Some exogenous event (displacement) provides new, profitable opportunities for investment in key sectors of the economy which increases output and prices initiating the upswing in the cycle. Rising prices, by raising profits, encourages more investment and also speculation for capital gain. The whole process is debt financed, primarily by bank loans, which in turn, by increasing deposits and the money supply, raise the price level. An overall sense of optimism raises velocity, fueling the expansion further. Moreover, the rising price level, by reducing the real value of outstanding debt encourages further borrowing. The process continues until a general and precarious state of "over-indebtedness" is reached. It exists when individuals, firms, and banks have insufficient cash flow to service their liabilities perhaps due to a shock to demand or supply. In such a situation, a crisis can be triggered by errors in judgment by debtors or creditors. Debtors, unable to pay debts when due or to refinance their positions, may be required to liquidate their assets.

Distress selling, if engaged in by a sufficiently large segment of the market, produces a decline in the price level because, as loans are extinguished and not renewed, bank deposits decline. Falling prices reduce net worth and profits, leading to bankruptcy. Both factors contribute to a decline in output and employment. In addition, while nominal interest rates fall with deflation, real rates increase, worsening the situation. The process continues until either widespread bankruptcy has eliminated the overindebtedness or at any stage reflationary monetary policy is adopted. However, once recovery begins, the whole process will repeat itself.

This approach has been revived since the financial crisis of 2007–09. Indeed some commentators have described the failure of Lehman Brothers in September 2007 as a "Minsky moment" (Brunnermeir and Oehmke, 2013).[f] It is also consistent with the credit boom approach of the BIS (Borio, 2012) and the long-run comparative empirical work on credit and asset price booms by Schularick and Taylor (2012) and Jordà et al. (2011).

3.1.3 The Business Cycle Approach

This approach views banking panics as more likely during a recession because the returns on bank assets are likely to fall as borrowers become less like likely to repay their loans (Mitchell, 1941). Depositors anticipating an increase in nonperforming loans will try to protect their wealth by withdrawing their deposits precipitating a bank run (Allen and Gale, 2007). Gorton (1988) following this approach finds that depositors anticipating a decline in income, and in an attempt to smooth their consumption, remove their funds from banks before the business cycle peak.

[f] See Wray (2015).

3.2 Recent Approaches to Banking Crises

3.2.1 Diamond and Dybvig: The Inherent Instability of Banking

In a seminal article, Diamond and Dybvig (1983) argue that banks transform illiquid claims by offering liabilities with a different smoother pattern of returns over time. Banks provide efficient risk sharing/insurance which the private market cannot provide. However, banks are vulnerable to runs because of the illiquidity of their assets. Thus there is a maturity mismatch. One equilibrium in this setup is a run which can be triggered, even on a sound bank, by a random event (a sunspot) because rational depositors, not wishing to be last in line, will rush to convert deposits into currency. Only the presence of deposit insurance or a lender of last resort can prevent banking instability.

An explosion of articles in the past two decades builds upon the Diamond and Dybvig model. A number of articles were critical of the sequential servicing constraint in the original Diamond and Dybvig model—that depositors had to wait their turn at the bank to access their cash. It was argued that as in the pre-1914 National Banking era, banks could suspend convertibility (Jacklin, 1987). On the other hand, Wallace (1988) justified the sequential constraint endogenously in his model. Other papers that rationalized the Diamond Dybvig (DD) sequential service constraint were Diamond and Rajan (2001) and Calomiris and Kahn (1991). Another issue was that of multiple equilibria leading to an inability to make strong predictions. In an influential article on currency crises, Morris and Shin (1998) used the global games approach to reach a unique equilibrium as a function of fundamentals without using a sunspot equilibrium as a coordinating device as in DD. Banking crises were analyzed in a similar way by Rochet and Vives (2004) and Goldstein and Pauzner (2005).

Subsequent literature extended the basic DD framework to encompass financial markets and the banking system (Allen and Gale, 1998, 2004); to include bubbles and crises (Allen and Gale, 2000); to include money and monetary policy in the basic DD type model (Diamond and Rajan, 2001, 2005, 2011, 2012); to include interbank markets (Bhattacharya and Gale, 1987). The DD model also is embedded in several articles justifying lender of last resort intervention to provide liquidity in a financial crisis (Holmström and Tirole, 1998; Gorton and Huang, 2004; Rochet and Vives, 2004).

3.2.2 Information Asymmetry

The explanation of banking panics that the asymmetric information approach offers is that depositors cannot costlessly value individual bank assets, and hence they have difficulty in monitoring the performance of banks (Jacklin and Bhattacharya, 1988; Chari and Jagannathan, 1989). On this view, a panic is a form of monitoring. Faced with new information, which raises the perceived riskiness of bank assets, depositors force out both sound and unsound banks by a system-wide panic.

3.3 Fiscal Crises

The canonical fiscal crisis is a debt crisis. It is a situation where a sovereign debtor is unable to service the interest and or principle as scheduled. A debt crisis arises when the fiscal authorities are unable to raise sufficient tax revenue in the present and the future to service and amortize the debt.

A debt crisis can then become a financial crisis when it impinges on the banking system and a currency crisis when it threatens the reserves of the central banks as was the case in the Asian crisis of the 1990s. Banking crises can feed into debt crises when the fiscal authorities bailout insolvent banks which then increases sovereign debt to a point where it becomes unsustainable. Debt crises can also spill into banking crises when banks hold significant amounts of sovereign debt whether by choice or because of government attempts to force banks to hold significant levels of government debt.

Later we survey the literature on sovereign debt crises and their linkages to financial (banking) crises.

3.3.1 Sovereign Debt Crises: Theory

Two seminal articles have driven much of the modern literature on sovereign debt crises.[g] Eaton and Gersovitz (1981) explained the existence of sovereign debt markets and the incentive of sovereign borrowers to repay their debt by access to credit markets. Debtors worried that a default could ruin their reputation and cutoff future access to the foreign capital needed to finance economic development and to smooth consumption over time. Bulow and Rogoff (1989a,b) argued that other methods of self-insurance can substitute for foreign borrowing and that the main reasons countries avoid default is because of the threat of sanctions. In the 19th century, the British (and other European lenders) would send in the gunboats or use other means to seize the defaulting country's customs revenues or other assets. Today, trade sanctions, withholding of trade credit and other legal interference could matter. Another early development was the analysis of excusable default. Grossman and van Huyck (1988) argue that countries that defaulted because of a large shock to their economy not of their own making were treated better by the credit markets than countries which defaulted because of bad economic policy decisions.

The subsequent literature was doubtful of sanctions in the post-World War II era (Cole and Kehoe, 1995; Eaton, 1996; Kletzer and Wright, 2000) although there is considerable historical evidence for this (Mitchener and Wiedenmeir, 2010) for the pre-World War I era. Emphasis was placed by some on the collateral damage to the economy from default (Cole and Kehoe, 1998).[h] Bulow and Rogoff (2015) defend the sanctions approach as a way to understanding recent events in Greece and Argentina.

[g] See Panizza et al. (2009) for a recent survey.

[h] Two recent models of sovereign defaults which occur following adverse shocks to the economy are Aguiar and Gopinath (2006) and Arellano (2008).

An additional development was the focus on serial default. Reinhart et al. (2003) showed that a number of defaulting emerging countries had a long historical record of debt default. This pattern of persistence extended to a number of European countries (eg, Spain and France) which had an earlier history of serial defaulting. Moreover, they found that countries which were serial defaulters also had *debt intolerance* (ie, that they would tend to default at significantly lower debt-to-GDP ratios than advanced countries). For example, Argentina defaulted in 2002 at a debt-to-GDP ratio of 35% whereas Japan today has a debt-to-GDP ratio well above 200% and it is not even close to defaulting.

Reinhart and Rogoff (2009) make an important distinction between domestic debt and foreign debt. They argue that domestic debt default by inflation, financial repression, redenomination, abrogation of gold clauses, etc., can have consequences as serious as external default. In addition, they argue that defaulting on high domestic debt may be a strong rationale for the use of the inflation tax in many countries.

3.4 Fiscal Crises and Financial Crises

After the breakdown of the Bretton Woods system and the liberalization of global financial markets, as well as domestic financial systems across the world, the stage was set for waves of systemic financial and fiscal crises. A key integrating element between financial and fiscal crises was the widespread use of guarantees by the government of the liabilities of the banking system.[i] The seminal article which lays out clearly the dynamics of fiscal–financial crises interaction was by Diaz-Alejandro (1985).[j] He describes the unfolding disaster that occurred in Chile from 1977 to 1982 after it liberalized its domestic financial system and opened up its capital account. Chile, like the other Latin American countries, had extensive controls over the domestic financial system as well as capital controls since the 1930s.

The Pinochet regime, under the influence of the "Chicago boys"—students of Al Harberger—liberalized every aspect of the economy. They reduced tariffs, eliminated controls over the domestic financial system, and removed capital controls. They also in 1977 reduced barriers to entry into banking, explicitly did not introduce deposit insurance, and forswore a bailout of the banking system in the event of trouble. They also pegged the Chilean peso to the US dollar.

The new liberalized regime encouraged massive capital inflows which led to increases in bank credit and fueled an asset price boom. A major bank failure in 1977 led to a bailout for fear of contagion. Afterwards, the government again forswore against future bailouts. The bailout which soon followed encouraged moral hazard and the credit boom

[i] See Schularick (2012) and Alessandri and Haldane (2009).

[j] See Reinhart (2015).

continued. In early 1982, more banks failed and their liabilities were guaranteed. This meant that the government had taken on a new contingent liability which in turn led to a growing fiscal deficit. The central bank financed the deficit with the inflation tax. This led to inflation and set the stage for a speculative attack on its reserves. A major banking and currency crisis ensued in the summer of 1982 leading Chile to abandon its peg and nationalize its banking system. It was followed by a debt crisis in 1983.[k]

McKinnon and Pill (1986) model the effects of liberalization and reform on a previously financially repressed emerging country. In their model, like in Diaz-Alejandro (1985), there is a large unsustainable lending boom financed by foreign capital, intermediated by the banks. The banks believe that their foreign loans are guaranteed by the government. This overborrowing phenomenon leads to rising domestic credit, an increase in money growth, inflation, and an asset price boom. A foreign shock leads to a collapse in the boom, a banking crisis, a currency crisis, and a reversal of the reforms.

3.4.1 The Japanese and Nordic Banking Crises 1990–92

The background to the Japanese banking crisis in 1890 was a boom–bust cycle, which began in the mid-1980s with a run up of real estate prices fueled by an increase in bank lending and loose monetary policy. The Bank of Japan began following a looser monetary policy in the aftermath of the Plaza Accord of 1985 which led to an appreciated yen and a weaker dollar (Funabashi, 1988). The resulting property price boom in turn led to a stock market boom as the increased value of property owned by the firms raised expected future profits and hence stock prices (Iwaisako and Ito, 1995). Both rising land prices and stock prices in turn increased firms' collateral encouraging further bank loans adding more fuel for the boom. The bust may have been triggered by the Bank of Japan's pursuit of a tight monetary policy in 1989 to stem the asset price boom.

The subsequent asset price collapse in the next 5 years led to a collapse in bank lending with a decline in the collateral backing corporate loans. The collapse in asset prices further impinged on the banking system's capital making many banks insolvent.[l] Lender of last resort policy prevented a classic banking panic, but regulatory forbearance propped up insolvent (zombie) banks. The bailout costs of the bank rescue and the slow economic growth that ensued swelled the already high Japanese debt-to-GDP ratio since then, but Japan has never defaulted on its debt. A fiscal crisis was avoided because Japanese sovereign debt is denominated in yen and is mainly domestically owned.

The Nordic financial crisis of 1991–92 involved a banking crisis, a currency crisis, and a large fiscal bailout. In the case of Norway, quantitative restrictions on bank lending were lifted in 1984. This led to a bank credit financed real estate boom and a serious

[k] Velasco (1987) provided a model of this experience.

[l] Many aspects of the Japanese experience resonate with the financial accelerator approach of Bernanke et al. (1999).

banking crisis (Steigum, 2009). The Swedish financial crisis of 1992 involved both the banking sector and the exchange rate. Liberalization of the financial sector and the capital account in the 1980s after decades of financial repression led to a bank credit fueled asset price boom (stocks and real estate). The deflationary shock of the ERM crisis triggered an asset price bust and a collapse of the banking sector as well as a massive currency crisis and devaluation. A fiscal bailout led to a run up of the debt-to-GDP ratio but not sufficient to trigger a fiscal crisis (Jonung et al., 2009).

A similar severe crisis occurred in Finland at the same time with the collapse of the Soviet Union a key real fundamental (Honkapohja, 2009). The loan losses in all three countries were considered large (Norway 6% of GDP; Sweden 7% of GDP; Finland 7% of GDP) but the fiscal resolutions in all three cases did not threaten a fiscal crisis (Drees and Pazarbasioglu, 1994). Thus the Nordic crisis may be the forerunner of the guarantee-induced fiscal crisis/financial crisis nexus earlier identified for emerging countries.

3.4.2 The Asian Crisis

The Asian crisis of 1997–98 involved banking, currency, and debt crises and these crises were all connected by government guarantees and an ostensibly new factor "original sin" or foreign currency liabilities.[m] A key mechanism by which foreign borrowing led to banking crises was that the Asian tigers (Thailand, Indonesia, Malaysia, and Korea) borrowed abroad extensively in foreign currency. They did this because they had not yet financially developed enough to issue debt in their own currencies as could the advanced countries. Borrowing abroad (eg, in dollars), gave access to foreign capital at low international interest rates. The risk associated with original sin is that if the country has a currency crisis and ends up devaluing its currency then it will have to generate greater tax revenues in domestic currency and export earnings to service its foreign debt. This in turn would depress the real economy and increase the likelihood of a sovereign default. The likelihood that exports could rise sufficiently depended on strong global demand and high elasticities. Moreover, the banking systems in these countries funded their loans with foreign securities (often short term) and after the devaluation, their balance sheets would become impaired increasing the likelihood of insolvency and a banking crisis.

The Asian crisis led to the creation of "third-generation" speculative attack models. They were an extension of both first- and second-generation speculative attack models. The first-generation model of currency crises (Krugman, 1979) posited that a speculative attack would inevitably occur when domestic fiscal and monetary fundamentals were inconsistent with adherence to a pegged exchange rate. The second-generation models (Obstfeld, 1995) posited that speculative attacks would occur when agents, who

[m] See Eichengreen and Hausmann (2005).

understood the weights that the government placed on the stability of the domestic economy and adhering to a peg, anticipated that the government would prefer domestic stability in the event of a crisis. Speculators would thereby sell the currency short and generate a crisis.

Several authors extended the first- and second-generation models to incorporate special features of the Asian crisis including moral hazard (guarantees), short-term borrowing in foreign currencies, and currency depreciation. Krugman (1998) argued that the currency and financial crises in Asia reflected the role of moral hazard as the progenitor of financial instability which in turn was a key cause of currency crises. According to his story, financial institutions in these countries engaged in risky lending on the assumption that they would be bailed out while at the same time they financed themselves with offshore loans at close to international interest rates. The capital inflow and domestic bank lending fueled an asset market boom which in turn encouraged the banks to lend more. This process encouraged a domestic investment and consumption boom and a growing current account deficit. When external factors revealed the exchange rate to be overvalued, a classic speculative attack led to devaluation. The devaluation in turn sparked a financial crisis as the banks' short term, foreign currency denominated loans mushroomed, making them both illiquid and insolvent. Bailouts of the financial system and especially of their dollar obligations in turn precipitated further speculative attacks and exhausted the monetary authorities' international reserves.

Dooley (2000) viewed the liabilities of the monetary authorities backing the financial safety net as an alternative claimant on emerging countries' international reserves. Market agents understood this and staged a speculative attack at the moment that net liabilities exceeded international reserves.

Krugman (1999) focused on the balance sheets of firms which borrowed abroad in foreign currencies. A speculative attack would occur when the market anticipates that a depreciating currency will lead to insolvency and contracting economic activity hence pulling out funds and precipitating the adverse chain of events.

Burnside et al. (2004) also emphasize the key role of government guarantees in explaining the Asian crisis. In their model, banks borrow in foreign currencies unhedged because their foreign debt is guaranteed by the government. But when a devaluation occurs, following an external shock, the banks default on their foreign debt and declare bankruptcy, but the government does not have the resources to pay for a bailout. This leads to both a banking crisis and a currency crisis when the central bank uses seigniorage to fund the fiscal deficit.

Corsetti et al. (1999) also model the Asian crisis. In their model, the government guarantees the banks' foreign currency loans which are used to finance domestic investment. This leads to a capital inflow boom, a current account deficit, and an investment boom. Private sector borrowers believe that they and the banks will be bailed out. When a shock occurs, this leads to both a banking crisis and a possible debt crisis as the contingent liabilities that the government has to cover increase the fiscal deficit.[n] Thus the Asian crisis had many elements of the Diaz-Alejandro story with guarantees that induced fiscal deficits and which were financed largely by money issue rather than increased sovereign borrowing.

3.4.3 The Eurozone Crisis

The Eurozone crisis of 2010–14 was a sequel to the global financial crisis of 2007–09 involving strong connections between banking and fiscal crises. Reinhart and Rogoff (2009, 2011) suggest that the link between banking and fiscal crises has strong historical roots. They show that banking crises often precede debt crises and that for a large panel of advanced and emerging countries in the 20th century that the debt-to-GDP ratio rises by 86% in the 3 years following a banking crisis setting the stage for a downgrading of credit and a possible default. Schularick (2012) notes that this has mainly occurred in the post-World War II period.

The Eurozone crisis seems to fit the prediction that fiscal and financial crises have a strong connection. In the aftermath of the subprime mortgage crisis, several European countries that had been connected to the US crisis or which had bank credit-driven house price booms, engaged in expensive bond financed bank bailouts. These bailouts and economic collapse increased the fiscal deficit leading to debt surges. The bailouts across Europe followed in some respects the example of Ireland which in September 2008 guaranteed its entire financial system. To fight the recession that accompanied the crisis, they also engaged in expansionary automatic fiscal policy which also increased the deficits.

Reinhart and Rogoff (2009) argue that the decline in tax revenues produced by the fall in output plus the expansionary government expenditures explained more of the run up in deficits and debt than the bailouts themselves. Laeven and Valencia (2012) provide a crude measure that separates out the rise in debt due to bailouts and resolution activity and a remaining portion due to discretionary and automatic fiscal expansion. In their sample, the median rise in the debt-to-GDP ratio after a crisis is 12 (percentage points) with the majority (6.8) attributable to fiscal rescue packages. For advanced economies, the figures are 21.4 and 3.8, and in emerging economies they are 9 and 10. Significant heterogeneity across countries is evident. Inference should recognize this fact and the historical record should be assessed in light of these data.

Against this background of weakening fiscal positions across the Eurozone, the announcement in 2009 that the Greek government had falsified its fiscal books set the stage for the Eurozone debt crisis which first involved the threat of a Greek default

[n] Other papers that model the Asian crisis and place emphasis on government guarantees include: Arellano and Kocherlakota (2014), Burnside et al. (2001), Burnside (2004), and Schneider and Tornell (2004).

and then contagion to other members via their banks which had significant holdings of Greek and other peripheral countries' sovereign debt.

The threatened sovereign default by Greece fed into a banking crisis because banks in Greece and the other financially integrated Eurozone countries held large amounts of Greek and other peripheral Eurozone sovereign debt. In the case of Ireland, a blanket guarantee of the Irish financial sector by the Irish government followed the collapse of a property price boom. This collapse made the Irish banks insolvent, and led to a fiscal crisis because markets expected that the Irish government would not be able to service the large run up in its debt that followed. An 85 billion euro international rescue by the IMF, the EU, and others followed in 2010. Later, some private sector actors were bailed-in.

In Spain, where another housing boom turned to bust, the crisis also led to fiscal problems. Spain introduced several costly bailout packages with enhanced guarantees, and took on a European bailout package. Throughout, international pressure—both political and market based—was harsh leading to higher risk premia. From 2010, Spain adopted a series of "austerity" plans coincident with these bailouts. In addition, Spanish banks increased demand for Spanish sovereign debt in order to take advantage of liquidity funding from the European Central Bank threatening an outcome whereby fiscal problems could be transmitted to the banks. Bond spreads in Portugal and Italy spiked after 2010, but countries such as France and Belgium also faced significant bond market pressure. European countries displayed vulnerabilities in the run up to the crisis, but the collapse of confidence in international bond markets for many European countries reflected the constraints of nations in a monetary union with no strong fiscal union, a weak/nonexistent banking union and (at least initially) hesitant monetary policy from the ECB.

The recent crisis presents several fine examples of the interconnection between fiscal and banking crises and new theoretical models and empirical evidence have been supportive of these links. Bolton and Jeanne (2011) model the interconnection between sovereign risk and the banking system in a currency union where the banks in each country diversify their portfolios by holding the sovereign debt of other member states. Holding government bonds serves as safe collateral which allows them to increase their leverage. The default by one member spreads to the others via the weakening of bank portfolios.[o]

Gennaioli et al. (2014) also model the interconnection between sovereign default and the banking system. As in Bolton and Jeanne, banks hold sovereign debt as collateral which allows them to increase their lending. A debt crisis leads to a credit crunch and a decline in real income. The authors demonstrate that the costs of a fiscal shock are higher for more financially developed countries.[p]

[o] Battistini et al. (2014) observe that in the Eurozone banks increase their holdings of domestic debt even when yields (and risk) rise and when systemic risk rises. Various policy implications for monetary unions like the EMU are discussed.

[p] Also see Uhlig (2013).

Acharya et al. (2014) model a two-way interconnection between fiscal crises and banking crises. Bank bailouts lead to an increase in sovereign risks because of the increase in fiscal deficits and debt ratios. This in turn weakens the banking system which holds sovereign debt as collateral.

They use the Irish bailout of 2008 as their example. Their model predicts that the spreads between bank CDSs and sovereign CDSs should rise during the banking crisis. Then after the bailout, bank CDSs should decline and sovereign CDSs should rise. This reflects the transfer of risk from the banks to the government. Empirical evidence for the advanced countries in the Eurozone backs this up. After the subprime crisis began in 2007, bank CDSs rise dramatically with no change in sovereign CDSs. Then after the Lehman collapse and the Irish guarantee at the end of September 2008, sovereign CDSs rise and bank CDSs decline.

Mody and Sandri (2012) examine the behavior of sovereign risk spreads of the Eurozone countries before and after the crisis of 2007–09. They show that after the creation of the Euro in 1999 sovereign spreads converged across the Eurozone. Then after the Bear Stearns bailout in March 2008 spreads increased in countries which had vulnerable financial sectors likely to be bailed out. After the Lehman failure in September 2008, spreads increased dramatically in countries that had higher debt ratios. Then, after the failure of Anglo Irish bank in January 2009, spreads increased across the Eurozone reflecting the increased vulnerability of the financial systems of all the member countries.

Martin and Philippon (2015) compare the behavior of member states of the Eurozone to that of the states in the United States during the Great Recession. The key difference between the Eurozone and the United States was the absence of a well-functioning fiscal union in the former (Bordo et al., 2013). What their analysis shows is that the United States and European cross-sectional experience in household debt and employment were quite similar in the period 2007–10. However, after 2010 there was a marked difference between the two currency areas. The peripheral Eurozone countries experienced a sudden stop in capital flows reflected in a spike in borrowing costs (spreads) and a drop in employment and growth. By contrast, the pattern of these variables across US states did not diverge. Past fiscal policy in the Eurozone countries, because of its effect on accumulated debt, impacted their economies both through the perceived risks to repayment and sustainability and the constraints on expansionary fiscal policy it generated after 2010.

Thus the Eurozone crisis represents the culmination of a guarantees-induced connection between financial crises and fiscal crises. The special characteristics of the Eurozone (the absence of fiscal and banking unions, the absence of floating exchange rates, and the ability to offset shocks with domestic monetary policy) made things worse.

4. EMPIRICS OF FINANCIAL CRISES OVER THE LONG RUN

In the following sections, we discuss the empirics of financial crises. We take a close look at defining and dating financial crises also exploring the coincidence of several types of crises. We highlight that a key concern for researchers should be classification uncertainty. Simply put, leading authors disagree on the definition of a crisis leading to discrepancies between authors and ultimately different conclusions about the impact and causes of crises. We discuss various methods and findings on the empirical determinants of financial crises in the next subsection. After this, we discuss how to measure output losses associated with financial crises and provide an overview of the literature. A closely related topic is recoveries. We explore the linkages between government debt and the fiscal costs of bailouts and guarantees. We note a new trade-off: when bailouts are costly, discretionary fiscal policy may be constrained especially in the face of a large financial crisis.

4.1 Dating of Financial Crises: A History of Comprehensive Chronologies

A number of different chronologies of financial crises exist. The crisis dates enumerated by each source are quite different as we will show. The coverage also varies in terms of the years and number of countries included in each sample. Because of all these discrepancies, the conclusions from each study are likely to differ and sometimes dramatically so. In this section, we survey the methodologies of the leading databases for dating financial crises.

Economists for the last 200 years have been drawn to major financial events and used them to learn about the macroeconomy. Conant (1915) surveys the history of central banking in many different nations in the early 20th century along the way detailing the prospective causes and impacts of financial events. The National Monetary Commission of the United States held lengthy hearings from leading financial experts, and significant amounts of evidence on the financial histories of many countries were submitted as evidence. Grossman (1994) was one of the first papers to systematically collect data on banking crises in the Great Depression.

Edwards and Santaella (1993) provide a chronology of currency devaluations from the Bretton Woods period. By the 1990s, researchers at the World Bank like Caprio and Klingebiel (1996) were providing dates for systemic banking crises in a large sample of countries. These crises were an economic phenomenon that had mainly disappeared between the 1940s and the early 1970s. By the late 1970s and into the early 1990s, such crises became increasingly commonplace first in Less Developed Countries and Emerging Market Economies (EMEs) and then in advanced countries. These events attracted significant interest by policy makers and academic researchers alike.

Kaminsky and Reinhart (1999) provide an account of banking, currency, and "twin" crises for nonadvanced countries. Laeven and Valencia (2008, 2012) compile a comprehensive dataset of banking, currency, and debt crises for the period 1970–2011. Laeven and Valencia's dataset covers the experience of 162 advanced, emerging, and less-developed economies.

For the long run, three major, comprehensive contributions stand out. Bordo et al. (2001) date banking, currency, and twin crises for all years between 1880 and 1997. For the years 1880–1945 their sample includes 21 now mostly advanced countries (with the exceptions of Argentina, Brazil, and Chile) and from 1945 data from 56 countries is available. Reinhart and Rogoff (2009) and Reinhart (2010) provide accessible data on banking, currency, and debt crises for 70 countries. Their record on sovereign debt crises extends back to the medieval period but only for a selected number of European polities. From 1800 Reinhart and Rogoff track banking, currency, and debt crises. Carmen Reinhart's website provides a set of open-access excel spreadsheets.[q] Finally, Taylor (2015), based on research with Jordà et al. (2011) provides the dates for "systemic" financial crises (mainly banking crises) for 17 countries 1870–2010.

Recently, Romer and Romer (2015) have collected a new set of dates for *financial distress* based on readings of the OECD Economic Outlook 1967–2007. While previous studies have mainly provided binary indicators of the various financial crises, Romer and Romer generate a measure based on a scale of 0–15. This measure is substantially different from traditional measures of crises, so we do not use it further in our analysis.

4.2 Crises Definitions

Table 1 gives the stated definitions for dating the various types of crises in each of the leading datasets: Bordo et al. (BEKM), Laeven and Valencia (LV), Reinhart and Rogoff (RR), and Jordà et al. (JST). As is evident, for banking and currency crises, the definitions vary by sets of authors leading to significant disagreements both about timing and whether there was or was not a crisis. In particular, for banking crises, authors disagree about how many banks must be closed or what percentage of the financial system's capital must be impaired for a crisis to be classified as systemic. Laeven and Valencia require that major policy interventions take place. Reinhart and Rogoff classify more crises than other authors likely because they only require bank runs to lead to the "closing of *one* or more financial institutions" (our emphasis).

Currency crises are generally defined as sharp declines in the nominal exchange rate. Many authors use a threshold decline (eg, 15% or 30%) in the nominal exchange rate possibly conditional on having only limited flexibility in the preceding years. BEKM use an exchange market pressure (EMP) index as developed in Eichengreen et al. (1995) where possible. Prior to the 1970s, and especially prior to the 1930s, the required data are relatively hard to obtain and so the emphasis is generally on nominal exchange rate movements. Laeven and Valencia follow Frankel and Rose (1996) as do Reinhart and Rogoff. There are some differences in the cutoffs used by the latter two sets of authors.

[q] http://www.carmenreinhart.com/.

Table 1 Crisis Definitions Four Leading Datasets

Authors	Sample	Banking Crisis Definition	Currency Crisis Definition	Debt Crisis Definition
Bordo et al. (2001)	1880–1939 21 Advanced countries 1945–97 21 Advanced countries + 35 less developed countries and emerging market economies.	Financial distress resulting in the erosion of most or all of aggregate banking system capital as in Caprio and Klingebiel (1996).	Forced change in parity, abandonment of a pegged exchange rate, or an international rescue. Or: an exchange market pressure (EMP) above a critical threshold (calculated as a weighted average of exchange rate change, short-term interest rate change, and reserve change relative to the same for the center country, the United Kingdom before 1913 and the United States after). A crisis is said to occur when this index exceeds a critical threshold. BEKM score an episode as a currency crisis when it shows up according to either or both of these indicators.	No debt crises are dated in this dataset.
Reinhart and Rogoff (2009)	1800–2011 70 Countries	A banking crisis occurs when there are one of two types of events: (1) bank runs that lead to the closure, merging, or takeover by the public sector of one or more financial institutions or (2) if there are no runs, the closure, merging, takeover, or large-scale government assistance of an important financial institution (or group of institutions), that marks the start of a string of similar outcomes for other financial institutions.	Reinhart (2010) refers to a working paper version of Reinhart and Rogoff (2011) stating they follow Frankel and Rose (1996). Frankel and Rose date a currency crisis as a period with a nominal depreciation of more than 25% which represents a greater than 10% increase in the rate of depreciation. Reinhart's website provides the following definition: "An annual depreciation versus the US Dollar… of 15 percent or more."	"External debt crises involve outright default on payment of debt obligations incurred under foreign legal jurisdiction, repudiation, or the restructuring of debt into terms less favorable to the lender than in the original" (Reinhart and Rogoff, 2011).

Continued

Table 1 Crisis Definitions Four Leading Datasets—cont'd

Authors	Sample	Banking Crisis Definition	Currency Crisis Definition	Debt Crisis Definition
Laeven and Valencia (2012)	1970–2011 162 Countries	Two conditions 1. "Significant signs of financial distress in the banking system (as indicated by significant bank runs, losses in the banking system, and/or bank liquidations)." 2. Significant banking policy intervention measures in response to significant losses in the banking system.	Nominal depreciation of the currency against the dollar of at least 30% that is also 10 percentage points higher than the rate of depreciation in the year before	"Default and restructuring" Data from Calomiris and Beim (2001), World Bank (2002), Sturzenegger and Zettelmeyer (2006), IMF staff reports and reports from rating agencies.
Taylor (2015)/ Jordà et al. (2011)	1870–2011 17 Countries	Taylor (2015) and Jordà et al. (2011) describe their coding as following Bordo et al., Reinhart and Rogoff, Laeven and Valencia and Cecchetti et al. (2009).	Not dated.	Not dated.

Comprehensive data on sovereign debt crises from the 19th century up to the 21st century come from Reinhart et al. (2003) and are also presented in the spreadsheets on Reinhart's website. Laeven and Valencia also provide their own dates based on a multitude of sources. The latter do not cite Reinhart and Rogoff as a source for their crisis dates. Laeven and Valencia date moments of sovereign default and restructuring. Reinhart and Rogoff date external debt crises when there is "outright default on payment of debt obligations incurred under foreign legal jurisdiction, repudiation, or the restructuring of debt into terms less favorable to the lender than in the original" (Reinhart and Rogoff, 2011).

As is visible, substantial disagreement across teams of authors exists. We revisit this later after exploring the record on the frequency of financial crises.

4.3 Financial Crises: The Historical Record

Fig. 1A–D shows the sample percentage of country-year observations for the first year of four different kinds of financial crisis. This variable is calculated as the ratio of the number of years in which the set of countries in the sample is in the first year of a banking, currency, debt, twin (banking and currency), or triple (banking, currency, debt) crisis to the total number of country-years.[r] We compare outcomes for various chronologies and across four time periods: The classical gold standard (1880–1913), the interwar period (1919–39), Bretton Woods (1945–72), and the recent period of globalization (1973–present). We note, as Bordo et al. (2001) do, the sample of countries does change over time within the BEKM dataset going from 21 to 56 countries in the post-1972 period which changes sample frequencies somewhat.

Currency crises are the most frequent variety of crisis followed by banking crises, debt crises, twin crises, and finally triple crises. By and large, all of the different chronologies agree on the trends. For the three datasets that cover the interwar period, only two out of three agree (Bordo et al. and Jordà et al.) that this period saw the highest frequency. Reinhart and Rogoff's data suggest that the recent period has a higher incidence of banking, triple, and debt crises (not pictured) than in the interwar period. Reinhart and Rogoff also show roughly the same frequency of twin crises in the interwar and the post-1973 period and a higher likelihood of a currency crisis in the Bretton Woods period and the post-1973 period. As in Bordo et al. (2001), there is little evidence that crises became more frequent over the long run with the possible exception of currency crises.

[r] Twin crises happen when a currency crisis event takes place within 1 year before or after a banking crisis. Triple crises are twin crises with an associated sovereign default within a 1-year window of either a currency or banking crisis. We avoid double counting by assigning a zero to all banking and currency crises that occur in the context of twin or triple crisis. Similarly any twin crisis that occurs with a sovereign default within a year is only counted as a triple crisis.

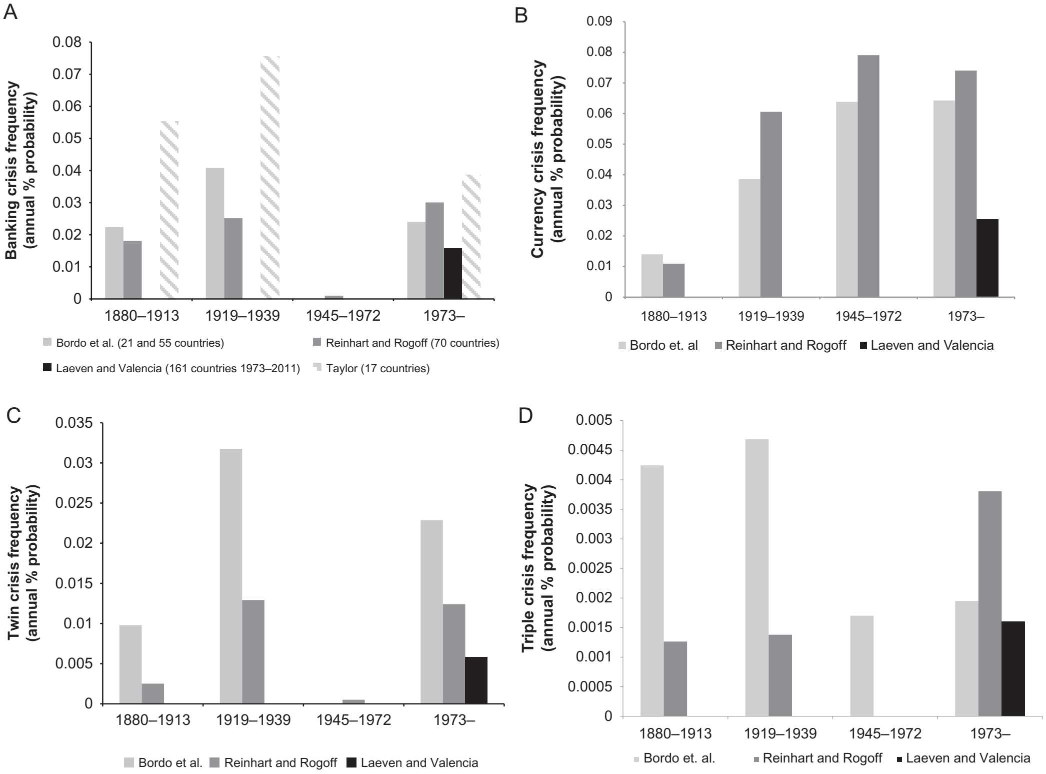


Fig. 1 (A) Banking crisis frequencies, 1880–2012. (B) Currency crisis frequencies, 1880–2009. (C) Twin crisis frequencies three datasets, 1880–2012. (D) Triple crisis frequencies, four datasets, 1880–2012. *Notes*: (A–D) Bars in (A) show the ratio of the number of country-years when a country was in the first year of a banking crisis to the total number of country-years in the sample. A banking crisis is defined differently according to each dataset. Banking crises are events not preceded or followed within 1 year by a currency crisis or a currency and debt crisis. Taylor (2015) studies "systemic crises." Laeven and Valencia have no data prior to 1970 so these data are excluded from the first three subsamples. Bars in (B) show the ratio of the number of country-years when a country was in the first year of a currency crisis to the total number of country-years in the sample. A currency crisis is defined differently according to each dataset. Currency crises are events not preceded or followed by banking crises or banking and debt crises. Laeven and Valencia have no data prior to 1970 so these data are excluded from the first three subsamples. Bars in (C) show the ratio of the number of country-years when a country was in the first year of a twin crisis to the total number of country-years in the sample. Currency and banking crises are defined differently according to each dataset. Twin crises are banking crises preceded or followed within 1 year by a currency crisis. Triple crises involving a debt default, banking, and currency crisis are excluded. Laeven and Valencia have no data prior to 1970 so these data are excluded from the first three subsamples. Bars in (D) show the ratio of the number of country-years when a country was in the first year of a triple crisis to the total number of country-years in the sample. Currency and banking crises are defined differently according to each dataset. Triple crises are banking crises preceded or followed within 1 year by a currency crisis and a debt crisis. Laeven and Valencia have no data prior to 1970 so these data are excluded from the first three subsamples.

Fig. 1B shows that currency crises shot up in probability in the interwar period and from then on have intensified slightly with Bordo et al. and Reinhart and Rogoff reporting probabilities in the range of 0.06–0.08. These two datasets are in strong disagreement with the Laeven and Valencia dataset in the recent period (1973–present). Even in samples where the years and countries overlap exactly, Laeven and Valencia report only half the currency crises that are recorded in Reinhart and Rogoff or Bordo et al.

In terms of time trends in twin crises, Bordo et al. find that their frequency was the highest in the interwar period (0.03) and the lowest in the Bretton Woods period. Reinhart and Rogoff's data disagree showing that a country would be equally likely to suffer a twin crisis in the interwar period as in the recent period (1973–2012). Laeven and Valencia date far fewer twin crises due to the comparatively low number of currency crises recorded.

Finally, for triple crises, both Bordo et al. and Reinhart and Rogoff agree that these are rare events and they occur in less than 1% of the country-years within sample. The datasets disagree with Reinhart and Rogoff showing that they are now more frequent than in the previous three periods while Bordo et al. show the pre-World War I period and the interwar as those with the highest likelihood of a triple crisis.[s] Once again, Laeven and Valencia do not concur with Reinhart and Rogoff for the 1973–2012 period suggesting that triple crises are much more rare than in the other two datasets.

Fig. 2A–E shows the number of crises that occur alone or coincident with other types of crises. With these diagrams, the connection between banking crises and outright sovereign default can be explored. The fraction of debt crises associated with banking crises (or both a banking crisis and a currency crisis) was nearly 0.21 in the years 1880–1913. In this constant country sample, this figure falls by over one half to 0.10 for the period 1919–39. Since 1973, the figure is 0.30 when we use crisis dates from Laeven and Valencia. Using Reinhart and Rogoff's data, the number stands at 0.29 for the 1973–2012 period.[t]

Of course research along the lines of Kaminsky and Reinhart (1999) provides evidence that currency crises frequently accompany banking crises in LDCs and Reinhart and Rogoff (2009) suggest that many debt crises are preceded by banking crises. It is interesting to note that according to our strict definitions "many" here equals only about 17% if a 1-year window is given. Bordo and Meissner (2006) discuss the historical relationship between banking, currency, and debt crises. They find that a significant

[s] Note we use Reinhart and Rogoff's debt crisis dates when dating a triple crisis within the Bordo et al. dataset.

[t] Some readers will note a difference between our numbers in Fig. 2D and those in the comparably designed fig. 4 of Laeven and Valencia (2012). There are some discrepancies within the Laeven and Valencia dataset which we have corrected.

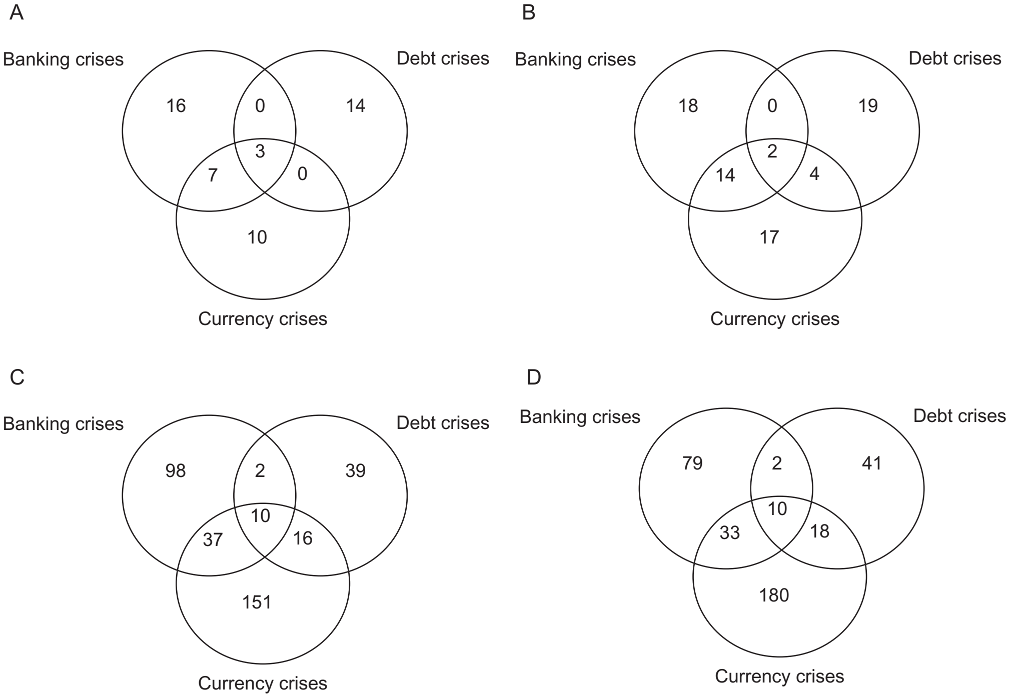


Fig. 2 (A) Coincidence of banking, currency, and debt crises, 1880–1913 (Bordo et al.). (B) Coincidence of banking, currency, and debt crises, 1919–39 (Bordo et al.). (C) Coincidence of banking, currency, and debt crises, 1970–2012 (Laeven and Valencia). (D) Coincidence of banking, currency, and debt crises, 1970–2012 (Reinhart and Rogoff dates). Notes: *(A–D) Source data for (A) and (B) are Bordo, M.D., Eichengreen, B., Klingebiel, D., Martinez-Peria, S. 2001. Is the crisis problem growing more severe? Econ. Policy 16 (32), 52–82. Source data for (C) is Laeven and Valencia (2013). Source data for (D) is Reinhart and Rogoff (2009).*

fraction of crises in the pre-World War I era could be classified as twin or triple crises. As Fig. 2A shows, 50% of the recorded currency crises prior to 1913 were accompanied by a banking crisis.

4.4 Classification Uncertainty: Definitions and Disagreements in Crisis Dates

In our view, the leading chronologies are those based on data underlying Bordo et al. (2001), Reinhart and Rogoff (2009), and Laeven and Valencia (2012). The dataset provided by Taylor (2015), which underlies Jordà et al. (2011), is somewhat limited and less comparable since it restricts attention to "systemic crises" only for a small set of 17 advanced countries. The other three datasets allow researchers to separate currency, banking, debt, twin, and triple crises, each of which are important phenomena in their own right. Two questions immediately arise. How well do these sources agree on their documented dates, and which source(s) is(are) the best?

In answer to the first question, regarding agreement, there is some significant evidence that the correlation between dating methodologies is not extremely high even within constant country samples. Tables 2a–2d show cross tabulations of banking crisis indicators for each of four sources (Bordo et al., Reinhart and Rogoff, Taylor, and Laeven and Valencia) for four different periods (1880–1913, 1919–39, 1945–72, and the years after 1973). We restrict attention in these tables to the first year of a banking crisis for a country.

In each subtable, we show the number of noncrisis country-years, and the number of country-years with a crisis in either of two datasets for the countries that are common to both datasets. The entry in row 2 column 2 of each table records the number of times both datasets are in agreement, and the last two columns provide a measure of the agreement between sources calculated as the percentage of all crisis-years dated within the period and the country sample in which the two sources agree. We provide this

Table 2a Comparison of Leading Crisis Chronologies, 1880–1913

	1880–1913	Reinhart and Rogoff		% Agree	
Pre-WWI	**Bordo et al. vs RR**	**No crisis**	**Banking crisis**	**Same year**	**±1 Year**
Bordo et al.	No crisis	681	17	0.33	0.38
	Banking crisis	5	11		
	21 Countries (21 in Bordo et al. and 70 in Reinhart and Rogoff)				
	1880–1913	**Taylor**		**% Agree**	
	RR vs Taylor	**No crisis**	**Banking crisis**	**Same year**	**±1 year**
Reinhart and Rogoff	No crisis	533	16	0.36	0.55
	Banking crisis	13	16		
	17 Countries (70 in Reinhart and Rogoff and 17 in Taylor)				
	1880–1913	**Taylor**		**% Agree**	
	Bordo et al. vs Taylor	**No crisis**	**Banking crisis**	**Same year**	**±1 Year**
Bordo et al.	No crisis	538	20	0.30	0.41
	Banking crisis	8	12		
	17 countries (21 in Bordo et al. and 17 in Taylor)				

Table 2b Comparison of Leading Crisis Chronologies, 1919–39

	1919–39	**Reinhart and Rogoff**		**% Agree**	
Interwar	**Bordo et al. vs RR**	**No crisis**	**Banking crisis**	**Same year**	**±1 Year**
Bordo et al.	No crisis	409	14	0.31	0.34
	Banking crisis	8	10		
	21 Countries (21 in Bordo et al. and 70 in Reinhart and Rogoff)				
	1919–39	**Taylor**		**% Agree**	
	RR vs Taylor	**No crisis**	**Banking crisis**	**Same year**	**±1 Year**
Reinhart and Rogoff	No crisis	321	2	0.69	0.74
	Banking crisis	9	25		
	17 Countries (17 in Taylor and 70 in Reinhart and Rogoff)				
	1919–39	**Taylor**		**% Agree**	
	Bordo et al. vs Taylor	**No crisis**	**Banking crisis**	**Same year**	**±1 Year**
Bordo et al.	No crisis	323	5	0.65	0.87
	Banking crisis	7	22		
	17 countries (21 in Bordo et al. and 17 in Taylor)				

Table 2c Comparison of Leading Crisis Chronologies, 1950–72

Bretton	**1950–72**	**Reinhart and Rogoff**		**% Agree**	
Woods	**Bordo et al. vs RR**	**No crisis**	**Banking crisis**	**Same year**	**±1 Year**
Bordo et al.	No crisis	539	0	1.00	1.00
	Banking crisis	0	0		
	21 Countries (21 in Bordo et al. and 70 in Reinhart and Rogoff)				
	1950–72	**Taylor**			
	RR vs Taylor	**No crisis**	**Banking crisis**	**% Agree**	**±1 Year**
Reinhart and Rogoff	No crisis	391	0	1.00	1.00
	Banking crisis	0	0		
	17 Countries (17 in Taylor and 70 in Reinhart and Rogoff)				
	1950–72	**Taylor**			
	Bordo et al. vs Taylor	**No crisis**	**Banking crisis**	**% Agree**	**±1 Year**
Bordo et al.	No crisis	391	0	1.00	1.00
	Banking crisis	0	0		
	17 countries (21 in Bordo et al. and 17 in Taylor)				

Table 2d Comparison of Leading Crisis Chronologies, 1973–2012

		Reinhart and Rogoff		% Agree	
Post-Bretton Woods	**1973–97** **Bordo et al. vs RR**	**No crisis**	**Banking crisis**	**Same year**	**±1 Year**
Bordo et al.	No crisis	1171	25	0.37	0.37
	Banking crisis	9	20		
	49 Countries (55 in Bordo et al. and 70 in Reinhart and Rogoff)				
		Taylor		**% Agree**	
	1973–2010 **RR vs Taylor**	**No crisis**	**Banking crisis**	**Same year**	**±1 Year**
Reinhart and Rogoff	No crisis	614	6	0.59	0.70
	Banking crisis	7	19		
	17 Countries (17 in Taylor and 70 in Reinhart and Rogoff)				
		LV		**% Agree**	
	1973–97 **Bordo et al. vs LV**	**No crisis**	**Banking crisis**	**Same year**	**±1 Year**
Bordo et al.	No crisis	1308	12	0.26	0.26
	Banking crisis	19	11		
	55 countries (55 in Bordo et al. and 162 in Laeven and Valencia)				
		Taylor		**% Agree**	
	1973–97 **Bordo et al. vs Taylor**	**No crisis**	**Banking crisis**	**Same year**	**±1 Year**
Bordo et al.	No crisis	407	6	0.39	0.39
	Banking crisis	5	7		
	17 Countries (55 in Bordo et al. and 17 in Taylor)				
		LV		**% Agree**	
	1973–2011 **RR vs LV**	**No crisis**	**Banking crisis**	**Same year**	**±1 Year**
Reinhart and Rogoff	No crisis	2520	24	0.26	0.29
	Banking crisis	51	27		
	70 Countries (70 in Reinhart and Rogoff and 162 in Laeven and Valencia)				
		LV		**% Agree**	
	1973–2010 **Taylor vs LV**	**No crisis**	**Banking crisis**	**Same year**	**±1 Year**
Taylor	No crisis	618	3	0.54	0.59
	Banking crisis	10	15		
	17 Countries (17 in Taylor and 162 in Laeven and Valencia)				

Notes: Tables present cross tabulations of banking crisis indicators for each of four sources (Bordo et al., Reinhart and Rogoff, Taylor, and Laeven and Valencia) in four periods. We restrict attention to the first year of a banking crisis for a country. In each entry, we show the number of noncrisis country-years and the number of country-years with a crisis in either of two datasets for the countries that are common to both datasets. The entry in row 2 column 2 of each table records the number of times both datasets agree. The last two columns provide a measure of the agreement between sources calculated as the percentage of all crisis-years dated within the period and the country sample in which the two sources agree. We provide this percentage for crises occurring in the same year and then allow for a 1-year window to allow for small variations in timing.

percentage for crises occurring in the same year and then allow for a 1-year window to allow for small, but reasonable variations in timing.

The minimum percentage in Tables 2a–2d is 0.26 (comparing Bordo et al. and Laeven and Valencia for 55 countries in the years 1973–97). For the years excluding the Bretton Woods period when all datasets are in agreement, the maxima are 0.69 (Jordà et al. vs Reinhart and Rogoff for 17 countries 1919–39) and 0.65 (Bordo et al. vs Jordà et al. for 17 countries 1919–39). The average percentage of times that the head-to-head comparisons agree is 0.42 excluding the Bretton Woods period where agreement is nearly complete.

Matters are obviously slightly worse in terms of correspondence than these numbers suggest since the figures are calculated based only on overlapping samples of countries. The fact that Reinhart and Rogoff and Laeven and Valencia provide much larger samples means that the absolute number of crises reported will be higher. Frequencies also vary as seen in Fig. 1A–D.[u]

Disagreement and classification uncertainty among datasets exists for several reasons. Datasets differ on their definitions of what constitutes a particular kind of crisis. As seen in Table 1, definitions of banking crises vary substantively across researchers. In addition, since we have divided the sample space or possible outcomes into non-overlapping categories (banking, currency, debt, twin, triple, etc.) disagreement for example can occur when one dataset codes a twin crisis but another dataset codes only a banking or currency crisis. A third reason datasets can disagree is due to near-miss timing when one set of authors has a particular date more than two or more years away from another set of authors. Finding the exact timing of an event is also a challenge in periods of high volatility. Latin American countries had prolonged periods of currency distress and banking instability in the 1970s, 1980s, and 1990s leading to such discrepancies.

An example comparing the Reinhart and Rogoff dataset with that of Bordo et al. is germane. The case of Argentina 1973–92 is particularly difficult. Reinhart and Rogoff date a currency crisis in every year from 1973 to 1992 and two banking crises that occur during this extended currency crisis (1980–82) and (1989–90). Bordo et al. have the following: currency crisis (1975–78), twin crisis (1980–82), currency crisis (1984–85), currency crisis (1987), and twin crisis (1989) which was not associated with the currency crisis in 1987. It is evident that different authors take different routes to finding the first year of a given type of crisis and discrepancies emanate from periods of great macroeconomic instability.

[u] Cecchetti et al. (2009) report that all crises in Laeven and Valencia are in Reinhart and Rogoff. However, this cannot strictly be true since Laeven and Valencia have a much larger sample of countries. In addition, we separate banking crisis from twin and triple crises which Cechetti et al. did not do.

Another issue with the historical dating of crises is that authors must rely on the research of other economic historians or significant amounts of scattered primary sources from multiple country-specific sources. Often historical sources are vague as to how many financial institutions were closed or faced runs which leads to discrepancies in the dating of banking crises. Jalil (2015) studies six leading chronologies of the American banking system in the 19th and early 20th centuries and observed major disagreements among them. Jalil argues that "quantitative sources alone are not sufficient to identify banking panics" and carries out an extensive and careful reading of contemporary sources to identify banking panics (as opposed to systemic banking crises).

Matters are significantly worse for dating currency crises in history especially in the 19th century. As it turns out, finding reliable exchange rate data for samples outside of the leading 21 countries is extremely difficult, if not impossible, since active and liquid markets in foreign exchange did not exist without some prior financial development. Reinhart and Rogoff (2009) provide an extensive list of dates for which nominal exchange rates are available with some relevant cases in point such as: Argentina (available from 1880 only), Finland (from 1900), Korea (1905), Greece (1872), New Zealand (1892), South Africa (1900), Uruguay (1900), etc. In other cases, using an EMP index will be difficult prior to the 1930s or even the 1950s due to missing reserve and interest rate data. In these cases, the Frankel and Romer approach of using a cutoff for changes in the nominal exchange rate will have to suffice. Relying exclusively on exchange rate changes however neglects many important episodes.

More disconcerting is the disagreement on sovereign defaults in the period since 1973. These data are mainly gathered from primary and secondary sources as noted in Laeven and Valencia (2012) and Reinhart and Rogoff (2009). While Reinhart and Rogoff find 64 defaults between 1973 and 2009, Laeven and Valencia (2012), *in the same set of countries*, only find 34. This is not simply a matter of widening the window of years in which a default is classified. Many defaults in Reinhart and Rogoff such as Algeria (1991), Brazil (2002), Uruguay (1990), etc., are not recorded in the Laeven and Valencia dataset. This is due to the fact that Laeven and Valencia record an event only when a payment is missed. Reinhart and Rogoff seem to follow a more lenient approach classifying events where there are ratings downgrades, loss of market access, issues of confidence, etc.[v] This classification is seemingly at odds with the seemingly stricter definition described in Table 1 and Reinhart and Rogoff (2011). Such discrepancies across authors obviously seriously impinge on interpretation of the historical record and inferences drawn about the frequency, duration, costs, and causes of crises.

[v] This conclusion is drawn from direct email correspondence with Luc Laeven on February 11, 2016. He cites the case of Brazil in 2002 as an example where no payments were missed but Reinhart and Rogoff declare a debt crisis.

4.5 Causes of Crises

With the reappearance of financial crises associated with financial liberalization in the 1970s and 1980s and better international data, researchers began to focus energy on isolating the leading determinants of financial crises. Theory and analytical frameworks developed in the 1970s and 1980s provided guidelines for the key variables of interest, but explicit structural tests of particular models still remain few and far between. Most of the research in this area focuses on a large set of macroeconomic, financial, and international variables and attempts to exclude variables with the lowest statistical power for predicting financial crises.

Subsequent to these early efforts, a new (near-consensus) view emerged, based in part on the experience of the 2007 crisis, assigning a primary role to credit booms as the key determinant and predictor of financial crises (eg, Borio and Drehman, 2009; Schularick and Taylor, 2012; Gourinchas and Obstfeld, 2012). Notwithstanding this view, it is appropriate to recognize that not all banking crises are driven by credit booms. It is also useful to recognize that not all housing, equity booms, or capital inflow bonanzas end in crises.[w] A more satisfactory approach to understanding the drivers of financial crises recognizes that the microstructure of the financial system matters as well as credit's interaction with a number of other macroeconomic determinants.

Four key approaches to understanding the causes of crises have been taken since the 1990s with subsequent refinements in recent years. The first approach uses cross-country data and limited dependent variable models such as logit or conditional logit to find statistically significant determinants (eg, Demirgüç-Kunt and Detragiache, 1998). Kaminsky et al. (1998), Kaminsky and Reinhart (1999), and Kaminsky (1999) show that the early warning indicators (EWI/EWS) methodology developed for predicting the business cycle can be satisfactorily employed. In addition, qualitative and descriptive analyses as well as "Big Data" methods have been used.

In a highly influential paper, Kaminsky and Reinhart (1999) apply and adapt the EWI approach to predict banking, currency, and twin crises in a sample of 20 countries between 1970 and 1995. From a large set of variables, they select 16 as the most important based on their changes in the months preceding and following the different types of crises in play. These variables are classified into four categories (financial sector, external sector, real sector, and fiscal sector). Kaminsky and Reinhart (1999) check whether a variable signaled a crisis within a particular window of time (12 months for a banking crisis and 24 months for a currency crisis) and then find thresholds by finding the level or

[w] See Bordo and Landon-Lane (2014), Jordà et al. (forthcoming) and Goetzmann (2015) for perspectives on housing booms and financial crises. Gorton and Ordoñez (2016) study *good booms* and *bad booms* arguing that not all credit booms end in crises.

change in a variable that minimizes the noise-to-signal ratio. In this way, a sophisticated use of information criteria is used to balance type I vs type II errors.

Two tradeoffs relevant to policy makers are immediately evident in the context of this strand of the EWI literature. First, what is the optimal size of the prediction window? Calling a crisis too early could put the brakes on an otherwise healthy economy, but failing to act at an early date might preclude avoiding a crisis. Another tradeoff concerns the loss function for policy makers. Babecký et al. (2014) note that minimizing the noise-to-signal indicator ignores the relative losses (to the policy maker) from missed crises vs false alarms. In this case, the optimal threshold would depend on the relative weights in a loss function as well as the predictive power of the signals. Obviously such a calculation has direct relevance for macroprudential policy. Jordà et al. (2011) illustrate the tradeoffs in finding an optimal threshold for any given indicator of a crisis with a correct classification frontier (CCF) akin to a production possibilities frontier trading off type I and type II errors. They suggest that the area under their CCF, which is equivalent to the AUC criteria, be used in determining whether a particular model has predictive power.

Returning to the pioneering study of Kaminsky and Reinhart, their data suggest that growth of money and interest rates are above trend before crises, while an appreciating real exchange rate and exports below trend help predict crises. In addition, in their sample, output falls below trend prior to a crisis. The best predictors (ie, those with the lowest noise-to-signal ratio) for banking crises are: appreciation of the real exchange rate, equity price booms, and the money multiplier. The lowest type I error (missed crises) is provided by high real interest rates which were strongly associated with financial liberalization especially in the 1980s.

Recent research in a similar vein (Babecký et al., 2014; Drehman et al., 2012; Gourinchas and Obstfeld, 2012) emphasizes the financial cycle highlighting above trend growth in the ratio of domestic private credit-to-GDP, equity, and property prices. The earlier literature (eg, Kaminsky and Reinhart, 1999) did not deny that credit was important, but in Kaminsky and Reinhart's sample the percentage of crises correctly called by this variable (when above its threshold) is only 50% while its noise-to-signal ratio was a relatively low 0.59. While these results are not equally comparable to the results in Schularick and Taylor (2012) who argue that credit is a very strong predictor of crises alone and of itself, it would appear that the role of credit depends on the particular sample chosen and definition of a crisis.

Another strand of the literature attempted to predict banking crises using logit analysis. Demirgüç-Kunt and Detragiache (1998) find the following:

– Low GDP growth, high real interest rates, and high inflation are significantly correlated with the occurrence of a banking crisis.
– Banking sector variables such as the ratio of broad money to foreign exchange reserves, credit to the private sector, and mismanaged liberalization are associated with banking crises.

- The level of GDP/capita is negatively related to crises.
- Deposit insurance which is overly generous may also be associated with moral hazard and banking instability.

Subsequent work by the same authors emphasized the role of financial liberalization in environments with weak regulatory capacity and generally weak institutions giving rise to corruption, weak rule of law, and poor contract enforcement. This result echoes the general experience we highlight above that deposit insurance and guarantees have fomented regulatory forbearance and in many instances this has led to banking crises.

Research over the past 5–10 years has made refinements to the general methods proposed earlier. For instance, Bussiere and Fratzscher (2006) note that a binary logit model that predicts the onset of a crisis may ignore the fact that macroeconomic variables and relationships may behave differently in the wake of a crisis. Using information from the quarters immediately following a crisis to predict a crisis may thus lead to poor predictions. Instead of running binary response models to predict banking crises, Bussiere and Fratzscher estimate a multinomial response which allows for three states of the world: noncrisis, crisis, and postcrisis. Better prediction is also achieved by using the tools from the model selection literature. In this vein, Babecký et al. (2014) use Bayesian averaging of regression estimates to select the strongest set of determinants rather than focusing on a small set of indicators.

Other authors have moved the goalposts slightly by incorporating more information to generate continuous indicators of crises. Rose and Spiegel (2011) and Babecký et al. (2013) cover banking crises while Frankel and Saravelos (2012) study currency crises. Here the indicators of crises incorporate information on the severity of the crisis. Rose and Spiegel (2012) study a multiple indicators multiple causes model to study the crisis of 2007–08. Their indicators are quite distinct from the simple binary banking crisis indicators in Table 1. Instead their indicators focus on the output drop between 2007 and 2008, exchange rate movements, credit ratings changes, and equity price changes. The advantages in such procedures are that one does not need to use limited dependent variable models which are biased in the case of rare events and deal poorly with unobservable heterogeneity. However, the regressand of interest has changed significantly in terms of the economic meaning compared to the previous literature on EWIs.

The conclusions from this strand of the literature are not entirely consistent with the earlier EWI literature. Frankel and Saravelos (2012) conclude that foreign exchange reserves, the real exchange rate, the growth rate of credit, GDP, and the current account are the most frequent statistically significant indicators (of currency crises) in the literature reviewed. While somewhat consistent with Kaminsky and Reinhart, the finding obviously has little light to shed on the correlates of banking crises. Rose and Spiegel (2011, 2012) use indicators that overlap with those in Kaminsky and Reinhart but find little predictive power for any of the leading causes explored in Kaminsky and Reinhart—most notably the level of the ratio of domestic credit-to-GDP. Since the crisis of 2007 manifested itself in different ways in different countries, it is not surprising that Rose and Spiegel (2011, 2012) find few reliable predictors to explain the diversity of experience. Exploration of this issue by Sayek and Taksin (2014) provides interesting evidence. In fact, the Eurozone crises were different from each other, but they bore some similarity to previous crises (ie, a small set of variables moved in the same way prior to these and previous crises). They emphasize that each recent crisis has a reasonable match to other historical crises in the model, but that each match is likely to be to a separate and distinct crisis. In other words, the causes of the most recent crisis are heterogeneous and not driven by one particular variable.

IMF (2009) discusses some of the differences in the way the recent crisis unfolded across countries, and hence why EWI analysis may be challenging. While some countries were exposed to offshore borrowing, others had significant housing booms. Still others had cross-border assets in the United States—the epicenter of the crisis, and several countries had unsustainable fiscal and financial problems (as discussed earlier) as well as policy constraints (eg, countries in the EMU). While one might infer from some of the recent theoretical and empirical literature that growth of the ratio of domestic credit-to-GDP is the key to understanding financial crises, this does not appear to be the only, nor the main, determinant of crises over the last few decades. Consequently, a focus solely on credit may not go very far in helping us understanding the recent crisis or future crises.

In addition, a large debate exists on the role of capital inflow surges vs credit booms. Both factors were cited as potential risks in the run up to the 2007 crisis and have perennially been in the spotlight in the empirical and theoretical literature. Many conceptual frameworks suggest that capital inflows fuel lending booms in open economies (Borio et al., 2014; Diaz-Alejandro, 1985; McKinnon and Pill, 1986). In a widely cited study, Jordà et al. (2011) find that prior to 1945 current account deficits are associated with systemic crises, but that after 1945 this is no longer the case. These authors cite the growth of the ratio of credit-to-GDP as a key determinant and a good predictor of systemic crises. In addition, they find no evidence that capital inflows which coincide with credit booms raise the probability of a systemic crisis.

In opposition to these findings, Caballero (2014) reexamines the connections between credit and capital flows in an interesting empirical treatment finding a role for both capital inflow surges and credit booms. Caballero (2014) uses a limited dependent variable model that allows for unobservable country level heterogeneity and yet allows for noncrisis countries to appear in the sample, unlike in Jordà et al. (2011) who use conditional logit models. Caballero's sample covers a large sample of countries (both developed and less developed) from 1973 to 2008 while Jordà et al. (2011) feature data from 14 countries 1870–2008.

Caballero finds that inflow bonanzas and credit booms are both statistically significant predictors of banking crises (using a similar definition of crises to Laeven and Valencia). However capital inflow bonanzas do not solely operate through bank intermediated credit booms. Moreover, portfolio-equity capital flows, but not bonanzas associated with a rise in net debt are statistically significant. The evidence here suggests capital inflow bonanzas may also generate instability by fueling asset price booms and enhancing liquidity. Based on these results, it seems imprudent to ignore the role of large capital inflows even if domestic private credit is not growing above trend.

Could monitoring of the credit-to-GDP ratio have been sufficient to avoid the recent global crisis? It is unlikely. Caballero (2014) as well Babecky et. al. (2014) and IMF (2009) suggest a more eclectic approach that simultaneously incorporates multiple variables. In addition, IMF (2009) emphasizes that while the subprime crisis is often thought of as a credit-driven event accompanied by unsustainable levels of leverage this characterization is at odds with the data. In the United States, for instance, private domestic credit-to-GDP did not grow strongly above trend because borrowing by corporates eased while household debt and leverage increased.[x] Contrariwise, the East Asian crisis was associated with strong growth of leverage in the corporate and financial sector and less so in the household sector (IMF, 2009). In addition, leverage among financial institutions in the 1990s was as high as in the years before 2007 (Portes, 2010). It should not be forgotten that in the recent crisis, leverage in many financial institutions was hidden in off-balance sheet vehicles and so forth giving credence to the idea that Goodhart's law is in play. Ideally, any surveillance of the financial system, any conclusions regarding causes of crises, and any macroprudential policy would pay close attention to where exactly risk was concentrated and where maturity mismatches are the most pronounced. Surveillance and policy must also carefully weigh the costs and benefits of imposing policy in light of the potential for type I and type II errors as highlighted in this literature.

4.6 Output Losses of Financial Crises

Financial crises are often associated with economic downturns and deviations of output from long-run trends. A large number of studies investigate the impact of crises on output, output growth, other macroeconomic aggregates, and even health indicators (Stuckler et al., 2012).

Table 3 provides a list of leading papers, methodologies, and baseline estimates for the impact of financial crises on output.[y] In this literature, most authors define output losses as

[x] Jordà et al. (2013) note that the level of "excess credit" was in the 60th percentile for all crisis events in their data and emphasize the issues related to shadow banking. They emphasize that a broader measure of credit shows a more significant boom.

[y] The Basel Committee on Banking Supervision, Bank for International Settlements (2010) also provides a similar table and calculates median and average output losses across studies (with no "permanent" effects) as 19%. See table A1.1 in that paper.

Table 3 Definitions and Values of Output Losses from Financial Crises

Authors	Sample	Crisis definition	Methodology for calculating the economic costs of financial crises	Average "losses"
Bordo et al. (2001)	1880–1939 21 Advanced countries 1945–97 21 Advanced countries + 35 LDCs and emerging markets	Banking crises	Cumulative loss of output between onset and recovery found by subtracting actual growth from pre-crisis trend growth. Recovery occurs when growth obtains its precrisis trend level.	7% (21 countries, 1973–97) 6.2% (56 countries, 1973–97)

Hoggarth et al. (2002)	1977–98
	47 Countries
	47 Banking crises
		Banking crises (systemic and borderline)	**1.** *GAP1* Sum of the differences between growth in potential output and actual output growth during the crisis period. Potential growth = arithmetic average of GDP growth in the 3 years prior to the crisis. End of crisis is when output growth returns to trend.
			2. *GAP2* Cumulative difference between level of potential output and actual output over the crisis period. Output potential is based on trend growth over the 10-year precrisis period using an HP filter.
				GAP1 = 14.5%
				GAP2 = 16.5%

Hutchison and Noy (2005)	1975–97 24 Emerging markets	Twin crises	Regressions of growth of real GDP on crisis indicators and lags.	Average loss of GDP of 15–18% over the average

Continued

Table 3 Definitions and Values of Output Losses from Financial Crises—cont'd

Authors	Sample	Crisis definition	Methodology for calculating the economic costs of financial crises	Average "losses"
				duration of 3–4 years after the onset of a crisis.
Dell'Ariccia et al. (2008)	1980–2000 41 Countries 48 Crises	Banking crises: there were extensive depositor runs; the government took emergency measures to protect the banking system, such as bank holidays or nationalization; the fiscal cost of the bank rescue was at least 2% of GDP; or nonperforming loans reached at least 10% of bank assets.	Marginal impact of banking crises on the annual growth rate of sectoral value added.	Growth rate is 1.1 percentage points lower in sectors highly dependent on external finance.
Angkinand (2009)	1970s–2003 35 Countries 47 Crises (systemic and nonsystemic)	Banking crises identified in Caprio and Klingebiel (2003).	Cumulative deviation in real GDP from an extrapolated HP trend. Calculated between the onset of a crisis and time when GDP reaches the trend.	3.13% (mean for all banking crises) 3.99% (mean for systemic banking crises)
Cecchetti et al. (2009)	1980–2007 Number of countries is not stated 40 crises	Banking crisis defined as in Laeven and Valencia.	Output loss is the cumulative loss in GDP from the onset of a crisis until GDP reaches the precrisis peak.	18.4% (mean) 9.2% (median)
Laeven and Valencia (2013)	1970–2011 162 Countries	Systemic banking crises possibly accompanied by currency, or debt crises or both.	Cumulative loss of real GDP between onset of crisis and 3 years after crisis starts calculated as the difference between actual output and the HP filter trend	23% (median) 32% (median advanced) 26% (median emerging markets)

			calculated over the 20 years prior to a crisis (or fewer years if data are not available)	
Jordà et al. (2013)	1870–2008 14 Countries	"Financial Recessions" (ie, recessions associated with systemic financial crises) with and without large growth in real credit.	Local projections from year $T+1$, to $T+5$ of log differences of GDP per capita in year t from peak year level.	16.9% Cumulative deviations from peak for "financial recessions" for $T+1$ to $T+5$ (table 7 row 1, p. 19)
Reinhart and Rogoff (2014)	1800–2011 70 Countries	100 Systemic banking crises defined as in Reinhart and Rogoff (2009) possibly accompanied by currency, or debt crises or both.	**1.** Peak to trough decline in GDP per capita. **2.** Severity index $=-1\times$ (peak to trough decline in GDP per capita) + number of years until peak level of GDP per capita is attained. This is defined as recovery time.	11.5% (mean) 8.8% (median) 8.3 Years peak to recovery (mean) 6.5 Years peak to recovery (median)
da Rocha and Solomou (2015)	24 Countries 1920–38 19 Crises	Systemic banking crises: "Classification is based on qualitative informed judgment, documenting the extent of financial distress in the banking system of a country."	Cumulative growth in real GDP and industrial production up to 7 years after a crisis starts.	33% Cumulative deviations from peak for T to $T+7$.

the deviations from a precrisis peak in output or a precrisis output trend. However, there is substantial variation in the methodologies used for calculating the costs of financial crises. Some authors study the marginal impact of crises and financial distress on growth rates. Others calculate the cumulative loss of output or GDP per capita from the peak of economic activity at various postpeak window lengths. Differences in methodologies, dependent variables, and samples lead to significant differences in the point estimates of the output costs of financial crises. Still, nearly all studies agree that financial crises are associated with economically significant downturns in output and output growth.

One major issue in determining the size of the output losses attributable to a crisis is causality or endogeneity. Real shocks may cause an output decline and problems in the financial sector, but equally, financial shocks are widely believed to generate output declines.[z] The problem comes down to identifying the sources of variation in outcomes when unobservable financial frictions and shocks matter and are correlated. More precisely demand and supply may change in response to the same shocks that contribute to financial problems thus making it difficult to cleanly identify the impact of the financial shock itself. Empirically, Reinhart and Rogoff (2014) observe in their sample that the peak of economic expansions usually coincides with banking crises but that in several instances the peak predates the crisis. Calomiris and Hubbard (1989) argue that output turns down prior to difficulties in the financial sector.

Two main approaches have been taken to deal with heterogeneity, unobservable factors, and endogeneity bias. Bordo et al. (2001) compare recessions without financial crises to recessions with financial crises. After controlling for a small set of observables, the authors find that financial crises are associated with higher output losses than recessions without financial crises. In a similar vein, Jordà et al. (2013) report statistically and economically significant differences between output downturns associated with financial crises and downturns not associated with financial crises even after conditioning on a number of predetermined macroeconomic variables. Jordà et al. (2011) also find that output losses in financial recessions are positively associated with the size of the precrisis rise in the ratio of credit-to-GDP.

Another way of dealing with causality is to put more theoretical structure on the data. Dell'Ariccia et al. (2008) argue that if financial sector distress matters then it should be the case that sectors which are more dependent on external finance should be the hardest hit when the banking sector is in trouble. Their evidence is consistent with this line of reasoning. Mladjan (2012) provides similar evidence for the Great Depression. In addition, Ziebarth (2013) found quasiexperimental evidence from the 1930s that where bank

[z] In general, the theoretical literature in macroeconomics shows how output losses due to shocks can be amplified in the presence of financial frictions. Financial shocks arise from capital quality shocks in a model with financial frictions as in Gertler and Karadi (2011). See also Bernanke et al. (1999).

failures were larger these were associated with greater declines in output, lower revenue, and a slower pace of entry by firms.

We provide some baseline estimates for output losses that are comparable in terms of methodology. We use crisis data from Bordo et al., Reinhart and Rogoff, and Laeven and Valencia, and data on output per capita for 42 countries between 1865 and 2009 from Barro and Ursúa (2008). For the period 2000–14, we use real GDP per capita (in local currency) from the World Economic Outlook database in order to calculate trends and output losses from the recent crisis that started in 2007. We calculate unconditional output losses in different periods using the crisis dates from the various datasets surveyed in Section 4. In particular, we study the cumulative percentage deviation of GDP per capita from the precrisis trend level of GDP per capita. The window we use is the year of the crisis to 3 years after the crisis starts. Precrisis trends are given by the average annual change of the logarithm of GDP per capita in up to 10 years prior to a crisis.[aa]

We provide these losses for banking, twin, and triple crises in Fig. 3A–D. Fig. 4A–F provides illustrations of specific country examples. Output losses, as we define them here, are economically very large. In the period 1880–1913, Fig. 3A shows that for banking crises average output losses are nearly 3% in BEKM's data (median = 0.20, standard deviation = 38.9) and 6% in RR (median = 5, standard deviation = 33). Losses are much larger in the interwar, largely driven by the Great Depression. Here, for the three different types of crises, losses are never lower than 40% in the BEKM dataset. In the post-Bretton Woods period, losses in the BEKM and RR datasets are smaller than the interwar period but larger than the 1880–1913 period. Here the average losses are on the order of 14% in the BEKM data (median = 18, standard deviation = 23), 21% in the RR data (median = 24 standard deviation = 28), and 29% in the LV data (median = 30, standard deviation = 28). The higher losses in the LV dataset stem from the inclusion of a wider range of countries and the inclusion of the crisis of 2007 which witnessed much higher output losses than previous crises.

Output losses are different in size due to different methodologies in calculating trends, in dating crises, defining the type of crisis of interest, and country/time coverage. When we restrict the sample to BEKM's 56 countries and the years 1973–97 which are covered in all datasets, then the output losses from banking crises are calculated as 14% (BEKM), 15% (RR), and 19% (LV). Using different sample years and countries lead to different headline numbers as is immediately obvious. In addition, LV use GDP and not per capita GDP, although, in practice this has only a minimal effect. LV also use an HP filter whereas we have opted for a simplified exponential detrending procedure. While we find

[aa] We eliminate crises that occur within 3 years of another crisis. Previous crises may have an impact on the trend and level of output. We also estimate losses separately for banking crises without currency and currency and debt crises so as to separate the sample space into mutually exclusive bins as above.

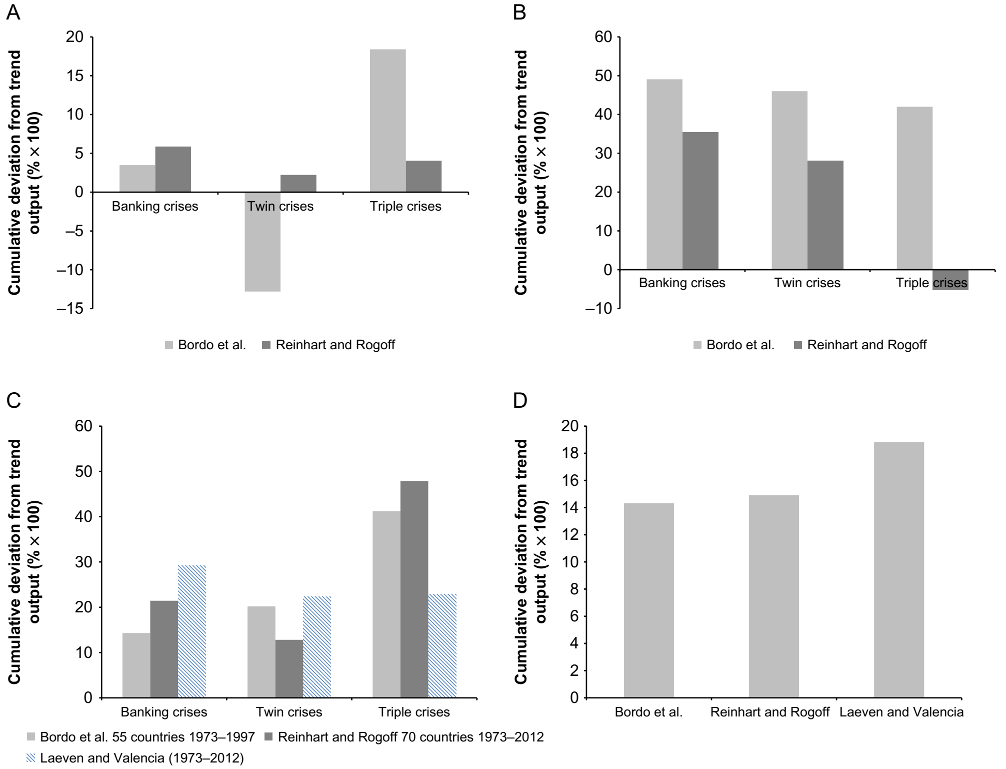


Fig. 3 (A) Output losses for three varieties of crises, 1880–1913 Bordo et al. vs Reinhart and Rogoff. (B) Output losses, three varieties of crises, 1919–39 Bordo et al. and Reinhart and Rogoff. (C) Output losses, three varieties of crises, 1973–97 (Bordo et al.), 1973–2012 (Reinhart and Rogoff), and 1973–2012 (Laeven and Valencia). (D) Output losses from banking crises, 1973–97 three datasets. *Notes*: (A–D) Output losses are calculated as the difference between the level of GDP per capita in the 3 years following a crisis and the extrapolated trend of GDP per capita. The trend is calculated as the average growth rate in the 10 years prior to crisis. See the text for additional information.

some instances where losses are not positive (ie, output per capita is not below trend), probably because the pretrend is already quite low, LV report no instances where this is the case. It appears that the lag length for calculating the trend also matters. Jordà et al. (2013) study deviations from the business cycle peak (table 5, p. 13, table 6, p. 15, and table 7 p. 19). This could lead them to offer smaller losses since no assumptions are made regarding the continuing trend of GDP per capita.

One surprise when looking at the long run is that output losses seem to be larger in the recent period compared to the pre-World War I period despite today's greater reliance on liquidity support, fiscal interventions, and other policies which attempt to remedy the market failures associated with financial shocks. However, compared to the interwar

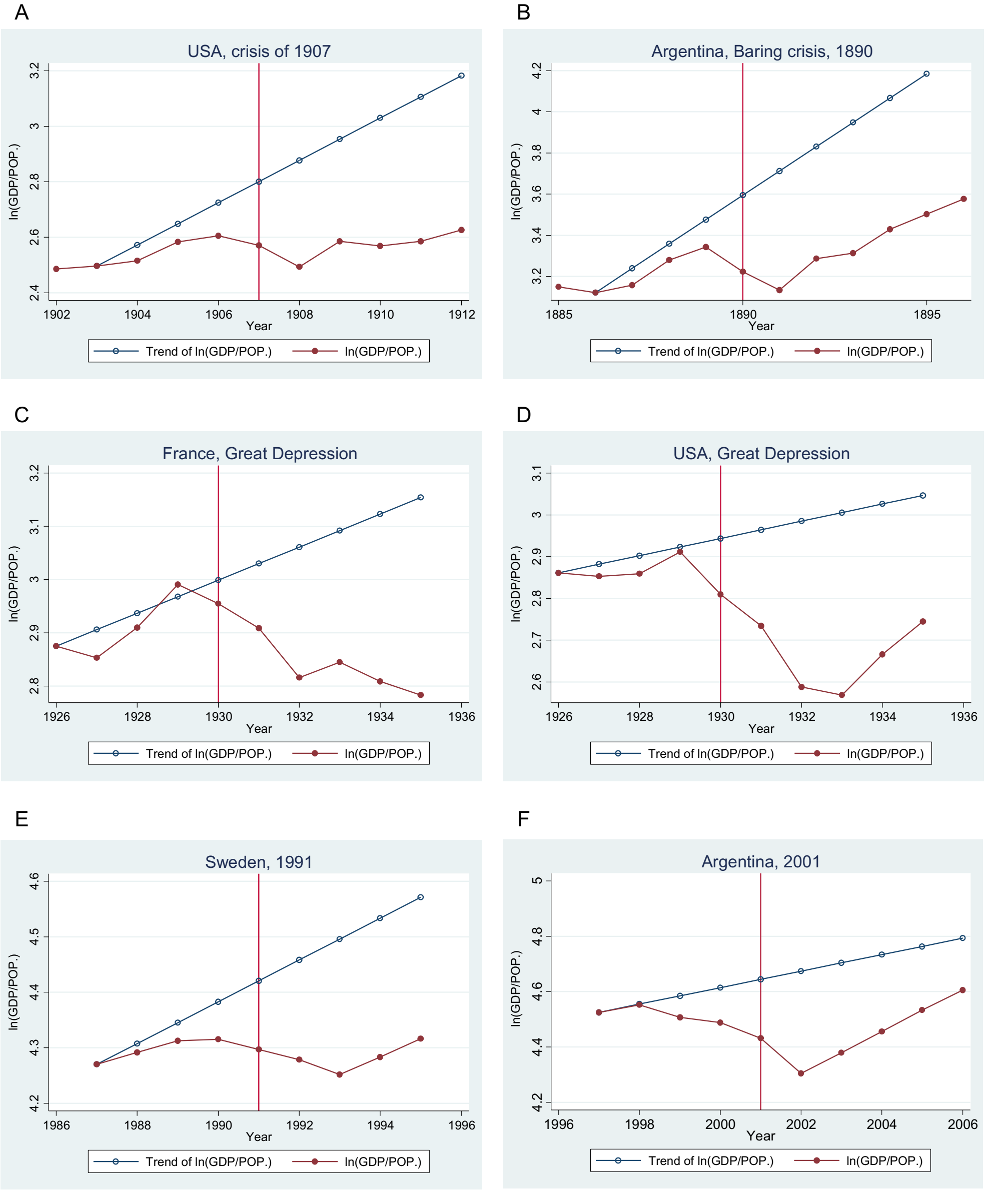


Fig. 4 (A) GDP per person actual and counterfactual, United States, 1907. (B) GDP per person actual and counterfactual, Argentina, Baring crisis. (C) GDP per person actual and counterfactual, France, Great Depression. (D) GDP per person actual and counterfactual, United States, Great Depression. (E) GDP per person actual and counterfactual, Sweden, 1991. (F) GDP per person actual and counterfactual, Argentina, 2001. Notes: *(A–F) Data are underlying Bordo et al. except for (F). Data for real GDP per capita for (F) are from the World Economic Outlook database. Trend (counterfactual) line is calculated based on simple extrapolation of the average growth rate in the previous 10 years.*

period/Great Depression years, a period when policy was counterproductive, the losses from banking crises are lower on average. Without recent interventions, output losses might have been higher—although without further work and careful research design to sort out endogeneity and selection biases we can take no firm stance on the causal impact of financial distress and systemic banking crises. As for the pre-World War I period, it may be the case that the economies of the time were more flexible or that the financial sector's size was limited mitigating the overall negative impact on output. In the historical period, countries avoided output drops comparable to those today even without a comprehensive crisis-fighting playbook beyond lender of last resort actions and ad hoc rescues. An interesting avenue for future empirical research is to study the size of output losses after properly accounting for variance in policy action.

4.7 The Speed of Recovery After Financial Crises

Reinhart and Rogoff (2009, 2014) posited that recessions with financial crises (ie, financial recessions) are followed by slow recoveries. These authors generally gauge time to recovery as the number of years until the level of real per capita GDP attains the prior peak it reached. Reinhart and Rogoff (2009) study a small sample of "severe" financial crises while Reinhart and Rogoff (2014) study 100 systemic banking crises 1857–2013. Reinhart and Rogoff find that recessions with systemic crises have longer times to recovery than those recessions which are not accompanied by a crisis.

In contrast to the earlier studies, Bordo and Haubrich (2012) posit that financial recessions generally are followed by faster recoveries. They start with Friedman's (1993) plucking model which shows that deep recessions will be followed by fast recoveries. Zarnowitz (1992) documents this stylized fact for the United States. Bordo and Haubrich then compare the recovery from recessions with crises to those that did not have crises for 22 business cycles 1880–2010 in the United States. Bordo and Haubrich measure the depth of the contraction as the percentage drop in quarterly GDP from the peak to the trough of NBER cycles. They then measure the "strength" of the recovery as the percentage change in GDP in the first four quarters of the expansion. Bordo and Haubrich also measured the recovery as the same number of quarters that output declined in the preceding downturn, so eg, if output declined for six quarters they measure the strength of the recovery as the percentage change in GDP in the first six quarters of the expansion. They find that recessions with financial crises (using crisis dates in BEKM) were 1 percentage point deeper than nonfinancial recessions and the recoveries were 1.5 percentage points stronger than recoveries in nonfinancial recessions. Other studies confirm Bordo and Haubrich including: Howard et al. (2011) and Romer and Romer (2015). Results in Jordà et al. (2013) show that the US recovery after 2007 was faster than what would have been predicted by their empirical models.

5. FISCAL CRISES, BANKING CRISES, AND THE FISCAL CRISIS TRILEMMA

Following the research of Reinhart and Rogoff (2009) and after observation of events in Europe, research has focused on the impact of banking crises on the probability of a debt crisis especially in advanced countries. While developing countries faced such troubles from the 1970s, advanced countries largely had fewer and smaller crises until recently. The exceptions being of course Japan, Sweden, and Finland in the 1990s. Reinhart and Rogoff suggest however that public debt (not the debt-to-GDP ratio) increased by about 86% in the wake of banking crises due to the impact of falling revenues. According to Reinhart and Rogoff, these increases, "in several cases," were not wholly due to the fiscal costs of bailouts. Schularick (2012) shows that the (systemic) crises of the late 20th century are associated with large rises in the debt-to-GDP ratio, but that in the same sample of 14 advanced countries, the crises prior to the 1970s were not associated with significant rises in this ratio.

Laeven and Valencia provide a systematic dataset on the rise in debt-to-GDP ratios for all of the banking crises in their dataset. The median rise in the debt-to-GDP ratio for all systemic crises in their data was 12% of GDP while in advanced economies this figure rises to 21.4% of GDP. Fiscal costs, measured as the rise in outlays due to restructuring the financial sector had a median of 6.8% of GDP. Laeven and Valencia subtract the rise in fiscal outlays due to restructuring from the rise in total debt to calculate a rough measure of the degree of discretionary fiscal policy. The median for this variable is 7% of GDP.

Tagkalakis (2013) empirically examines the feedback loop from fiscal policy to financial markets and back in a sample of 20 OECD countries 1990–2010. Fiscal instability leads to financial instability and financial instability leads to fiscal instability via bailouts. Fratzscher and Rieth (2015) using structural VARs with daily financial markets data for 2003–13 confirm the two-way causality between sovereign risk shocks and bank risk. They find that sovereign risk shocks are more important in explaining bank risk than the reverse. In another report carried out by the European Commission (2009), the average unconditional postcrisis rise in the debt-to-GDP ratio was 18.9% of GDP. This figure is cumulative until the "end" of each crisis in the sample and covers 49 crises (Laeven and Valencia dates) for advanced and emerging economies 1970–2007.

The findings in Tagkalakis (2013) are intriguing since it appears that the rise in debt following a financial crisis is larger the bigger the size of the financial sector relative to total output. Laeven and Valencia (2013) also argue that the largest fiscal costs of crises since the 1970s have been in Ireland, Iceland, Israel (1977), Greece, and Japan (1990s).

Putting all of these findings together suggests the possibility that there is a potential tradeoff for countries along the lines of a trilemma. This financial/fiscal trilemma suggests that countries have *two* of the following three choices: a large financial sector, a large bailout package, and a strong discretionary reaction to the downturn associated with financial crises. The logic is as follows by way of an example. Assume a country with a large financial sector faces a banking crisis. If so, then the government can provide a bailout package of a size that is commensurate with the size of the financial sector. If so it uses up its fiscal space. Otherwise, it could lower the size of the bailout and devote its fiscal space to discretionary fiscal policy. With a smaller financial sector, and the same amount of fiscal space, since the size of the bailout would by definition be smaller, the size of the rise in debt due to expansionary policy could rise.

The cases of the United States and Greece post-2007 are illustrative. The United States had a large financial sector, but its bailout, as measured by the fiscal costs was relatively small (4.5% of GDP). On the other hand, the rise in the debt-to-GDP ratio not attributable to the gross costs of the bailout was on the order of 19% of GDP (Laeven and Valencia, 2012). While Greece had a rise in the ratio of debt-to-GDP (after accounting for the fiscal rescue costs) of about 17%, its downturn was much larger and likely merited, based on past experience, a much larger discretionary response. Greece's fiscal costs of the bailout are reported by LV to be 27% of GDP. Obviously, the ability of countries to finance either a bailout or a discretionary package depends on the willingness of capital markets to fund deficits. In this regard, the trilemma would be more applicable or more binding for countries which had better debt sustainability measures at the beginning of their crisis events.

To test the idea of a financial trilemma we used data from Laeven and Valencia (2012) on the change in the ratios of total government debt-to-GDP in the 3 years following a banking, twin, or triple crisis, the fiscal costs of bailouts to GDP and a residual which is the difference between the change in the debt-to-GDP ratio and the ratio of the fiscal costs-to-GDP. We used data for 19 banking crises in the advanced countries since 1970. We omit the case of Switzerland in 2008 since it had a decline in the overall debt-to-GDP ratio and our econometric model is in logs. Also emerging economies had many episodes of declines in the debt-to-GDP due to inflation which poses some issues for our initial exploration in a log-linear regression.

In the spirit of measuring a tradeoff we run the following regression:

$$\ln\left(\Delta\frac{Debt_{it}}{GDP_{it}}\right) = \kappa + \theta_1\left[\ln\left(\Delta\frac{Fiscal\ costs_{it}}{GDP_{it}}\right)\right] + \theta_2\left[\ln\left(\Delta\frac{Discretion_{it}}{GDP_{it}}\right)\right] + \varepsilon_{it} \qquad (1)$$

We do not use panel data methods in this case. Instead we study 19 episodes for the years 1970–2012. Data are for 18 countries (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Japan, Luxembourg, Netherlands, Norway, Portugal, Sweden, United Kingdom, and United States). Fiscal costs are given by Laeven and Valencia and "Discretion" is the change in the debt-to-GDP ratio minus the ratio of fiscal costs-to-GDP. Of course, countries have differing abilities and desires for the target variable "debt" depending on market conditions, political economy issues, and the size of the financial crisis. Here we assume that crises create a need for bailouts, but that when a country spends resources on a bailout that this is associated with a tradeoff in the size of the “discretionary” response.

The results of the regression, with robust standard errors underneath the coefficients in parentheses are:

$$\ln\left(\Delta\frac{Debt_{it}}{GDP_{it}}\right)=\underset{(0.13)}{0.69}+\underset{(0.03)}{0.25}\left[\ln\left(\Delta\frac{Fiscal\ costs_{it}}{GDP_{it}}\right)\right]+\underset{(0.04)}{0.74}\left[\ln\left(\Delta\frac{Discretion_{it}}{GDP_{it}}\right)\right]$$

The results suggest that the coefficients on the two regressors add up to one and imply a tradeoff between bailouts and discretion. In Fig. 5, we plot the predicted iso-lines at given levels of the change in the ratio of debt-to-GDP. Alongside these iso-lines, we also plot the data for the 18 countries and 19 crises in our sample. The rise in the ratio of debt-to-GDP matches the data relatively well especially in the mid-range of the changes in the debt-to-GDP ratio (the R^2 of Eq. (1) is 0.97).

We also checked whether the tradeoff is apparent by interacting the fiscal costs variable with the size of the financial sector (the ratio of domestic private credit-to-GDP from IMF IFS). If the interaction term is positive then it implies that the countries with large financial sectors devote more of their fiscal space to bailouts. This is indeed what we find as seen in the following results:

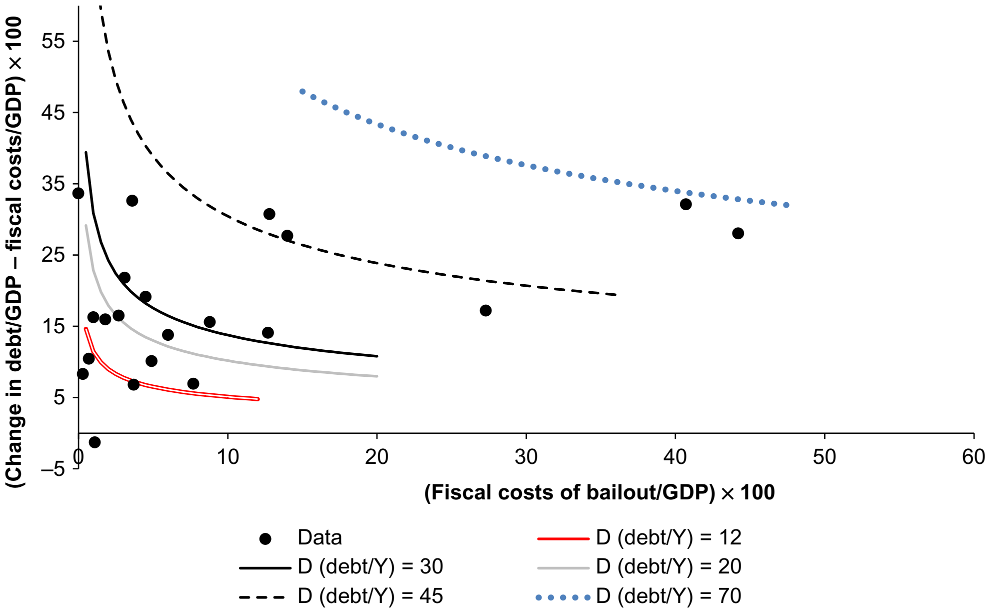


Fig. 5 Fiscal costs of bailouts vs the rise in government debt/GDP from other nonbailout costs, 19 crises, 1970–2012. Notes: *Data are from Laeven, L., Valencia, F., 2012. Systemic banking crises database: an update. IMF working paper no. 12/163. Iso-lines are the predicted values for the debt-to-GDP ratio from Eq. (1).*

$$\ln\left(\Delta\frac{Debt_{it}}{GDP_{it}}\right) = \underset{(0.49)}{1.72} + \underset{(0.24)}{-0.27}\left[\ln\left(\Delta\frac{Fiscal\ costs_{it}}{GDP_{it}}\right)\right] + \underset{(0.05)}{0.11}\Bigg[\ln\left(\Delta\frac{Fiscal\ costs_{it}}{GDP_{it}}\right)\times$$

$$\ln\left(\frac{Domestic\ credit_{it}}{GDP_{it}}\right)\Bigg] + \underset{(0.04)}{0.72}\left[\ln\left(\Delta\frac{Discretion_{it}}{GDP_{it}}\right)\right] - \underset{(0.10)}{0.22}\left[\ln\left(\frac{Domestic\ credit_{it}}{GDP_{it}}\right)\right]$$

Further investigation from a univariate regression reveals that the share of the rise in the ratio of debt-to-GDP accounted for by bailouts was a positive function of the size of the financial sector though this coefficient is not highly statistically significant. Results of the regression are seen in Fig. 6. Overall then, we find that as the size of fiscal bailouts increases, that what might be termed the discretionary component of the fiscal response is often smaller. A third factor generates a trilemma. Large financial sectors necessitate larger bailouts. If countries had small financial sectors, the constraints on discretionary fiscal action would be less binding.

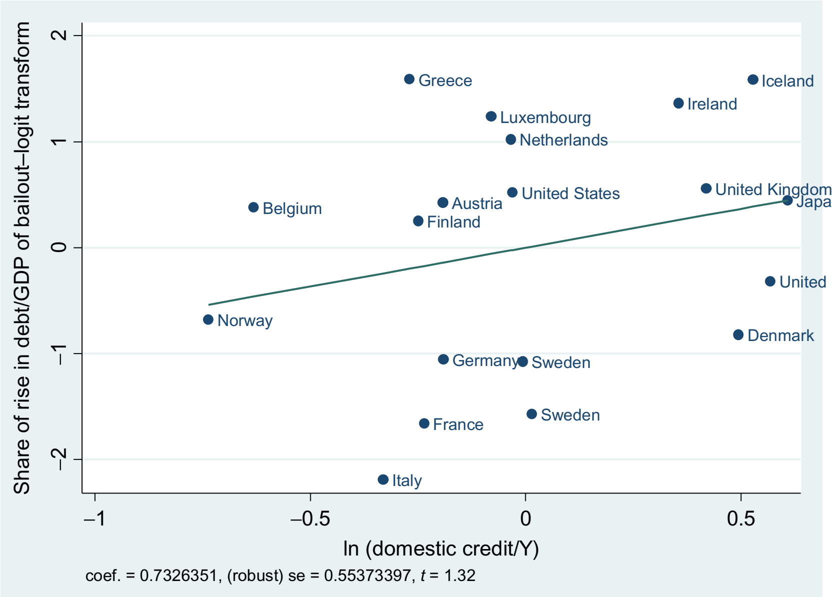


Fig. 6 Fiscal costs of a bailout as a share of the rise in debt-to-GDP vs size of the financial sector. *Notes*: Figure presents the predicted regression line/partial regression plot from a univariate regression of the share in the rise in debt as a percentage of GDP against the logarithm of the level of private domestic credit-to-GDP. We perform a logit transform on the dependent variable prior to estimation. *Debt data are from Laeven and Valencia and the credit data are from IMF IFS.*

6. CONCLUSIONS

This chapter surveyed the history, theory, and empirics of financial crises, fiscal crises and their interconnections. The history of the last two centuries shows clearly the presence of financial crises, currency crises, and debt crises somewhere in the world about every decade with five global systemic crises since the advent of globalization in the 19th century. The connection between financial crises and fiscal crises is primarily a more recent event, at least since the 1930s, although there were a number of such events in emerging market countries going back to the late 20th century. The key link between the two types of crises has been the increased use of government guarantees of financial institutions. These have surged in incidence and magnitude greatly since the Great Depression and especially since the 1980s. Governments after the Great Slump realized that banking panics were very costly events both in economic and political terms, and they have gone to great lengths to avoid the classic banking panics of the 19th and early 20th centuries and to avoid the perception of inaction. The consequence has been both more virulent modern banking crises with an increasingly strong likelihood of fiscal resolution and the accompanying fiscal resolution costs. This reflects the general phenomenon that when government intervenes to prevent costly events like forest fires and floods from occurring that economic agents adjust their behavior accordingly and use more of the protected resource than is in the long-run optimal (Ip, 2015). This has been the case with banking crises where the establishment of a safety net based on deposit insurance and other guarantees has led to regulatory forbearance and moral hazard and increased leverage by the protected financial institutions. Thus there is a tradeoff between the costs of the financial crises that accompany financial development and growth and the moral hazard costs of insurance. Under many plausible assumptions, eliminating financial crises entirely is not necessarily an achievable nor a desirable outcome (Tornell and Westermann, 2005). But neither is letting crises burn out on their own as was common in the early 19th century an ideal strategy. The optimal amount of financial crisis insurance in a world rife with market and regulatory imperfections is a subject for ongoing and future research.[bb]

The theoretical literature has evolved greatly since the mid-20th century in its treatment of different crisis phenomenon incorporating the tools of rational expectations, game theory, and dynamics. There was a burst of literature explaining banking panics in qualitative terms after the Great Depression (Friedman and Schwartz, 1963) among others. Then, after the opening up of global financial markets and the liberalization of financial markets from the post-Great Depression controls and repression, a wave of currency and banking crises swept the global economy. New innovations in theory including the Diamond–Dybvig model and first-generation speculative attack models were developed. The emerging market crises of the 1990s led to a spate of new theory

[bb] Allen and Gale (2007) discuss these issues from a theoretical perspective.

with emphasis on multiple equilibria and endogenous self-fulfilling crises. Since the 1990s, most macroeconomic models emphasize the interplay between real shocks and financial frictions with increasing sophistication. In addition, dynamic general equilibrium models are beginning to incorporate a banking sector with bank runs and liquidity.[cc] The recent subprime mortgage crisis followed by the Eurozone crisis has led to new literature focusing on the link between financial and fiscal crises linked together by government guarantees. Many of the ideas developed recently stemmed from work done after the Asian crisis of 1997. Judging from the explosion in theoretical modeling that followed the earlier waves of crises, more work will likely be done in the future on the fiscal–financial crisis nexus. Some questions that might be posed include:

- What do we know about optimal bank regulation, macroprudential policy, and the political economy of resolution? What do we know about the market failures that generate a need for such interventions?
- If it is hard to predict financial crises, can macroprudential policy and fiscal rules be reliable? Empirical research based on cross-country panel datasets has only just begun here (eg, Cerutti et al., forthcoming).
- What role does fiscal space play in the resolution phase of systemic financial crises?
- Is the way in which resolution proceeds dependent upon initial conditions and other institutional constraints?[dd]
- What kinds of fiscal union are feasible both economically and politically in a monetary union and how important are fiscal constraints under such arrangements? What fiscal arrangements are feasible and efficient in a monetary union facing systemic shocks?

Our survey of the empirical evidence on financial and fiscal crises led to our uncovering two very basic controversies: (1) Classification uncertainty: how do we define different types of financial crises and how do we date them? (2) What do we know about the costs and causes of crises? Our review of the literature and our own results based on a multi-country and multiyear database reveal that there are crucial differences over the definition of crises among the leading approaches taken in the literature. This has led to very different chronologies of the incidence of crises. This creates a serious problem for theorists and policy makers. Who should you believe? Picking the wrong approach can lead to misleading models calibrated to the wrong targets and ultimately to incorrect or misguided policy prescriptions.

If economists and policy makers truly believed that crises were an important phenomenon to understand and possibly avoid then it might be the case that an independent crisis dating committee could help set the standard in much the same way the NBER business

[cc] See, for example, Gertler and Kiyotaki (2015), Boissay et al. (forthcoming), and Paul (2016).

[dd] Steinkamp and Westermann (2014) show that the way in which resolution lending proceeds especially as regards junior or senior creditor status is associated with the country interest rate.

cycle dating committee works. The advantage of following this model is that the NBER is a respected nongovernmental, nonpartisan organization. Other organizations such as the IMF are not sufficiently politically independent. If crises are becoming increasingly global and crisis fighting is a global public good, then the importance of such a reform should be obvious. Such a committee could, if initiated, choose not only how to sensibly define crises in a uniform and consistent way, but also with an agreed definition this could help predict crises and to inform about the costs of crises.

With respect to measuring output losses, there are great differences in methodologies taken and techniques used. However, despite these differences, all of the studies agree that the output losses of financial crises are economically significant. This suggests that the stakes are high and the need for new theoretical and policy approaches to mitigate the crisis problem more compelling. The literature has some initial evidence that crises can be more severe when guarantees are not safeguarded or embedded in a reliable institutional framework. As of yet, we do not have a clear understanding of the magnitude of the impact of policies intended to mitigate crises (monetary policy, bailouts, fiscal policy). This surely must be a priority for research going forward. Any work in this direction must surely strive to meet the empirical standards set by the policy evaluation literature using credible research designs and/or sensible structural models of the phenomena in question.

Other empirical issues open up the door for further work. The question whether financial recessions lead to slow recoveries has not been resolved. Determining the leading causes of financial crises also is an open question. It is not at all obvious from the historical record that credit financed asset price boom–busts (ie, what has come to be known as the financial cycle) have always been, or will always be, the key explanation despite the recent emphasis on that explanation. Given the complexities of the financial ecosystem, perhaps some very general precepts should be at hand such as what is the level of risk and where are the risks residing?[ee] Overemphasis on one or a handful of indicators can be misleading if not dangerous for economic and financial stability. Due regard to the interconnections and systemic risks is required. Finally a question that needs more research is the connection between financial development, fiscal resolutions of crises, and overall fiscal policy goals.

Answering these questions is of the utmost importance for public policy toward financial crises. Understanding theoretically the causes and mechanics of crises and how they impinge on the real economy are crucial for the development of reasonable policies for crisis prevention, crisis management, and crisis resolution. But of course, getting the historical facts straight is also crucial. It is vital to avoid making rash generalizations which are based on overreading or misreading of economic history. Such analysis leads to pitfalls for theorizing based on stylized events that may be very far from reality and for policies

[ee] See Haldane and May (2011) on complexity and interconnections in the financial system.

designed to fight the next crisis based on a misunderstanding of what happened the last time.

The bottom line from our study is more work needs to be done on getting the historical facts correct in measuring the incidence and impact of financial crises and in understanding the true causes of crises and how they impact the real economy.

ACKNOWLEDGMENTS

We thank Sarah Quincy for invaluable research assistance. We thank conference and seminar participants at the BIS, Bonn, Bank of France, Copenhagen Business School, Hoover Institute, Kyoto World Economic History Conference, Lund, and Rutgers. Chris Cumming, Lars Jonung, Oscar Jordà, Yannick Kalantzis, John Landon-Lane, Ashoka Mody, Moritz Schularick, Anders Ögren, John B. Taylor, Harald Uhlig, and Eugene White provided very helpful comments. We remain responsible for any errors. Data and other related files are available at: https://sites.google.com/site/chrismmeissner/data.

REFERENCES

Acharya, V., Drechsler, I., Schnabl, P., 2014. A Pyrrhic victory? Bank bailouts and sovereign credit risk. J. Financ. 69 (6), 2689–2739.

Aguiar, M., Gopinath, G., 2006. Defaultable debt, interest rates and the current account. J. Int. Econ. 69 (1), 64–83.

Akerlof, G.A., Romer, P., 1993. Looting: the economic underworld of bankruptcy for profit. Brook. Pap. Econ. Act. 199 (2), 1–73.

Alessandri, P., Haldane, A.G., 2009. Banking on the state. In: Federal Reserve Bank of Chicago 12th Annual International Banking Conference, 2009.

Allen, F., Gale, D., 1998. Optimal financial crises. J. Financ. 53 (4), 1245–1284.

Allen, F., Gale, D., 2000. Bubbles and crises. Econ. J. 110 (460), 236–255.

Allen, F., Gale, D., 2004. Financial intermediaries and markets. Econometrica 72 (4), 1023–1061.

Allen, F., Gale, D., 2007. Understanding Financial Crises. Oxford University Press, New York, NY.

Angkinand, A.P., 2009. Banking regulation and the output cost of banking crises. J. Int. Financ. Mark. Inst. Money 19 (2), 240–257.

Arellano, C., 2008. Default risk and income fluctuations in emerging economies. Am. Econ. Rev. 98 (3), 690–712.

Arellano, C., Kocherlakota, N., 2014. Internal debt crises and sovereign defaults. J. Monet. Econ. 68 (Suppl.), S68–S80.

Arnold, B., Borio, C., Ellis, L., Moshirian, F., 2012. Systemic risk, macroprudential policy frameworks, monitoring financial systems and the evolution of capital adequacy. J. Bank. Financ. 36 (12), 3125–3132.

Babecký, J., Havránek, T., Matějů, J., Rusnák, M., Smídková, K., Vašíček, B., 2013. Leading indicators of crisis incidence: evidence from developed countries. J. Int. Money Financ. 35 (1), 1–19.

Babecký, J., Havránek, T., Matějů, J., Rusnák, M., Smídková, K., Vašíček, B., 2014. Banking, debt, and currency crises in developed countries: stylized facts and early warning indicators. J. Financ. Stab. 15, 1–17.

Bagehot, W., 1873. Lombard Street: A Decision of the Money Market. HS King, London.

Barro, R.J., Ursúa, J.F., 2008. Macroeconomic crises since 1870. Brook. Pap. Econ. Act. (Spring 2008), 255–335.

Basel Committee on Banking Supervision, Bank for International Settlements, 2010. An Assessment of the Long-Term Economic Impact of Stronger Capital and Liquidity Requirements. Mimeo BIS Basel, Switzerland.

Battistini, N., Pagano, M., Simonelli, S., 2014. Systemic risk, sovereign yields and bank exposures in the euro crisis. Econ. Policy 29 (78), 203–251.

Bernanke, B., 1983. Nonmonetary effects of the financial crisis in the propagation of the Great Depression. Am. Econ. Rev. 73 (1), 257–276.

Bernanke, B., Gertler, M., Gilchrist, S., 1999. The financial accelerator in a quantitative business cycle framework. In: Taylor, J., Woodford, M. (Eds.), In: Handbook of Macroeconomics, vol. 1, pp. 1341–1393.

Bhattacharya, S., Gale, D., 1987. Preference shocks, liquidity and central bank policy. In: Barnett, W., Singleton, K. (Eds.), New Approaches to Monetary Economics. Cambridge University Press, New York, NY, pp. 69–88.

Boissay, F., Collard, F. Smets, F., forthcoming. Booms and banking crises. J. Polit. Econ.

Bolton, P., Jeanne, O., 2011. Sovereign default risk and bank fragility in financially integrated economies. IMF Econ. Rev. 59, 162–194.

Bordo, M.D., 2006. Sudden stops, financial crises, and original sin in emerging countries: Déjà vu? NBER working paper 12393.

Bordo, M.D., Eichengreen, B., 1999. Is our international economic environment unusually crisis prone? In: Gruen, D., Gower, L. (Eds.), Capital Flows and the International Financial System. Reserve Bank of Australia, Sydney, pp. 18–74.

Bordo, M.D., Flandreau, M., 2003. Core, periphery, exchange rate regimes and globalization. In: Bordo, M. D., Taylor, A.M., Williamson, J.G. (Eds.), Globalization in Historical Perspective. University of Chicago Press for the NBER, Chicago, IL, pp. 417–472.

Bordo, M.D., Haubrich, J., 2012. Deep recessions, fast recoveries, and financial crises: evidence from the American record. NBER working paper 18194, June.

Bordo, M.D., Landon-Lane, J., 2010. The banking panics in the United States in the 1930s: some lessons for today. Oxf. Rev. Econ. Policy 26, 486–509.

Bordo, M.D., Landon-Lane, J., 2012. The global financial crisis: is it unprecedented? In: Obstfeld, M., Cho, D., Mason, A. (Eds.), Global Economic Crisis: Impacts, Transmission and Recovery. Edward Elgar, Northampton, MA, pp. 19–56.

Bordo, M.D., Landon-Lane, J., 2014. What explains house price booms? History and empirical evidence. In: Kouretras, G., Papadopoulos, A.P. (Eds.), Macroeconomic Analysis and International Finance. International symposia in Economic Theory and Econometrics. Emerald Publishers, Bingley, UK, pp. 1–36.

Bordo, M.D., Meissner, C.M., 2006. The role of foreign currency debt in financial crises: 1880–1913 vs 1972–1997. J. Bank. Financ. 30 (12), 3299–3329.

Bordo, M.D., Schwartz, A.J., 1999. Why currency clashes between internal and external stability goals end in currency crises, 1797–1995. Open Econ. Rev. 7 (1), 437–468.

Bordo, M.D., Eichengreen, B., Klingebiel, D., Martinez-Peria, S., 2001. Is the crisis problem growing more severe? Econ. Policy 16 (32), 52–82.

Bordo, M.D., Cavallo, A., Meissner, C.M., 2010. Sudden stops: determinants and output effects in the first era of globalization, 1880–1913. J. Dev. Econ. 91 (2), 227–241.

Bordo, M.D., Jonung, L., Markiewicz, A., 2013. A fiscal union for the euro: some lessons from history. CESifo Econ. Stud. 61 (3–4), 449–488.

Borio, C., 2012. The financial cycle and macroeconomics: what have we learnt? BIS working paper 395.

Borio, C., Drehman, M., 2009. Assessing the risk of banking crises: revisited. BIS Q. Rev. (March), 29–46.

Borio, C., James, H., Shin, H.S., 2014. The international monetary and financial system: a capital account historical perspective. Federal Reserve Bank of Dallas Globalization and Monetary Policy Institute Working Paper No. 204.

Brunnermeir, M., Oehmke, M., 2013. Bubbles, financial crises, and systemic risk. In: Constantinides, M.H., Stulz, R.M. (Eds.), In: Handbook of the Economics of Finance, vol. 2. North Holland Elsevier, Oxford, pp. 1221–1288.

Bulow, J., Rogoff, K.S., 1989a. A constant recontracting model of sovereign debt. J. Polit. Econ. 97 (1), 155–178.

Bulow, J., Rogoff, K.S., 1989b. Sovereign debt: is to forgive to forget? Am. Econ. Rev. 79 (1), 43–52.

Bulow, J., Rogoff, K.S., 2015. Why sovereigns repay debts to external creditors and why it matters. vox EU, 10 June 2015. http://www.voxeu.org/article/why-sovereigns-repay-debts-external-creditors-and-why-it-matters.

Burnside, C., 2004. Currency crises and contingent liabilities. J. Int. Econ. 62 (1), 25–52.

Burnside, C., Eichenbaum, M., Rebelo, S., 2001. Prospective deficits and the Asian currency crisis. J. Polit. Econ. 109 (6), 1155–1197.

Burnside, C., Eichenbaum, M., Rebelo, S., 2004. Government guarantees and self-fulfilling speculative attacks. J. Econ. Theory 119 (1), 31–63.

Bussiere, M., Fratzscher, M., 2006. Towards a new early warning system of financial crises. J. Int. Money Financ. 25 (6), 953–973.

Caballero, J.A., 2014. Do surges in international capital inflows influence the likelihood of banking crises? Econ. J., 1–36.

Calomiris, C., Beim, D., 2001. Emerging Financial Markets. Irwin Professional Publishers, New York, NY.

Calomiris, C.W., Hubbard, R.G., 1989. Price flexibility, credit availability, and economic fluctuations: evidence from the United States, 1894–1909. Q. J. Econ. 104 (3), 429–452.

Calomiris, C., Kahn, C., 1991. The role of demandable debt in structuring optimal banking arrangements. Am. Econ. Rev. 93 (5), 1615–1646.

Calomiris, C., Mason, J., 2003. Fundamentals, panic and bank distress during the depression. Am. Econ. Rev. 93 (5), 1615–1647.

Caprio Jr., G., Klingebiel, D., 1996. Bank insolvencies: cross-country experience. Policy Research working paper 1620. The World Bank, Washington, DC.

Caprio, C., Klingebiel, D., 2003. Episodes of Systemic and Borderline Financial Crises. The World Bank, Washington, DC.

Carlson, M., Mitchener, K.J., Richardson, G., 2011. Arresting banking panics: federal reserve liquidity provision and the forgotten panic of 1929. J. Polit. Econ. 119 (5), 889–924.

Cecchetti, S.G., Kohler, M., Upper, C., 2009. Financial crises and economic activity. NBER working paper 15379.

Cerutti, E, Claessens, S., Laeven, L., forthcoming. The use and effectiveness of macroprudential policies: new evidence. J. Financ. Stab.

Chari, V.V., Jagannathan, R., 1989. Banking panics, information and rational expectations equilibrium. J. Financ. 43 (3), 749–761.

Cole, H., Kehoe, P., 1995. The role of institutions in reputation models of sovereign debt. J. Monet. Econ. 35 (1), 45–46.

Cole, H., Kehoe, P., 1998. Models of sovereign debt: partial versus general reputation. Int. Econ. Rev. 29 (1), 55–70.

Conant, C., 1915. A History of Modern Banks of Issue, fifth ed. G.P. Putnam's and Sons, New York, NY.

Corsetti, G., Pesenti, P., Roubini, N., 1999. Paper tigers? A model of the Asian crisis. Eur. Econ. Rev. 43 (7), 1211–1236.

da Rocha, B.T., Solomou, S., 2015. The effects of systemic banking crises in the inter-war period. J. Int. Money Financ. 54, 35–49.

Dell'Ariccia, G., Detragiache, E., Rajan, R., 2008. The real effect of banking crises. J. Financ. Intermed. 17 (1), 89–112.

Demirgüç-Kunt, A., Detragiache, E., 1998. The determinants of banking crises: evidence from developing and developed countries. IMF Staff Pap. 45, 81–109.

Diamond, D., Dybvig, P., 1983. Bank runs, deposit insurance, and liquidity. J. Polit. Econ. 91 (3), 401–419.

Diamond, D., Rajan, R., 2001. Liquidity risk, liquidity creation and financial fragility: a theory of banking. J. Polit. Econ. 109 (2), 2431–2465.

Diamond, D., Rajan, R., 2005. Liquidity shortages and banking crisis. J. Financ. 60 (2), 615–647.

Diamond, D., Rajan, R., 2011. Fear of fire sales, illiquidity seeking and credit squeezes. Q. J. Econ. 126 (2), 557–591.

Diamond, D., Rajan, R., 2012. Illiquid banks, financial stability and interest rate policy. J. Polit. Econ. 120 (3), 552–591.

Diaz-Alejandro, C., 1985. Good bye financial repression, hello financial crash. J. Dev. Econ. 19 (1–2), 1–24.

Dooley, M., 2000. A model of crises in emerging markets. Econ. J. 110 (460), 256–272.
Drees, B., Pazarbasioglu, C., 1994. The Nordic banking crisis: pitfalls in financial liberalization. IMF occasional paper. International Monetary Fund, Washington, DC.
Drehman, M., Borio, C., Tsatsaronis, K., 2012. Characterising the financial cycle: don't lose sign of the medium term! BIS working paper no. 380.
Eaton, J., 1996. Sovereign debt, reputation and credit terms. Int. J. Financ. Econ. 1 (1), 25–35.
Eaton, J., Gersovitz, M., 1981. Debt with potential repudiation: theoretical and empirical analysis. Rev. Econ. Stud. 48 (2), 289–309.
Edwards, S., Santaella, J., 1993. Devaluation controversies in the developing countries: lessons from the Bretton woods era. In: Bordo, M.D., Eichnegreen, B. (Eds.), A Retrospective on the Bretton Woods System: Lessons for International Monetary Reform. University of Chicago Press, Chicago, IL, pp. 405–460.
Eichengreen, B., 1992. Golden Fetters. Oxford University Press, Oxford.
Eichengreen, B., Hausmann, R., 2005. Other People's Money: Debt Denomination and Financial Instability in Emerging Market Economies. University of Chicago Press, Chicago, IL.
Eichengreen, B., Rose, A.K., Wyplosz, C., 1995. Exchange market mayhem: the antecedents and aftermath of speculative attacks. Econ. Policy 10 (21), 249–312.
European Commission, 2009. Public Finances in EMU 2009. European Commission, Luxembourg.
Fisher, I., 1932. Booms and Depressions. Adelphi, New York, NY.
Fisher, I., 1933. The debt deflation theory of Great Depressions. Econometrica 1 (4), 337–357.
Frankel, J.A., Rose, A.K., 1996. Currency crashes in emerging markets: an empirical treatment. J. Int. Econ. 41 (3), 351–366.
Frankel, J.A., Saravelos, G., 2012. Can leading indicators assess country vulnerability? Evidence from the 2008–09 global financial crisis. J. Int. Econ. 87 (2), 216–231.
Fratzscher, M., Rieth, M., 2015. Monetary policy, bank bailouts and the Sovereign-bank risk Nexus in the Euro Area. CEPR working paper no. 10370.
Friedman, M., 1993. The 'plucking model' of business fluctuations revisited. Econ. Inq. 31, 171–177.
Friedman, M., Schwartz, A.J., 1963. A Monetary History of the United States 1867 to 1960. Princeton University Press, Princeton, NJ.
Funabashi, Y., 1988. Managing the Dollar: From the Plaza to the Louvre. Institute for International Economics, Washington, DC.
Gennaioli, N., Martin, A., Rossi, S., 2014. Sovereign default, domestic banks, and financial institutions. J. Financ. 69 (2), 819–886.
Gertler, M., Karadi, P., 2011. A model of unconventional monetary policy. J. Monet. Econ. 58 (1), 17–34.
Gertler, M., Kiyotaki, N., 2015. Banking liquidity and bank runs in an infinite horizon economy. Am. Econ. Rev. 105 (7), 2011–2043.
Goetzmann, W., 2015. Bubble investing: learning from history. NBER working paper 21693.
Goldstein, I., Pauzner, A., 2005. Demand-deposit contracts and the probability of bank runs. J. Financ. 60 (3), 1293–1327.
Gorton, G., 1988. Banking panics and business cycles. Oxf. Econ. Pap. 40 (4), 751–781.
Gorton, G., Huang, I., 2004. Liquidity, efficiency and bank bailouts. Am. Econ. Rev. 94 (3), 455–483.
Gorton, G., Ordoñez, G., 2016. Good booms, bad booms. NBER working paper 22008.
Gourinchas, P.O., Obstfeld, M., 2012. Stories of the twentieth century for the twenty-first. Am. Econ. J. Macroecon. 4 (1), 226–265.
Grossman, R.S., 1994. The shoe that didn't drop: explaining banking stability during the Great Depression. J. Econ. Hist. 54 (3), 654–682.
Grossman, R.S., 2010. Unsettled Account: The Evolution of Banking in the Industrialized World Since 1820. Princeton University Press, Princeton, NJ.
Grossman, H., van Huyck, J., 1988. Sovereign debt as a contingent claim: excusable default, repudiation and reputations. Am. Econ. Rev. 78 (5), 1088–1097.
Haldane, A., May, R.M., 2011. Systemic risk in banking ecosystems. Nature 469, 351–355.
Hoggarth, G., Ricardo, R., Saporta, V., 2002. Costs of banking system instability: some empirical evidence. J. Bank. Financ. 26 (5), 825–855.

Holmström, B., Tirole, J., 1998. Private and public supply of liquidity. J. Polit. Econ. 106 (1), 1–40.

Honkapohja, S., 2009. The 1990s financial crisis in Nordic countries. Bank of Finland discussion paper.

Howard, G., Martin, R., Wilson, B., 2011. Are recoveries from banking and financial crises really so different? International Finance discussion papers. Federal Reserve Board of Governors, Washington, DC.

Hutchison, M.M., Noy, I., 2005. How bad are twins? Output costs of currency and banking crises. J. Money Credit Bank. 725–752.

International Monetary Fund, IMF, 2009. Lessons of the Global Crisis for Macroeconomic Policy. Mimeo, International Monetary Fund Research Department, Washington, D.C.

Ip, G., 2015. Foolproof: Why Safety Can Be Dangerous and How Danger Makes Us Safe Little. Brown and Company, New York, NY.

Iwaisako, T., Ito, T., 1995. Explaining asset bubbles in Japan. NBER working paper 5350.

Jacklin, C., 1987. Demand deposits, trading restrictions, and risk sharing. In: Prescott, E., Wallace, N. (Eds.), Contractual Arrangements for Intertemporal Trade. University of Minnesota Press, Minneapolis, MN, pp. 26–47.

Jacklin, C., Bhattacharya, S., 1988. Distinguishing panics and information based bank runs: welfare and policy implications. J. Polit. Econ. 96 (3), 568–592.

Jalil, A., 2015. A new history of banking panics in the United States, 1825–1929: construction and implications. Am. Econ. J. Macroecon. 7 (3), 295–330.

Jonung, L., Kiander, J., Vartia, P., 2009. The great financial crisis in Finland and Sweden: the dynamics of boom, bust and recovery 1985–2000. In: Jonung, L., Kiander, J., Vartia, P. (Eds.), The Great Financial Crisis in Finland and Sweden: The Nordic Experience of Financial Liberalization. Edward Elgar Publishers, Cheltenham, UK, pp. 19–70.

Jordà, Ò., Schularick, M., Taylor, A.M., 2011. Financial crises, credit booms, and external imbalances: 140 years of lessons. IMF Econ. Rev. 59 (2), 340–378.

Jordà, O., Schularick, M., Taylor, A.M., 2013. When credit bites back. J. Money Credit Bank. 45 (2), 3–28.

Jordà, O., Schularick, M., Taylor, A.M., forthcoming. The great mortgaging: housing finance, crises, and business cycles. Econ. Policy. 31 (85).

Kaminsky, G., 1999. Currency and banking crises: the early warnings of distress. IMF working paper no. 99/178.

Kaminsky, G., Reinhart, C.M., 1999. The twin crises: the causes of banking and balance-of-payments problems. Am. Econ. Rev. 89 (3), 473–500.

Kaminsky, G.L., Vega-García, P., 2016. Systemic and idiosyncratic sovereign debt crises. J. Eur. Econ. Assoc. 14 (1), 80–114.

Kaminsky, G., Lizondo, S., Reinhart, C.M., 1998. Leading indicators of currency crises. Staff Pap. Int. Monet. Fund 45 (1), 1–48.

Kaufman, H., 1986. Debt: The Threat to Economic and Financial Stability. In: Debt, Financial Stability and Public Policy. Federal Reserve Bank of Kansas City, Kansas City, MO.

Kindleberger, C., 1978. Manias, Panics and Crashes: A History of Financial Crises. Wiley and Sons, New York, NY.

Kletzer, K., Wright, B., 2000. Sovereign debt as intertemporal barter. Am. Econ. Rev. 90 (3), 621–639.

Krugman, P., 1979. A model of balance of payments crises. J. Money Credit Bank. 11 (3), 311–325.

Krugman, P., 1998. Currency Crises. Mimeo, Princeton University, Princeton, NJ.

Krugman, P., 1999. Balance sheets, the transfer problem and financial crises. In: Isard, P., Razin, A., Rose, A.K. (Eds.), International Finance and Financial Crises: Essays in Honor of Robert B. Flood. Springer, New York, NY, pp. 31–55.

Laeven, L., Valencia, F., 2008. Systemic banking crises: a new database. IMF working paper no. 08/224.

Laeven, L., Valencia, F., 2012. Systemic banking crises database: an update. IMF working paper no. 12/163.

Marichal, C., 1989. A century of debt crisis in Latin America: from independence to the Great Depression, 1820–1930. Princeton University Press, Princeton, NJ.

Martin, P., Philippon, T., 2015. Inspecting the mechanism: leverage and the great recession in the Eurozone. NBER working paper 20572.

McKinnon, R., Pill, H., 1986. Credible liberalizations and international capital flows: the over borrowing syndrome. In: Ito, T., Kreuger, A. (Eds.), Financial Deregulation and Integration in East Asia. University of Chicago Press, Chicago, IL, pp. 7–50.

Minsky, H., 1977. A theory of systemic fragility. In: Altman, E.J., Sametz, A. (Eds.), Financial Crises: Institutions and Markets in a Fragile Environment. Wiley, New York, NY, pp. 138–152.

Mitchell, W.C., 1941. Business Cycles and Their Causes. University of California Press, Berkeley, CA.

Mitchener, K.J., Richardson, G., 2014. Shadowy Banks and the Interbank Amplifier During the Great Depression. Mimeo, UC Irvine, Irvine, CA.

Mitchener, K.J., Wiedenmeir, M., 2010. Supersanctions and sovereign debt repayment. J. Int. Money Financ. 29 (1), 19–36.

Mladjan, M., 2012. Accelerating into the Abyss: Financial Dependence and the Great Depression. Mimeo. EBS Business School, Wiesbaden, Germany.

Mody, A., Sandri, D., 2012. The Eurozone crisis: how banks and sovereigns came to be joined at the hip. Econ. Policy 27 (70), 201–230.

Morris, S., Shin, H.S., 1998. Unique equilibrium in a model of self-fulfilling currency attack. Am. Econ. Rev. 88 (3), 587–597.

Obstfeld, M., 1995. The logic of currency crises. In: Eichengreen, B., Frieden, J., von Hagen, J. (Eds.), Monetary and Fiscal Policy in an Integrated Europe. Springer, Heidelberg, pp. 62–90.

Panizza, U., Sturzenegger, F., Zettelmeyer, J., 2009. The economics and law of sovereign debt and default. J. Econ. Lit. 47 (3), 651–669.

Paul, P., 2016. Financial Crises and Debt Rigidities. Mimeo, University of Oxford, Oxford, UK.

Portes, R., 2010. Comments on Claessens, S., Dell'Ariccia, G., Igan, D., and Laeven, L. Econ. Policy 25 (62), 267–293.

Reinhart, C.M., 2010. This time is different chartbook: country histories on debt, default and financial crises. NBER working paper 15815.

Reinhart, C.M., 2015. The antecedents and aftermath of financial crises as told by Carlos F. Diaz-Alejandro. NBER working paper 21350.

Reinhart, C.M., Rogoff, K.S., 2009. This Time is Different: Eight Centuries of Financial Folly. Princeton University Press, Princeton, NJ.

Reinhart, C.M., Rogoff, K.S., 2011. From financial crash to debt crisis. Am. Econ. Rev. 101 (5), 1676–1706.

Reinhart, C.M., Rogoff, K.S., 2014. Recovery from financial crises: evidence from 100 episodes. Am. Econ. Rev. 104 (5), 50–55.

Reinhart, C., Rogoff, K.S., Savastano, M., 2003. Debt intolerance. Brook. Pap. Econ. Act. 1, 1–62.

Richardson, G., 2007. Categories and causes of bank distress during the Great Depression, 1920–1935: the liquidity and insolvency debate revisited. Explor. Econ. Hist. 44 (4), 588–607.

Richardson, G., Troost, W., 2009. Monetary intervention mitigated banking panics during the Great Depression: quasi-experimental evidence from a federal reserve district border, 1929–1933. J. Polit. Econ. 117 (6), 1031–1073.

Rochet, J., Vives, X., 2004. Coordination failures and the lender of last resort; was Bagehot right after all? J. Eur. Econ. Assoc. 2 (6), 1116–1147.

Rockoff, H., 2014. It is Always the Shadow Banks: The Failures that Ignited America's Financial Panics. Mimeo, Rutgers University, New Brunswick, NJ.

Romer, C., Romer, D., 2015. New evidence on the impact of financial crises in advanced countries. NBER working paper 21021.

Rose, A.K., Spiegel, M.M., 2011. Cross-country causes and consequences of the 2008 crisis: an update. Eur. Econ. Rev. 55 (3), 309–324.

Rose, A.K., Spiegel, M.M., 2012. Cross-country causes and consequences of the 2008 crisis: early warning. Jpn World Econ. 24, 1–16.

Sayek, S., Taksin, F., 2014. Financial crises: lessons from history for today. Econ. Policy 29 (79), 447–493.

Schneider, M., Tornell, A., 2004. Balance sheet effects, bailout guarantees and financial crises. Rev. Econ. Stud. 74, 883–913.

Schularick, M., 2012. Public debt and financial crises in the twentieth century. Eur. Rev. Hist. 19 (6), 881–897.

Schularick, M., Taylor, A.M., 2012. Credit booms gone bust: monetary policy, leverage cycles, and financial crises, 1870–2008. Am. Econ. Rev. 102 (2), 1029–1061.

Schwartz, A.J., 1987. The lender of last resort and the federal safety net. J. Financ. Serv. Res. 1, 77–111.

Steigum, E., 2009. The boom and bust cycle in Norway. In: Jonung, L., Kiander, J., Vartia, P. (Eds.), The Great Financial Crisis in Finland and Sweden: The Nordic Experience of Financial Liberalization. Edward Elgar Publishers, Cheltenham, UK, pp. 202–244.

Steinkamp, S., Westermann, F., 2014. The role of creditor seniority in Europe's sovereign debt crisis. Econ. Policy 29 (79), 495–552.

Stuckler, D., Meissner, C.M., Fishback, P., Basu, S., McKee, M., 2012. Banking crises and mortality during the Great Depression: evidence from US urban populations, 1929–1937. J. Epidemiol. Community Health 66 (5), 410–419.

Sturzenegger, F., Zettelmeyer, J., 2006. Debt Defaults and Lessons from a Decade of Crises. MIT Press, Cambridge, MA.

Tagkalakis, A., 2013. The effects of financial crisis on fiscal positions. Eur. J. Polit. Econ. 29, 197–213.

Taylor, A.M., 2015. Credit, stability and the macroeconomy. Annu. Rev. Econ. 7 (1), 309–339. http://dx.doi.org/10.1146/annurev-economics-080614-115437.

Temin, P., 1976. Did Monetary Forces Cause the Great Depression? WW Norton, New York, NY.

Tornell, A., Westermann, F., 2005. Boom Bust Cycles and Financial Liberalization. MIT Press, Cambridge, MA.

Uhlig, H., 2013. Sovereign default risk and banks in a monetary union. Ger. Econ. Rev. 15 (1), 23–41.

Velasco, A., 1987. Financial crises and balance of payments crises: a simple model of the southern cone experience. J. Dev. Econ. 27 (1–2), 263–283.

Wallace, N., 1988. Another attempt to explain an illiquid banking system: the Diamond Dybvig model with sequential service taken seriously. Q. Rev. FRB Minneapolis 12 (4), 3–16.

White, E.N., 2000. Banking and finance in the twentieth century. In: Gallman, R., Engerman, S. (Eds.), Cambridge Economic History of the United States. Cambridge University Press, New York, NY, pp. 742–802.

White, E.N., 2015. Rescuing a SIFI, halting a panic: the Barings crisis of 1890. In: Paper Presented at the Banque de France. December, 2015.

World Bank, 2002. Global Development Finance. In: Appendix on Commercial Debt Restructuring. World Bank, Washington, D.C.

Wray, L.W., 2015. Why Minsky Matters: An Introduction to the Work of a Maverick Economist. Princeton University Press, Princeton, NJ.

Zarnowitz, V., 1992. Business Cycles: Theory, History Indicators, and Forecasting. University of Chicago Press, Chicago, IL.

Ziebarth, N., 2013. Identifying the effects of bank failures from a natural experiment in Mississippi during the Great Depression. Am. Econ. J. Macroecon. 5 (1), 81–101.

SECTION 2

The Methodology of Macroeconomics

CHAPTER 8

Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics[☆]

J.H. Stock[*,‡], **M.W. Watson**[†,‡]
[*]Harvard University, Cambridge, MA, United States
[†]The Woodrow Wilson School, Princeton University, Princeton, NJ, United States
[‡]The National Bureau of Economic Research, Cambridge, MA, United States

Contents

[☆] Replication files and the Supplement are available on Watson's Website, which also includes links to a suite of software for estimation and inference in DFMs and structural DFMs built around the methods described in this chapter.

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.04.002

Abstract

This chapter provides an overview of and user's guide to dynamic factor models (DFMs), their estimation, and their uses in empirical macroeconomics. It also surveys recent developments in methods for identifying and estimating SVARs, an area that has seen important developments over the past 15 years. The chapter begins by introducing DFMs and the associated statistical tools, both parametric (state-space forms) and nonparametric (principal components and related methods). After reviewing two mature applications of DFMs, forecasting and macroeconomic monitoring, the chapter lays out the use of DFMs for analysis of structural shocks, a special case of which is factor-augmented vector autoregressions (FAVARs). A main focus of the chapter is how to extend methods for identifying shocks in structural vector autoregression (SVAR) to structural DFMs. The chapter provides a unification of SVARs, FAVARs, and structural DFMs and shows both in theory and through an empirical application to oil shocks how the same identification strategies can be applied to each type of model.

Keywords

State-space models, Structural vector autoregressions, Factor-augmented vector autoregressions, Principal components, Large-model forecasting, Nowcasting, Structural shocks

JEL Classification Codes

C32, C38, C55, E17, E37, E47

1. INTRODUCTION

The premise of dynamic factor models (DFMs) is that the common dynamics of a large number of time series variables stem from a relatively small number of unobserved (or latent) factors, which in turn evolve over time. Given the extraordinary complexity and regional and sectoral variation of large modern economies, it would seem surprising a priori that such a simple idea would have much empirical support. Remarkably, it does.

Fig. 1 shows a key result for a single-factor DFM fit to 58 quarterly US real activity variables (sectoral industrial production (IP), sectoral employment, sales, and National Income and Product Account (NIPA) series); the details are discussed in Section 6. A single common factor for these series was estimated using principal components analysis, a least-squares method for estimating the unobserved factors nonparametrically discussed in Section 2. The figure shows the detrended[a] four-quarter growth rates of four measures of aggregate economic activity (real Gross Domestic Product (GDP), total nonfarm employment, IP, and manufacturing and trade sales), along with the fitted value from a regression of the quarterly growth rate of each series on the single common factor. None of the four series plotted in Fig. 1 were used to estimate the factor: although disaggregated NIPA variables like consumption of durables, of nondurables, and of services were used, total consumption, GDP, and other high-level aggregates were not. As can be seen in the figure, the single factor explains a large fraction of the four-quarter variation in these four series. For these four series, the R^2s of the four-quarter fits range from 0.73 for GDP to 0.92 for employment. At the same time, the estimated factor does not equal any one of these series, nor does it equal any one of the 58 series used to construct it.

DFMs have several appealing properties that drive the large body of research on methods and applications of DFMs in macroeconomics. First, as Fig. 1 suggests and as is discussed in more detail later, empirical evidence supports their main premise: DFMs fit the data. The idea that a single index describes the comovements of many macroeconomics variables arguably dates at least to Burns and Mitchell (1946), and additional early references are discussed in Section 2.

Second, as is discussed in the next section, the key DFM restriction of a small number of latent factors is consistent with standard dynamic equilibrium macroeconomic theories.

Third, techniques developed in the past 15 years have allowed DFMs to be estimated using large datasets, with no practical or computational limits on the number of variables. Large datasets are now readily available,[b] and the empirical application in this chapter uses a 207-variable DFM. Estimation of the factors, DFM parameters, and structural DFM impulse response functions (IRFs) takes only a few seconds. Forecasts based on large

[a] Following Stock and Watson (2012a) and as discussed in Section 6.1, the trends in the growth rates were estimated using a biweight filter with a bandwidth of 100 quarters; the displayed series subtract off these trends.

[b] For example, McCracken and Ng (2015) have compiled an easily downloaded large monthly macroeconomic dataset for the United States (FRED-MD), which is available through the Federal Reserve Bank of St. Louis FRED data tool at https://research.stlouisfed.org/econ/mccracken/fred-databases/.

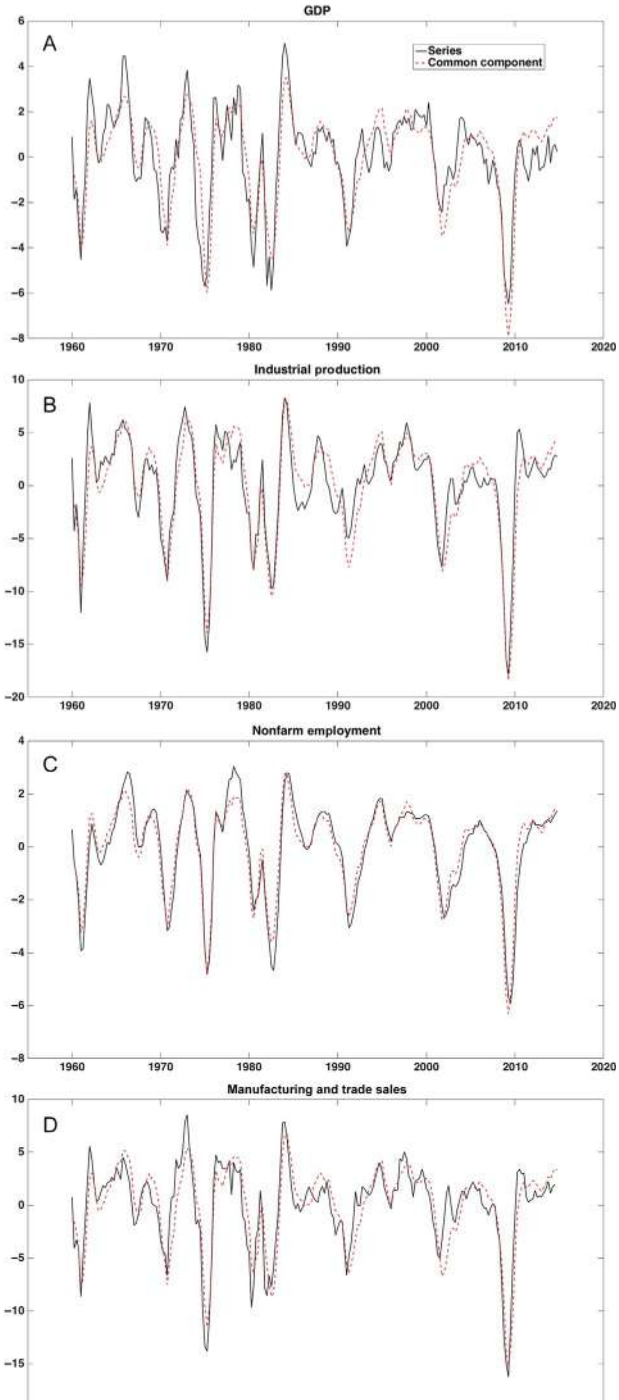


Fig. 1 Detrended four-quarter growth rates of US GDP, industrial production, nonfarm employment, and manufacturing and trade sales (*solid line*), and the common component (*fitted value*) from a single-factor DFM (*dashed line*). The factor is estimated using 58 US quarterly real activity variables. Variables all measured in percentage points.

DFMs have rich information sets but still involve a manageably small number of predictors, which are the estimates of the latent factors, and do so without imposing restrictions such as sparsity in the original variables that are used by some machine learning algorithms. As a result, DFMs have been the main "big data" tool used over the past 15 years by empirical macroeconomists.

Fourth, DFMs are well suited to practical tasks of professional macroeconomists such as real-time monitoring, including construction of indices from conceptually similar noisy time series.

Fifth, because of their ability to handle large numbers of time series, high-dimensional DFMs can accommodate enough variables to span a wide array of macroeconomic shocks. Given a strategy to identify one or more structural shocks, a structural DFM can be used to estimate responses to these structural shocks. The use of many variables to span the space of the shocks mitigates the "invertibility problem" of structural vector autoregressions (SVARs), in which a relatively small number of variables measured with error might not be able to measure the structural shock of interest.

The chapter begins in Section 2 with an introduction to structural dynamic factor models (SDFMs) and methods for estimating DFMs, both parametric (state-space methods) and nonparametric (principal components and related least-squares methods). This discussion includes extensions to data irregularities, such as missing observations and mixed observation frequencies, and covers recent work on detecting breaks and other forms of instability in DFMs.

The chapter then turns to a review of the main applications of DFMs. The first, macroeconomic monitoring and forecasting, is covered in Section 3. These applications are mature and many aspects have been surveyed elsewhere, so the discussion is relatively brief and references to other surveys are provided.

Sections 4 and 5 examine estimation of the effects of structural shocks. One of the main themes of this chapter is that the underlying identification approaches of SVARs carry over to structural DFMs. This is accomplished through two normalizations, which we call the unit effect normalization for SVARs and the named factor normalization for DFMs. These normalizations set the stage for a unified treatment, provided in these sections, of structural DFMs, factor-augmented VARs (FAVARs), and SVARs.

The basic approaches to identification of structural shocks are the same in SVARs, FAVARs, and SDFMs. Section 4 therefore surveys the identification of structural shocks in SVARs. This area has seen much novel work over the past 10 years. Section 4 is a stand-alone survey of SVAR identification that can be read without reference to other sections of this chapter and complements Ramey (2016). Section 4 discusses another of the main themes of this chapter: as modern methods for identification of structural shocks in SVARs become more credible, they raise the risk of relying on relatively small variations in the data, which in turn means that they can be weakly identified. As in applications with microdata, weak identification can distort statistical inference using both Bayes and frequentist methods. Section 4 shows how weak identification can arise in various SVAR identification strategies.

Section 5 shows how these SVAR identification schemes extend straightforwardly to SDFMs and FAVARs. Section 5 also develops another main theme of this chapter that structural DFMs, FAVARs, and SVARs are a unified suite of tools with fundamentally similar structures that differ in whether the factors are treated as observed or unobserved. By using a large number of variables and treating the factors as unobserved, DFMs "average out" the measurement error in individual time series, and thereby improve the ability to span the common macroeconomic structural shocks.

Sections 6 and 7 turn to an empirical illustration using an eight-factor, 207-variable DFM. Section 6 works through the estimation of the DFM, first using only the real activity variables to construct a real activity index, then using all the variables.

Section 7 uses the 207-variable DFM to examine the effect of oil market shocks on the US economy. The traditional view is that unexpected large increases in oil prices have large and negative effects on the US economy and have preceded many postwar US recessions (Hamilton, 1983, 2009). Subsequent work suggests, however, that since the 1980s oil shocks have had a smaller impact (eg, Hooker, 1996; Edelstein and Kilian, 2009; Blanchard and Galí, 2010), and moreover that much of the movement in oil prices is due to demand shocks, not oil supply shocks (eg, Kilian, 2009). We use a single large DFM to illustrate how SVAR identification methods carry over to structural DFMs and to FAVARs, and we compare structural DFM, FAVAR, and SVAR results obtained using two different methods to identify oil market shocks. The structural DFM results are consistent with the main finding in the modern literature that oil supply shocks explain only a fraction of the variation in oil prices and explain a very small fraction of the variation in major US macroeconomic variables since the mid-1980s.

In Section 8, we step back and assess what has been learned, at a high level, from the large body of work on DFMs in macroeconomics. These lessons include some practical recommendations for estimation and use of DFMs, along with some potential pitfalls.

There are several recent surveys on aspects of DFM analysis which complement this chapter. Bai and Ng (2008) provide a technical survey of the econometric theory for principal components and related DFM methods. Stock and Watson (2011) provide an overview of the econometric methods with a focus on applications. Bańbura et al. (2013) survey the use of DFMs for nowcasting. The focus of this chapter is DFMs in macroeconomics and we note, but do not go into, the vast applications of factor models and principal components methods in fields ranging from psychometrics to finance to big data applications in the natural and biological sciences and engineering.

2. DFMs: NOTATION AND SUMMARY OF ECONOMETRIC METHODS

2.1 The DFM

The DFM represents the evolution of a vector of N observed time series, X_t, in terms of a reduced number of unobserved common factors which evolve over time, plus uncorrelated disturbances which represent measurement error and/or idiosyncratic dynamics of the individual series. There are two ways to write the model. The dynamic form represents the dependence of X_t on lags (and possibly leads) of the factors explicitly, while the static form represents those dynamics implicitly. The two forms lead to different estimation methods. Which form is more convenient depends on the application.

The DFM is an example of the much larger class of state-space or hidden Markov models, in which observable variables are expressed in terms of unobserved or latent variables, which in turn evolve according to some lagged dynamics with finite dependence (ie, the law of motion of the latent variables is Markov). What makes the DFM stand out for macroeconometric applications is that the complex comovements of a potentially large number of observable series are summarized by a small number of common factors, which drive the common fluctuations of all the series.

Unless stated explicitly otherwise, observable and latent variables are assumed to be second-order stationary and integrated of order zero; treatment of unit roots, low-frequency trends, and cointegration are discussed in Section 2.1.4. In addition, following convention all data series are assumed to be transformed to have unit standard deviation.

Throughout this chapter, we use lag operator notation, so that $a(\mathrm{L}) = \sum_{i=0}^{\infty} a_i \mathrm{L}^i$, where L is the lag operator, and $a(\mathrm{L})X_t = \sum_{i=0}^{\infty} a_i X_{t-i}$.

2.1.1 Dynamic Form of the DFM

The DFM expresses a $N \times 1$ vector X_t of observed time series variables as depending on a reduced number q of unobserved or latent factors f_t and a mean-zero idiosyncratic component e_t, where both the latent factors and idiosyncratic terms are in general serially correlated. The DFM is,

$$X_t = \lambda(\mathrm{L}) f_t + e_t \tag{1}$$

$$f_t = \Psi(\mathrm{L}) f_{t-1} + \eta_t \tag{2}$$

where the lag polynomial matrices $\lambda(\mathrm{L})$ and $\Psi(\mathrm{L})$ are $N \times q$ and $q \times q$, respectively, and η_t is the $q \times 1$ vector of (serially uncorrelated) mean-zero innovations to the factors. The idiosyncratic disturbances are assumed to be uncorrelated with the factor innovations at all leads and lags, that is, $Ee_t\eta'_{t-k} = 0$ for all k. In general, e_t can be serially correlated. The ith row of $\lambda(\mathrm{L})$, the lag polynomial $\lambda_i(\mathrm{L})$, is called the dynamic factor loading for the ith series, X_{it}.

The term $\lambda_i(\mathrm{L})f_t$ in (1) is the *common component* of the ith series. Throughout this chapter, we treat the lag polynomial $\lambda(\mathrm{L})$ as one sided. Thus the common component of each series is a distributed lag of current and past values of f_t.[c]

The idiosyncratic disturbance e_t in (1) can be serially correlated. If so, models (1) and (2) are incompletely specified. For some purposes, such as state-space estimation discussed later, it is desirable to specify a parametric model for the idiosyncratic dynamics. A simple and tractable model is to suppose that the ith idiosyncratic disturbance, e_{it}, follows the univariate autoregression,

[c] If $\lambda(\mathrm{L})$ has finitely many leads, then because f_t is unobserved the lag polynomial can without loss of generality be rewritten by shifting f_t so that $\lambda(\mathrm{L})$ is one sided.

$$e_{it} = \delta_i(\mathrm{L})e_{it-1} + \nu_{it}, \tag{3}$$

where ν_{it} is serially uncorrelated.

2.1.1.1 Exact DFM

If the idiosyncratic disturbances e_t are uncorrelated across series, that is, $Ee_{it}e_{js} = 0$ for all t and s with $i \neq j$, then the model is referred to as the *exact dynamic factor model.*

In the exact DFM, the correlation of one series with another occurs only through the latent factors f_t. To make this precise, suppose that the disturbances (e_t, η_t) are Gaussian. Then (1) and (2) imply that,

$$\begin{aligned} E\left[X_{it}|X_t^{-i}, f_t, X_{t-1}^{-i}, f_{t-1}, \ldots\right] &= E\left[\lambda_i(L)f_t + e_{it}|X_t^{-i}, f_t, X_{t-1}^{-i}, f_{t-1}, \ldots\right] \\ &= E\left[\lambda_i(L)f_t|X_t^{-i}, f_t, X_{t-1}^{-i}, f_{t-1}, \ldots\right] \\ &= \lambda_i(L)f_t, \end{aligned} \tag{4}$$

where the superscript "$-i$" denotes all the series other than i. Thus the common component of X_{it} is the expected value of X_{it} given the factors and all the other variables. The other series X_t^{-i} have no explanatory power for X_{it} given the factor.

Similarly, in the exact DFM with Gaussian disturbances, forecasts of the ith series given all the variables and the factors reduce to forecasts given the factors and X_{it}. Suppose that e_{it} follows the autoregression (3) and that (ν_t, η_t) are normally distributed. Under the exact DFM, $E\nu_{it}\nu_{jt} = 0$, $i \neq j$. Then

$$\begin{aligned} E[X_{it+1}|X_t, f_t, X_{t-1}, f_{t-1}, \ldots] &= E[\lambda_i(L)f_{t+1} + e_{it+1}|X_t, f_t, X_{t-1}, f_{t-1}, \ldots] \\ &= \alpha_i^f(\mathrm{L})f_t + \delta_i(\mathrm{L})X_{it}, \end{aligned} \tag{5}$$

where $\alpha_i^f(\mathrm{L}) = \lambda_{i0}\Psi(\mathrm{L}) - \delta_i(\mathrm{L})\lambda_i(\mathrm{L}) + \mathrm{L}^{-1}(\lambda_i(\mathrm{L}) - \lambda_0)$.[d]

If the disturbances (e_t, η_t) satisfy the exact DFM but are not Gaussian, then the expressions in (4) and (5) have interpretations as population linear predictors.

Eqs. (4) and (5) summarize the key dimension reduction properties of the exact DFM: for the purposes of explaining contemporaneous movements and for making forecasts, once you know the values of the factors, the other series provide no additional useful information.

2.1.1.2 Approximate DFM

The assumption that e_t is uncorrelated across series is unrealistic in many applications. For example, data derived from the same survey might have correlated measurement error,

[d] Substitute (2) and (3) into (1) to obtain, $X_{it+1} = \lambda_{i0}(\Psi(\mathrm{L})f_t + \eta_{t+1}) + \sum_j \lambda_{ij}f_{t-j+1} + \delta_i(\mathrm{L})e_{it} + \nu_{it+1}$. Note that $\sum_j \lambda_{ij}f_{t-j+1} = \mathrm{L}^{-1}(\lambda_i(\mathrm{L}) - \lambda_{i0})f_t$ and that $\delta_i(\mathrm{L})e_{it} = \delta_i(\mathrm{L})(X_{it} - \lambda_i(\mathrm{L})f_t)$. Then $X_{it+1} = \lambda_{i0}(\Psi(\mathrm{L})f_t + \eta_{t+1}) + \mathrm{L}^{-1}(\lambda_i(\mathrm{L}) - \lambda_{i0})f_t + \delta_i(\mathrm{L})(X_{it} - \lambda_i(L)f_t) + \nu_{it+1}$. Eq. (5) obtains by collecting terms and taking expectations.

and multiple series for a given sector might have unmodeled sector-specific dynamics. Chamberlain and Rothschild's (1983) *approximate factor model* allows for such correlation, as does the theoretical justification for the econometric methods discussed in Section 2.2. For a discussion of the technical conditions limiting the dependence across the disturbances in the approximate factor model, see Bai and Ng (2008).

Under the approximate DFM, the final expressions in (4) and (5) would contain additional terms reflecting this limited correlation. Concretely, the forecasting Eq. (5) could contain some additional observable variables relevant for forecasting series X_{it}. In applications, this potential correlation is best addressed on a case-by-case basis.

2.1.2 Static (Stacked) Form of the DFM

The *static*, or *stacked*, form of the DFM rewrites the dynamic form (1) and (2) to depend on *r static factors* F_t instead of the q dynamic factors f_t, where $r \geq q$. This rewriting makes the model amenable to principal components analysis and to other least-squares methods.

Let p be the degree of the lag polynomial matrix $\lambda(\mathrm{L})$ and let $F_t = \left(f_t', f_{t-1}', \ldots, f_{t-p}'\right)'$ denote an $r \times 1$ vector of so-called "static" factors—in contrast to the "dynamic" factors f_t. Also let $\Lambda = (\lambda_0, \lambda_1, \ldots, \lambda_p)$, where λ_h is the $N \times q$ matrix of coefficients on the hth lag in $\lambda(\mathrm{L})$. Similarly, let $\Phi(\mathrm{L})$ be the matrix consisting of 1s, 0s, and the elements of $\Psi(\mathrm{L})$ such that the vector autoregression in (2) is rewritten in terms of F_t. With this notation the DFM (1) and (2) can be rewritten,

$$X_t = \Lambda F_t + e_t \tag{6}$$

$$F_t = \Phi(\mathrm{L}) F_{t-1} + G\eta_t, \tag{7}$$

where $G = \begin{bmatrix} \mathrm{I}_q & 0_{q\times(r-q)} \end{bmatrix}'$.

As an example, suppose that there is a single dynamic factor f_t (so $q=1$), that all X_{it} depend only on the current and first lagged values of f_t, and that the VAR for f_t in (2) has two lags, so $f_t = \Psi_1 f_{t-1} + \Psi_2 f_{t-2} + \eta_t$. Then the correspondence between the dynamic and static forms for X_{it} is,

$$X_{it} = \lambda_{i0} f_t + \lambda_{i1} f_{t-1} + e_{it} = \begin{bmatrix} \lambda_{i0} & \lambda_{i1} \end{bmatrix} \begin{bmatrix} f_t \\ f_{t-1} \end{bmatrix} + e_{it} = \Lambda_i F_t + e_{it}, \tag{8}$$

$$F_t = \begin{bmatrix} f_t \\ f_{t-1} \end{bmatrix} = \begin{bmatrix} \Psi_1 & \Psi_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} f_{t-1} \\ f_{t-2} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_t = \Phi F_{t-1} + G\eta_t, \tag{9}$$

where the first expression in (8) writes out the equation for X_{it} in the dynamic form (1), $\Lambda_i = \begin{bmatrix} \lambda_{i0} & \lambda_{i1} \end{bmatrix}$ is the ith row of Λ, and the final expression in (8) is the equation for X_{it} in the static form (6). The first row in Eq. (9) is the evolution equation of the dynamic factor in (2) and the second row is the identity used to express (2) in first-order form.

In the static form of the DFM, the common component of the ith variable is $\Lambda_i F_t$, and the idiosyncratic component is e_{it}.

With the additional assumptions that the idiosyncratic disturbance follows the autoregression (3) and that the disturbances (ν_t, η_t) are Gaussian, the one step ahead forecast of the ith variable in the static factor model is,

$$E[X_{it+1}|X_t, F_t, X_{t-1}, F_{t-1}, \ldots] = \alpha_i^F(\mathrm{L})F_t + \delta_i(\mathrm{L})X_{it}, \tag{10}$$

where $\alpha_i^F = \Lambda_i\Phi(\mathrm{L}) - \delta_i(\mathrm{L})\Lambda_i$. If the disturbances are non-Gaussian, the expression is the population linear predictor.

The forecasting Eq. (10) is the static factor model counterpart of (5). In both forms of the DFM, the forecast using all the series reduces to a distributed lag of the factors and the individual series. The VAR (7) for F_t can be written in companion form by stacking the elements of F_t and its lags, resulting in a representation in which the stacked factor follows a VAR(1), in which case only current values of the stacked vector of factors enter (10).

Multistep ahead forecasts can be computed either by a direct regression onto current and past F_t and X_{it}, or by iterating forward the AR model for e_{it} and the VAR for F_t (Eqs. (3) and (7)).

In general, the number of static factors r exceeds the number of dynamic factors q because F_t consists of stacked current and past f_t. When $r > q$, the static factors have a dynamic singularity, that is, $q - r$ linear combinations of F_t are perfectly predictable from past F_t. In examples (8) and (9), there is a single dynamic factor and two static factors, and the perfectly predictable linear combination is $F_{2t} = F_{1t-1}$.

When the numbers of static and dynamic factors are estimated using macroeconomic data, the difference between the estimated values of r and q is often small, as is the case in the empirical work reported in Section 6. As a result, some applications set $r = q$ and $G = I$ in (7). Alternatively, if $q < r$, the resulting covariance matrix of the static factor innovations, that is, of $F_t - \Phi(\mathrm{L})F_{t-1} = G\eta_t$, has rank q, a constraint that can be easily imposed in the applications discussed in this chapter.

2.1.3 Normalization of the Factors

Because the factors are unobserved, they are identified only up to arbitrary normalizations. We first consider the static DFM, then the dynamic DFM.

In the static DFM, the space spanned by F_t is identified, but F_t itself is not identified: $\Lambda F_t = (\Lambda \mathrm{Q}^{-1})\ (\mathrm{Q}F_t)$, where Q is any invertible $r \times r$ matrix. For many applications, including macro monitoring and forecasting, it is necessary only to identify the space spanned by the factors, not the factors themselves, in which case Q in the foregoing expression is irrelevant. For such applications, the lack of identification is resolved by imposing a mathematically convenient normalization. The two normalizations discussed in this chapter are the "principal components" normalization and the "named factor" normalization.

2.1.3.1 Principal Components Normalization

Under this normalization, the columns of Λ are orthogonal and are scaled to have unit norm:

$$N^{-1}\Lambda'\Lambda = I_r \text{ and } \Sigma_F \text{ diagonal ("principal components" normalization)} \tag{11}$$

where $\Sigma_F = E\left(F_t F_t'\right)$.

The name for this normalization derives from its use in principal components estimation of the factors. When the factors are estimated by principal components, additionally the diagonal elements of Σ_F are weakly decreasing.

2.1.3.2 Named Factor Normalization

An alternative normalization is to associate each factor with a specific variable. Thus this normalization "names" each factor. This approach is useful for subsequent structural analysis, as discussed in Section 5 for structural DFMs, however it should be stressed that the "naming" discussed here is only a normalization that by itself it has no structural content.

Order the variables in X_t so that the first r variables are the naming variables. Then the "named factor" normalization is,

$$\Lambda^{NF} = \begin{bmatrix} I_r \\ \Lambda^{NF}_{r+1:n} \end{bmatrix}, \quad \Sigma_F \text{ is unrestricted ("named factor" normalization).} \tag{12}$$

Under the named factor normalization, the factors are in general contemporaneously correlated.[e]

The named factor normalization aligns the factors and variables so that the common component of X_{1t} is F_{1t}, so that an innovation to F_{1t} increases the common component of X_{1t} by one unit and thus increases X_{1t} by one unit. Similarly, the common component of X_{2t} is F_{2t}, so the innovation to the F_{2t} increases X_{2t} by one unit.

For example, suppose that the first variable is the price of oil. Then the normalization (12) equates the innovation in the first factor with the innovation in the common component of the oil price. The innovation in the first factor and the first factor itself therefore can be called the oil price factor innovation and the oil price factor.

The named factor normalization entails an additional assumption beyond the principal components normalization, specifically, that matrix of factor loadings on the first r variables (the naming variables) is invertible. That is, let $\Lambda_{1:r}$ denote the $r \times r$ matrix of factor loadings on the first r variables in the principal components normalization. Then $\Lambda^{NF}_{r+1:N} = \Lambda^{-1}_{1:r}\Lambda_{r+1:N}$. Said differently, the space of innovations of the first r common components must span the space of innovations of the static factors. In practice, the naming variables must be sufficiently different from each other, and sufficiently representative

[e] Bai and Ng (2013) refer to (11) and (12) normalizations as the PC1 and PC3 normalizations, respectively, and also discuss a PC2 normalization in which the first $r \times r$ block of Λ is lower triangular.

of groups of the other variables, that the innovations to their common components span the space of the factor innovations. This assumption is mild and can be satisfied by suitable choice of the naming variables.

2.1.3.3 Timing Normalization in the Dynamic Form of the DFM

In the dynamic form of the DFM, an additional identification problem arises associated with timing. Because $\lambda(\mathrm{L})f_t = [\lambda(\mathrm{L})q(\mathrm{L})^{-1}][q(\mathrm{L})f_t]$, where $q(\mathrm{L})$ is an arbitrary invertible $q \times q$ lag polynomial matrix, a DFM with factors f_t and factor loadings $\lambda(\mathrm{L})$ is observationally equivalent to a DFM with factors $q(\mathrm{L})\,f_t$ and factor loadings $\lambda(\mathrm{L})q(\mathrm{L})^{-1}$. This lack of identification can be resolved by choosing q variables on which f_t loads contemporaneously, without leads and lags, that is, for which $\lambda_i(\mathrm{L}) = \lambda_{i0}$.

2.1.4 Low-Frequency Movements, Unit Roots, and Cointegration

Throughout this chapter, we assume that X_t has been preprocessed to remove large low-frequency movements in the form of trends and unit roots. This is consistent with the econometric theory for DFMs which presumes series that are integrated of order zero (I(0)).

In practice, this preprocessing has two parts. First, stochastic trends and potential deterministic trends arising through drift are removed by differencing the data. Second, any remaining low-frequency movements, or long-term drifts, can be removed using other methods, such as a very low-frequency band-pass filter. We use both these steps in the empirical application in Sections 6 and 7, where they are discussed in more detail.

If some of the variables are cointegrated, then transforming them to first differences loses potentially important information that would be present in the error correction terms (that is, the residual from a cointegrating equation, possibly with cointegrating coefficients imposed). Here we discuss two different treatments of cointegrated variables, both of which are used in the empirical application of Sections 6 and 7.

The first approach for handling cointegrated variables is to include the first difference of some of the variables and error correction terms for the others. This is appropriate if the error correction term potentially contains important information that would be useful in estimating one or more factors. For example, suppose some of the variables are government interest rates at different maturities, that the interest rates are all integrated of order 1 (I(1)), that they are all cointegrated with a single common I(1) component, and the spreads also load on macro factors. Then including the first differences of one rate and the spreads allows using the spread information for estimation of their factors.

The second approach is to include all the variables in first differences and not to include any spreads. This induces a spectral density matrix among these cointegrated variables that is singular at frequency zero, however that frequency zero spectral density matrix is not estimated when the factors are estimated by principal components. This approach is appropriate if the first differences of the factors are informative for the common trend but the cointegrating residuals do not load on common factors. For example, in the empirical example in Sections 7 and 8, multiple measures of real oil prices are included in first differences. While there is empirical evidence that these oil prices, for example Brent and WTI, are cointegrated, there is no a priori reason to believe that the WTI-Brent spread is informative about broad macro factors, and rather that spread reflects details of oil markets, transient transportation and storage disruptions, and so forth. This treatment is discussed further in Section 7.2.

An alternative approach to handling unit roots and cointegration is to specify the DFM in levels or log levels of some or all of the variables, then to estimate cointegrating relations and common stochastic trends as part of estimating the DFM. This approach goes beyond the coverage of this chapter, which assumes that variables have been transformed to be I(0) and trendless. Banerjee and Marcellino (2009) and Banerjee et al. (2014, 2016) develop a factor-augmented error correction model (FECM) in which the levels of a subset of the variables are expressed as cointegrated with the common factors. The discussion in this chapter about applications and identification extends to the FECM.

2.2 DFMs: A Brief Review of Early Literature

Factor models have a long history in statistics and psychometrics. The extension to DFMs was originally developed by Geweke (1977) and Sargent and Sims (1977), who estimate the model using frequency-domain methods. Engle and Watson (1981, 1983) showed how the DFM can be estimated by maximum likelihood using time-domain state-space methods. An important advantage of the time domain over the frequency-domain approach is the ability to estimate the values of the latent factor using the Kalman filter. Stock and Watson (1989) used these state-space methods to develop a coincident real activity index as the estimated factor from a four-variable monthly model, and Sargent (1989) used analogous state-space methods to estimate the parameters of a six-variable real business cycle model with a single common structural shock.

Despite this progress, these early applications had two limitations. The first was computational: estimation of the parameters by maximum likelihood poses a practical limitation on the number of parameters that can be estimated, and with the exception of the single-factor 60-variable system estimated by Quah and Sargent (1993), these early applications had only a handful of observable variables and one or two latent factors. The second limitation was conceptual: maximum likelihood estimation requires specifying a full parametric model, which in practice entails assuming that the idiosyncratic components are mutually independent, and that the disturbances are normally distributed, a less appealing set of assumptions than the weaker ones in Chamberlain and Rothschild's (1983) approximate DFM.[f] For these reasons, it is desirable to have methods that can

[f] This second limitation was, it turns out, more perceived than actual if the number of series is large. Doz et al. (2012) show that state-space Gaussian quasi-maximum likelihood is a consistent estimator of the space spanned by the factors under weak assumptions on the error distribution and that allow limited correlation of the idiosyncratic disturbances.

handle many series and higher dimensional factor spaces under weak conditions on distributions and correlation among the idiosyncratic terms.

The state-space and frequency-domain methods exploit averaging both over time and over the cross section of variables. The key insight behind the nonparametric methods for estimation of DFMs, and in particular principal components estimation of the factors, is that, when the number of variables is large, cross-sectional variation alone can be exploited to estimate the space spanned by the factors. Consistency of the principal components (PC) estimator of F_t was first shown for T fixed and $N \to \infty$ in the exact factor model, without lags or any serial correlation, by Connor and Korajczyk (1986). Forni and Reichlin (1998) formalized the cross-sectional consistency of the unweighted cross-sectional average for a DFM with a single factor and nonzero average factor loading dynamics. Forni et al. (2000) showed identification and consistency of the dynamic PC estimator of the common component (a frequency-domain method that entails two-sided smoothing). Stock and Watson (2002a) proved consistency of the (time domain) PC estimator of the static factors under conditions along the lines of Chamberlain and Rothschild's (1983) approximate factor model and provided conditions under which the estimated factors can be treated as observed in subsequent regressions. Bai (2003) derived limiting distributions for the estimated factors and common components. Bai and Ng (2006a) provided improved rates for consistency of the PC estimator of the factors. Specifically, Bai and Ng (2006a) show that as $N \to \infty$, $T \to \infty$, and $N^2/T \to \infty$, the factors estimated by principal components can be treated as data (that is, the error in estimation of the factors can be ignored) when they are used as regressors.

2.3 Estimation of the Factors and DFM Parameters

The parameters and factors of the DFM can be estimated using nonparametric methods related to principal components analysis or by parametric state-space methods.

2.3.1 Nonparametric Methods and Principal Components Estimation

Nonparametric methods estimate the static factors in (6) directly without specifying a model for the factors or assuming specific distributions for the disturbances. These approaches use cross-sectional averaging to remove the influence of the idiosyncratic disturbances, leaving only the variation associated with the factors.

The intuition of cross-sectional averaging is most easily seen when there is a single factor. In this case, the cross-sectional average of X_t in (6) is $\bar{X}_t = \overline{\Lambda} F_t + \bar{e}_t$, where $\bar{X}_t$, $\overline{\Lambda}$, and $\bar{e}_t$, denote the cross-sectional averages $\bar{X}_t = N^{-1}\sum_{i=1}^{N} X_{it}$, etc. If the cross-sectional correlation among $\{e_{it}\}$ is limited, then by the law of large numbers $\bar{e}_t \xrightarrow{p} 0$, that is, $\bar{X}_t - \overline{\Lambda} F_t \xrightarrow{p} 0$. Thus if $\overline{\Lambda} \neq 0$, $\bar{X}_t$ estimates F_t up to scale. With more than one factor, this argument carries through using multiple weighted averages of X_t. Specifically, suppose that $N^{-1}\Lambda'\Lambda$ has a nonsingular limit; then the weighted average

$N^{-1}\Lambda' X_t$ satisfies $N^{-1}\Lambda' X_t - N^{-1}\Lambda'\Lambda F_t \xrightarrow{p} 0$, so that $N^{-1}\Lambda' X_t$ asymptotically spans the space of the factors. The weights $N^{-1}\Lambda$ are infeasible because Λ is unknown, however principal components estimation computes the sample version of this weighted average.

2.3.1.1 Principal Components Estimation

Principal components solve the least-squares problem in which Λ and F_t in (6) are treated as unknown parameters to be estimated:

$$\min_{F_1,\ldots,F_T,\Lambda} V_r(\Lambda, F), \quad \text{where } V_r(\Lambda, F) = \frac{1}{NT}\sum_{t=1}^{T}(X_t - \Lambda F_t)'(X_t - \Lambda F_t), \tag{13}$$

subject to the normalization (11). Under the exact factor model with homogeneous idiosyncratic variances and factors treated as parameters, (13) is the Gaussian maximum likelihood estimator (Chamberlain and Rothschild, 1983). If there are no missing data, then the solution to the least-squares problem (13) is the PC estimator of the factors, $\hat{F}_t = N^{-1}\hat{\Lambda}' X_t$, where $\hat{\Lambda}$ is the matrix of eigenvectors of the sample variance matrix of X_t, $\hat{\Sigma}_X = T^{-1}\sum_{t=1}^{T} X_t X_t'$, associated with the r largest eigenvalues of $\hat{\Sigma}_X$.

2.3.1.2 Generalized Principal Components Estimation

If the idiosyncratic disturbances have different variances and/or some are cross correlated, then by analogy to generalized least squares, efficiency gains should be possible by modifying the least-squares problem (13) for a more general weight matrix. Specifically, let Σ_e denote the error variance matrix of e_t; then the analogy to generalized least-squares regression suggests that F_t and Λ solve a weighted version of (13), where the weighting matrix is Σ_e^{-1}:

$$\min_{F_1,\ldots,F_T,\Lambda} T^{-1}\sum_{t=1}^{T}(X_t - \Lambda F_t)'\Sigma_e^{-1}(X_t - \Lambda F_t). \tag{14}$$

A solution to (14) is the infeasible generalized PC estimator, $\tilde{F}_t = N^{-1}\tilde{\Lambda}' X_t$, where $\tilde{\Lambda}$ are the scaled eigenvectors corresponding to the r largest eigenvalues of $\Sigma_e^{-1/2}\hat{\Sigma}_X\Sigma_e^{-1/2'}$.[g]

The feasible generalized PC estimator replaces the unknown Σ_e in (14) with an estimator $\hat{\Sigma}_e$. Choi (2012) shows that if $\hat{\Sigma}_e$ is consistent for Σ_e then the feasible generalized PC estimator of $\{F_t\}$ and Λ is asymptotically more efficient than principal components. Several estimators of Σ_e have been proposed. The limited amount of evidence from simulation and empirical work comparing their performance suggests that a reasonable approach is to use Boivin and Ng's (2006) two-step diagonal weight matrix approach, in which the first step is principal components (that is, identity weight matrix) and

[g] As stated in the beginning of this section, the series in X are typically preprocessed to have unit standard deviation, so in this sense the unweighted principal components estimator (13) implicitly also has weighting if it is expressed in terms of the nonstandardized data.

the second step uses a diagonal $\hat{\Sigma}_e$, where the diagonal element is the sample variance of the estimated idiosyncratic component from the first step.

Other approaches include Forni et al.'s (2005), which allows for contemporaneous covariance across the idiosyncratic terms but does not adjust for serial correlation, and Stock and Watson's (2005) and Breitung and Tenhofen's (2011), which adjusts for serial correlation and heteroskedasticity in e_{it} but not cross correlation. See Choi (2012) for additional discussion.

2.3.1.3 Extension to Restrictions on Λ

The principal components methods described in Sections 2.3.1.1 and 2.3.1.2 apply to the case that Λ and F are exactly identified using the principal components normalization. If there are additional restrictions on Λ, then principal components no longer applies but the least-squares concept does. Specifically, minimization can proceed using (13), however Λ is further parameterized as $\Lambda(\theta)$ and minimization now proceeds over θ, not over unrestricted Λ.

In general this minimization with respect to θ entails nonlinear optimization. In some leading cases, however, closed-form solutions to the least-squares problem are available. One such case is a hierarchical DFM in which there are common factors that affect all variables, and group-level factors that affect only selected variables; for example, suppose the groups are countries, the group factors are country factors, and the cross-group common factors are international factors. If the factors are normalized to be orthogonal, the first-level factors can be estimated by principal components using all the series, then the factors unique to the gth group can be estimated by principal components using the residuals from projecting the group-g variables on the first-level factors. A second case is when the restrictions are linear, so that $\text{vec}(\Lambda) = R\theta$, where R is a fixed known matrix; in this case, standard regression formulas provide an explicit representation of the minimizer $\hat{\theta}$ given $\{\hat{F}_t\}$ and vice versa.

2.3.2 Parametric State-Space Methods

State-space estimation entails specifying a full parametric model for X_t, e_t, and f_t in the dynamic form of the DFM, so that the likelihood can be computed.

For parametric estimation, additional assumptions need to be made on the distribution of the errors and the dynamics of the idiosyncratic component e_t in the DFM. A common treatment is to model the elements of e_t as following the independent univariate autoregressions (3). With the further assumptions that the disturbances ν_{it} in (3) are i.i.d. $\text{N}(0, \sigma^2_{\nu_i})$, $i = 1, \ldots, N$, η_t is i.i.d. $\text{N}(0, \Sigma_\eta)$, and $\{\nu_t\}$ and $\{\eta_t\}$ are independent, Eqs. (1)–(3) constitute a complete linear state-space model. Alternatively, the static DFM can be written in state-space form using (6), (7), and (3).

Given the parameters, the Kalman filter can be used to compute the likelihood and the Kalman smoother can be used to compute estimates of f_t given the full-sample data on

$\{X_t\}$. The likelihood can be maximized to obtain maximum likelihood estimates of the parameters. Alternatively, with the addition of a prior distribution, the Kalman filter can be used to compute the posterior distribution of the parameters and posterior estimates of the unobserved factors can be computed from the Kalman smoother. The fact that the state-space approach uses intertemporal smoothing to estimate the factors, whereas principal components approaches use only contemporaneous smoothing (averaging across series at the same date) is an important difference between the methods.

Parametric state-space methods have several advantages, including the use of quasi-maximum likelihood estimation, the possibility of performing Bayes inference, efficient treatment of missing observations (this latter point is discussed further in the next section), and the use of intertemporal smoothing to estimate the factors. However, state-space methods also have drawbacks. Historically, their implementation becomes numerically challenging when N is large because the number of parameters grows proportionately to N, making maximum likelihood estimation of the parameter vector prohibitive.[h] In addition, state-space methods require specifying the degree of the factor loading lag polynomial and models for the factors and for the idiosyncratic terms. These modeling choices introduce potential misspecification which is not reflected in the model-based inference, that is, standard errors and posterior coverage regions are not robust to model misspecification.

2.3.3 Hybrid Methods and Data Pruning

2.3.3.1 Hybrid Methods

One way to handle the computational problem of maximum likelihood estimation of the state-space parameters is to adopt a two-step hybrid approach that combines the speed of principal components and the efficiency of the Kalman filter (Doz et al., 2011). In the first step, initial estimates of factors are obtained using principal components, from which the factor loadings are estimated and a model is fit to the idiosyncratic components. In the second step, the resulting parameters are used to construct a state-space model which then can be used to estimate F_t by the Kalman filter. Doz et al. (2011) show that, for large N and T, the resulting estimator of the factors is consistent for the factor space and is robust to misspecification of the correlation structure of the idiosyncratic components, and thus has a nonparametric interpretation.

2.3.3.2 Pruning Datasets and Variable Selection

The discussion so far assumes that all the variables have been chosen using a priori knowledge to include series that are potentially valuable for estimating the factors. Because the emphasis is on using many variables, one possibility is that some extraneous variables

[h] Durbin and Koopman (2012, section 6.5) discuss computationally efficient formulae for Kalman filtering when N is large.

could be included, and that it might be better to eliminate those variables. Whether this is a problem, and if so how to handle it, depends on the empirical application. If there is a priori reason to model the factors as applying to only some variables (for example, there are multiple countries and interest is in obtaining some country-specific and some international factors) then it is possible to use a hierarchical DFM. In effect this prunes out variables of other countries when estimating a given country factors. Another approach is to use prescreening methods to prune the dataset, see for example Bai and Ng (2006a). Alternatively, sparse data methods can be used to eliminate some of the variables, for example using a sparsity prior in a state-space formulation (eg, Kaufmann and Schumacher, 2012).

2.3.4 Missing Data and Mixed Data Sampling Frequencies

Missing data arise for various reasons. Some series might begin sooner than others, the date of the final observation on different series can differ because of timing of data releases, and in some applications the series might have different sampling frequencies (eg, monthly and quarterly). The details of how missing data are handled differ in principal components and state-space applications. All the procedures in common use (and, to the best of our knowledge, all the procedures in the literature) adopt the assumption that the data are missing at random. Under the missing-at-random assumption, whether a datum is missing is independent of the latent variables (no endogenous sample selection). The missing-at-random assumption arguably is a reasonable assumption for the main sources of missing data in DFMs in most macroeconomic applications to date.

2.3.4.1 Principal Components Estimation with Missing Data

The solution to the least-squares problem (13) in terms of the eigenvalues of $\hat{\Sigma}_X$ holds when all NT observations are nonmissing, that is, when the panel is balanced. When there are missing observations, least-squares still can be used to estimate F_t and Λ, however the solution must be obtained numerically. Specifically, the modification of (13) when there is missing data is,

$$\min_{F_1,\ldots,F_T,\Lambda} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} S_{it}(X_{it} - \Lambda_i F_t)^2, \tag{15}$$

where $S_{it} = 1$ if an observation on X_{it} is available and $S_{it} = 0$ otherwise and where Λ_i is the ith row of Λ. The objective function in (15) can be minimized by iterations alternating with Λ given $\{F_t\}$ then $\{F_t\}$ given Λ; each step in the minimization has a closed-form expression. Starting values can be obtained, for example, by principal component estimation using a subset of the series for which there are no missing observations. Alternatively, Stock and Watson (2002b) provide an EM algorithm for handling missing observations.

Given an estimate of the factor loadings and factors based on missing data, the estimated common component for the ith series remains $\hat{\Lambda}_i\hat{F}_t$ and the one step ahead forecast is given by (10), where the parameters of (10) are estimated treating $\hat{F}_t$ as data.

2.3.4.2 State-Space Estimation with Missing Data

The state-space framework can be adapted to missing data by allowing the measurement Eq. (1) to vary depending on what data are available at a given date t; see Harvey (1989, p. 325). Alternatively, the dimension of the measurement equation can be kept the same by including a proxy value for the missing observation while adjusting the model parameters so that the Kalman filter places no weight on the missing observation. See Giannone et al. (2008), Mariano and Murasawa (2010), and Marcellino and Sivec (2014) for variations on this latter approach.

For large N, one computational challenge is keeping the dimension of the state vector small as N grows, which is more complicated with missing observations than with all observations nonmissing; see Jungbacker et al. (2011) and Bańbura and Modugno (2014) for discussion and proposed computationally efficient solutions.

One theoretical advantage of the state-space approach to mixed frequencies is that it can pin down when precisely the measurement occurs (eg, the US establishment survey measures payroll employment during the week including the 12th of the month). A second theoretical advantage of the state-space approach is that it can explicitly differentiate between stock variables (observed at a point in time, like employment) and flow variables (temporal averages, like GDP). In practice, dealing with flows is complicated, however, because the flow aggregation identities are in levels but the variables being measured, such as sectoral output, are typically best modeled in growth rates. These complications require approximations and can substantially increase the dimension of the latent state variable. For an application with mixed sampling frequencies and mixed stock and flow variables, see Aruoba et al. (2009). See Foroni and Marcellino (2013) for a survey of methods for handling mixed-frequency data, including DFMs and alternative approaches.

There appears to be little research comparing the performance of parametric and nonparametric approaches to mixed-frequency data.

2.3.5 Bayes Methods

An alternative approach to estimating DFMs is to use Bayes methods. In Bayesian estimation, the DFM parameters are treated as random draws from a prior distribution. Because the factors are unobserved and multiplied by the coefficients, Bayesian inference is more complicated than it is in the standard regression model with observed regressors and conjugate priors, and Bayesian DFM estimation requires using modern numerical techniques.

The first Bayesian treatments of DFMs of which we are aware are Kim and Nelson (1998) and Otrok and Whiteman (1998), who both estimated a small single-factor system using Markov Chain Monte Carlo methods. Kim and Nelson (1998) also incorporated Markov switching in the process for the latent factor. In other early work, Kose et al. (2003) extend Otrok and Whiteman (1998) to a 180-variable system with international macroeconomic data, using a hierarchical regional/country structure. Aguilar and West (2000) developed Bayes methods for estimating dynamic factor models with stochastic volatility, which they apply to multivariate financial time series.

A theoretical advantage of Bayes methods is that the mean squared error of some functions of the estimated parameters (such as in forecast functions) can be reduced by shrinkage. Koopman and Mesters (forthcoming) take an empirical Bayes approach to estimating the efficient amount of shrinkage. Their algorithm iterates between estimation of the factors by Gaussian signal extraction (Kalman smoother) and Bayes estimation of the parameters given the consistently estimated factors.

To date, the dominant methods used in macro applications are Frequentist, especially the computationally straightforward methods based on principal components. This chapter therefore focuses on Frequentist methods for estimation of DFMs. However, because the number of parameters in Λ is large, Bayes methods for DFMs are a promising area for improving estimator and forecast performance from a Frequentist perspective.

2.4 Determining the Number of Factors

2.4.1 Estimating the Number of Static Factors r

The number of static factors r can be determined by a combination of a priori knowledge, visual inspection of a scree plot, and the use of information criteria and other statistical measures.

2.4.1.1 Scree Plots

A *scree plot* displays the marginal contribution of the kth principal component to the average R^2 of the N regressions of X_t against the first k principal components. This marginal contribution is the average additional explanatory value of the kth factor. When there are no missing data, the scree plot is a plot of the ordered eigenvalues of $\hat{\Sigma}_X$, normalized by the sum of the eigenvalues.

2.4.1.2 Information Criteria

Information criteria, such as the Akaike information criterion, use a penalized objective function to trade off the benefit of including an additional parameter against the cost of increased sampling variability. Bai and Ng (2002) extend this idea to including an additional factor using the penalized sum of squares,

$$IC(r) = \ln V_r\left(\hat{\Lambda}, \hat{F}\right) + rg(N, T), \tag{16}$$

where $V_r\left(\hat{\Lambda}, \hat{F}\right)$ is the least-squares objective function in (13) evaluated at the PCs $\left(\hat{\Lambda}, \hat{F}\right)$, and where $g(N,T)$ is a penalty factor such that $g(N,T) \rightarrow 0$ and $\min(N,T)g\ (N,T) \rightarrow \infty$ as $N,\ T \rightarrow \infty$. Bai and Ng (2002) provide conditions under which the value of r that minimizes an information criterion with $g(N,T)$ satisfying these conditions is consistent for the true value of r. A commonly used penalty function is the Bai and Ng (2002) IC_{p2} penalty, for which $g(N,T) = [(N+T)/NT]\ln[\min(N,T)]$. When $N = T$, this penalty simplifies to two times the BIC penalty, $T^{-1}\ln T$. Monte Carlo evidence suggests that this penalty function works well in designs calibrated to macroeconomic data.

2.4.1.3 Other Approaches

Onatski (2010) provides an alternative consistent estimator of r which estimates r as the largest value of k for which the difference between eigenvalues k and $k+1$ of $\hat{\Sigma}_X$ exceeds a threshold provided in that paper; this estimator corresponds to finding the final "cliff" in the scree plot larger than that threshold. Similarly, Ahn and Horenstein (2013) show that an alternative consistent estimator of r is obtained as the maximizer of the ratio of eigenvalue k to eigenvalue $k+1$; their estimator corresponds to locating the largest "relative cliff" in the scree plot. Onatski (2009) takes a different approach and considers tests as opposed to estimation of r by information criteria.

Practical experience suggests that different methods frequently give different estimates. There is limited research comparing the performance of the different methods. This sensitivity suggests that it is important to augment the statistical estimators with inspection of the scree plot and with judgment informed by the application at hand.

2.4.2 Estimating the Number of Dynamic Factors q

In principle, the number of dynamic factors can be less than the number of static factors and if so, the static factors follow a singular dynamic process. Framed in terms of (7), these singularities arise because the covariance matrix of the innovations to F_t (that is, $G\eta_t$ in (7)) is singular with rank $q < r$. This implies that the spectral density matrix of F_t is singular. Estimation of q given r entails estimating the rank of this singularity. Although in principle an information criterion could be used to estimate the number of dynamic factors based on the likelihood of the dynamic form of the DFM, estimating q given r has the advantage that it is unnecessary to compute that likelihood.

There are three related methods for consistently estimating q given r. Amengual and Watson (2007) first compute the residual of the projection of X_t onto lagged values of the PC estimator of F_t, then apply the Bai and Ng (2002) information criterion to the covariance matrix of those residuals. Bai and Ng (2007) work directly with the factors and use an information criterion to estimate the rank of the residual covariance matrix of a VAR estimated using the r principal components. In contrast to these two approaches, Hallin and Liška (2007) propose a frequency-domain procedure which uses an information criterion to estimate the rank of the spectral density matrix of X_t. There seems to be limited research comparing these methods.

2.5 Breaks and Time-Varying Parameters

The discussion so far has considered DFMs with time-invariant parameters. In many applications, however, there is at least the possibility of parameter instability. This section reviews the robustness of PC estimator of the factors to small breaks. If, however, the instability is large and widespread, the full-sample PC estimator breaks down. As a result, in many applications it is important to check for and/or model structural instability in the factor loadings. There are two broad approaches to handling instability in DFMs: positing a break in the parameters, and modeling the parameters as evolving stochastically.

2.5.1 Robustness of PC to Limited Instability

If the amount of instability is small and/or limited across variables, the PC estimator of the factors remains consistent. The intuition behind this initially surprising result can be seen by returning to the example of Section 2.3.1 of the cross-sectional average when there is a single factor. Suppose that the static factor loading matrix is time dependent, so that Λ in (6) is replaced by Λ_t. Then $\bar{X}_t = \bar{\Lambda}_t F_t + \bar{e}_t$, where $\bar{\Lambda}_t$ is the cross-sectional average of Λ_t. Let $\bar{\bar{\Lambda}}$ denote the time average of $\bar{\Lambda}_t$. Then $\bar{X}_t - \bar{\bar{\Lambda}} F_t = \left(\bar{\Lambda}_t - \bar{\bar{\Lambda}}\right) F_t + \bar{e}_t$. If only a vanishing fraction of series have a break in their factor loadings, or if the breaks in Λ_{it} are stochastic, have limited temporal dependence, and are uncorrelated across series, or if Λ_{it} has persistent drift which has mean zero and is uncorrelated across series, then by the law of large numbers $\bar{\Lambda}_t - \bar{\bar{\Lambda}} \xrightarrow{p} 0$ and $\bar{e}_t \xrightarrow{p} 0$ so that $\bar{X}_t - \bar{\bar{\Lambda}} F_t \xrightarrow{p} 0$. Thus, despite this nontrivial instability, if $\bar{\bar{\Lambda}}$ is nonzero, $\bar{X}_t$ estimates the factor up to scale.

Bates et al. (2013) provide general conditions on parameter instability under which the PC estimator remains consistent. They show, for example, that the factor estimates remain consistent if there is a large discrete break in the factor loadings for a fraction $O(N^{-1/2})$ of the series, or if the factor loadings follow independent random walks with relatively small innovations, as long as those innovations are independent across series.[i] For these instabilities, tests for stability of Λ would reject with probability tending to one in large samples but the PC estimator remains consistent.[j]

Despite these robustness results for the estimated factors, the coefficients in any specific equation could have large drift or breaks. Stock and Watson (2009) provide evidence

[i] Specifically, Bates et al. (2013) show that if $\Lambda_t = \Lambda_0 + h_{NT}\xi_t$, where $h_{NT} = O(1/\min[N^{1/4}T^{1/2}, T^{3/4}])$, then the estimated factors achieve the Bai and Ng (2002) mean square consistency rate of $1/\min(N,T)$.

[j] Stock and Watson (2009) provide some empirical evidence that suggests the relevance of such breaks. In a pseudo out-of-sample forecasting exercise using US macroeconomic data, they find evidence of a break in 1984 in the factor loadings, but also find that the best forecasts are produced by estimating the factors over the full data span but estimating the factor loadings over the post-1984 subset.

that allowing for such instability can be important in practice when interest is in a specific series (say, for forecasting), even if full-sample principal components estimates of the factors are used.

2.5.2 Tests for Instability

Despite this insensitivity of the PC estimator to some forms of instability in the factor loadings, principal components is not robust to widespread large breaks or to large time variation in Λ that is systematically correlated across series. Following Stock and Watson (2009) and Breitung and Eickmeier (2011), consider the case in which Λ takes on two values:

$$X_t = \Lambda_t F_t + e_t, \quad \Lambda_t = \begin{cases} \Lambda^{(1)} \text{ if } t < \tau \\ \Lambda^{(2)} \text{ if } t \geq \tau \end{cases}. \tag{17}$$

For this discussion, suppose the dynamics of the factor structure does not change. Thus the DFM holds in both regimes, with the same r factors, but with different factor loadings. As shown by Stock and Watson (2009) and Breitung and Eickmeier (2011), if the break in Λ is widespread across the series, the split-sample PC estimators of the factors will differ from each other. Moreover, if there are r factors in each subsample and a widespread break in Λ, then in the full sample it will appear as though there are $2r$ factors. Breitung and Eickmeier (2011) provide Monte Carlo evidence that as a result the Bai and Ng (2002) procedure would systematically overestimates the number of factors.

There are now a number of tests for breaks in the factor loadings. Stock and Watson (2009) consider the problem of breaks in a single equation and suggest regressing each variable on the estimated factors and implementing break tests for each regression. Breitung and Eickmeier (2011) consider a related Lagrange multiplier test that handles breaks in a fixed finite number of DFM equations; their test appears to improve size control, relative to the Stock and Watson (2009) approach. Tests proposed by Chen et al. (2014) and Han and Inoue (2015) test for a general break in Λ (all equations) by noting that, if Λ changes, the covariance matrix of the full-sample PC estimator will change at the break date in Λ. Chen et al.'s (2014) test entails testing for a break in the regression of one of the estimated factors on the others. Han and Inoue (2015) test for a break in the full covariance matrix of the PC estimator of the factors. All the foregoing break tests generalize to unknown break dates using standard methods. Cheng et al. (Forthcoming) take a different approach and extend LASSO methods to consider changes in the factor loadings and/or changes in the number of factors.

Care must be taken when interpreting these break tests for at least two reasons. First, although these tests are for a discrete break, break tests have power against other types of parameter instability, in particular against drifting parameters.[k]

[k] See, for example, Stock and Watson (1998) and Elliott and Müller (2006).

Second, a more subtle issue of interpretation is that, although these tests are designed to detect breaks in Λ and thus breaks in the factor space, at least some of them will have power against heteroskedasticity in the factor innovations and/or breaks in the VAR process followed by the factors. This power against heteroskedasticity in some tests but not others arises because of different normalizations used in the tests. In principle, these different sources of instability—breaks in Λ, heteroskedasticity in the factor innovations, and breaks in the VAR process for F_t—are separately identified. These tests are new and their relative power against different types of breaks has not been studied in any detail. Because the modeling and substantive implications of a widespread break in Λ are quite different from those of a change in the volatility of the factor innovations, interpretation of rejections must be sensitive to this ambiguity.[1]

2.5.3 Incorporating Time-Varying Factor Loadings and Stochastic Volatility

Although tests for stability can detect breaks or evolution of the DFM parameters, the empirical significance of that instability must be assessed by estimating the model taking into account the instability.

The most straightforward way to estimate the DFM taking into account the instability is through subsample estimation. However, doing so presumes a single common break date, and in many applications one might be concerned about continuous parameter drift, volatility clustering, or breaks for different series at different dates. If so, then it is appropriate to use a more flexible model of parameter change than the single common break model.

An alternative approach to time variation is to model the parameters as evolving stochastically rather than breaking at a single date. If parameter variation is small, this approach can be implemented in two steps, first estimating the factors by least squares, then estimating a time-varying model treating the factors as observed. See, for example, Cogley and Sargent (2005) for time-varying parameter VAR methods for observed variables; for recent contributions and references see Korobilis (2014). Eickmeier et al. (2015)

[1] Empirical work applying break tests to DFMs suggests that DFM parameters have changed over the postwar sample. In particular, there is evidence of a break in the factor loadings around onset of the Great Moderation. Stock and Watson (2009) find evidence of a break in 1984, the only date they consider. Breitung and Eickmeier (2011) apply their tests for breaks at an unknown date and find breaks in multiple equations with estimated break dates around 1984. Chen et al. (2014) also find breaks around 1980. Stock and Watson (2012a) and Cheng et al. (Forthcoming) find evidence of breaks at the onset of the 2007 recession. Stock and Watson (2012a) find that this break is in the variances of the factor innovations (in Σ_η), whereas Cheng et al. find that the breaks are in Λ. However, the Cheng et al. normalization imposes homoskedasticity in the factor innovations, so in their test a change in Σ_η would appear as a change in Λ; thus both sets of results are consistent with the break being in Σ_η. All these papers examine quarterly US data.

work through the details of this two-step approach to time variation in DFMs. Using the results in Bates et al. (2013) as motivation, Eickmeier et al. (2015) suggest estimating the factors by principal components and treating them as observed. The time variation in the DFM is now easily handled equation-by-equation. They apply these methods in a time-varying FAVAR, but the methods equally apply to DFMs once one treats the estimated factors as observed.

If, however, the parameter variation is large then (as discussed in the previous section) this approach will yield misleading estimates of the factors. Consequently, recent work has focused on treating the factors as unobserved while allowing for and estimating time-varying stochastic processes for the factor loadings. An additional extension is to stochastic volatility in the innovations to the factors and idiosyncratic terms, which allows both for additional time variation in the implied filter and for volatility clustering in the data.

Much of the current work on time-varying DFMs uses or extends the model of del Negro and Otrok (2008). Their model allows the factor loadings to evolve according to a random walk: $\Lambda_{it} = \Lambda_{it-1} + \sigma_{\Delta\Lambda,i}\zeta_{\mathrm{it}}$, where ζ_{it} is an i.i.d. $N(0,1)$ disturbance. They also allow for time variation in the factor VAR coefficients and in the autoregressive coefficients describing the idiosyncratic dynamics. Finally, del Negro and Otrok (2008) allow for stochastic volatility in the innovations to the factors and to the idiosyncratic disturbances. The result of these extensions of the DFM is that the state evolution equation is a nonlinear function of the state variables so that while it remains a hidden Markov model, it can no longer be estimated by the Kalman filter. Del Negro and Otrok (2008) show how the model can instead be estimated by numerical Bayes methods. Papers that apply this algorithm or variants to DFMs with time-varying parameters include Mumtaz and Surico (2012), Bjørnland and Thorsrud (2015a), and Stock and Watson (2015). The details of these methods go beyond the scope of this chapter.

3. DFMs FOR MACROECONOMIC MONITORING AND FORECASTING

Two classic applications of DFMs are to real-time macroeconomic monitoring and to forecasting. The early hope of some researchers for DFMs—initially small DFMs and later "big data" high-dimensional DFMs—was that their ability to extract meaningful signals (factors) from noisy data would provide a breakthrough in macroeconomic forecasting. This early optimism turned out to be misplaced, arguably mainly because so many of the shocks that matter the most for the economy, such as the invasion of Kuwait by Iraq in August 1990 and the financial crisis in the fall of 2008, are simply not known in advance. This said, DFMs have resulted in meaningful forecasting improvements, especially for measures of real economic activity. They have also proven particularly useful for the important task of macroeconomic monitoring, that is, tracking economies in real time. The literature on using DFMs for forecasting and macro monitoring is vast. This section provides a selective survey of that literature, discusses some technical issues at a high level, and provides references for readers interested in the technical details.

3.1 Macroeconomic Monitoring

Economists at central banks, executive branches of government, and in the private sector track the evolution of the economy in real time, that is, they monitor the macroeconomy. A key part of macroeconomic monitoring is following and interpreting data releases to glean insights as to where the economy is at present, and where the economy is going. Macroeconomic monitoring has two salient challenges. First, data releases are peppered throughout the month and quarter, so that the available data change from day to day or even within a day, a feature referred to as the "ragged edge" problem. Second, the number of data releases and series contained within those releases is vast. Handling this flow of large volumes of disparate data requires judgment and knowledge of idiosyncratic events. Increasingly, the job of macroeconomic monitoring has also benefited from systematic high-dimensional modeling in the form of DFMs.

DFMs are used for two related macro monitoring tasks. The first is the construction of indices that distill the currently available data into a concise summary of economic conditions. The second is *nowcasting*, which is the task of "forecasting" the current value of a specific series which has not yet been released, for example, forecasting the value of fourth-quarter GDP in November.

3.1.1 Index Construction

A natural application of DFMs is to a classic problem in empirical macroeconomics, the construction of an index of indicators of economic activity. In the DFM, the latent factor summarizes the comovements of the observed variables, so in a DFM with a single factor, the estimate of the latent factor is a natural index of the movements of the relevant time series.

The first application of DFMs for real-time macromonitoring was the Stock and Watson (1989, 1991) experimental coincident index (XCI), which was released monthly through the National Bureau of Economic Research from May 1989 to December 2003. The XCI was the Kalman filter estimate of the single common factor among four monthly coincident indices: total nonfarm employment, the index of IP, real manufacturing and trade sales, and real personal income less transfers. The DFM was estimated by maximum likelihood in state-space form. This system handled the "ragged edge" problem of one of the series (real personal income less transfers) being available with a substantial delay, so the initial release of the index used a reduced-dimension measurement equation for the final observation. Retrospective analysis of the real-time experience showed that the XCI was successful in contemporaneous monitoring and (using a companion model for the probability of recessions) in real-time detection of the recession of 1990, however, the XCI and its associated leading index did not forecast the recession at the target 6-month horizon (Stock and Watson, 1993).

Subsequent work with small state-space DFMs include the construction of monthly real activity indices for US states (Crone and Clayton-Matthews, 2005), which has been released in real time by the Federal Reserve Bank of Philadelphia since 2005. Mariano and Murasawa (2003) extended the XCI to mixed-frequency data by including quarterly GDP. Aruoba et al. (2009) developed a weekly index using mixed-frequency data (weekly, monthly, and quarterly), and the resulting "ADS" index is released in real time by the Federal Reserve Bank of Philadelphia.

Much of the recent work on index construction has focused on higher dimensional systems. Since January 2001, the Federal Reserve Bank of Chicago has released in real time the monthly Chicago Fed National Activity Index (CFNAI), which is the principal components estimate of the common factor in 85 real activity variables based on the real activity index constructed in Stock and Watson (1999). Since January 2002, the UK Centre for Economic Policy Research has released in real time the monthly EuroCOIN index of EU real economic activity. EuroCOIN was developed by Altissimo et al. (2001) and initially incorporated 951 Euro-area activity variables.[m] The index was updated in Altissimo et al. (2010); that version entails estimating the factors by principal components using 145 Euro-area real activity variables.

3.1.2 Nowcasting

Nowcasting focuses on predicting the current value of observable variables, such as current-quarter GDP. Nowcasting has long been done by economists using methods that allow the use of mixed-frequency data and intermittent releases. The older methods do not specify joint distributions and in general are variable-specific, often without a model structure tying together nowcasts across variables or over time as data become available. In contrast, DFMs permit specifying an internally consistent model that can be used for nowcasting multiple variables while placing appropriate weight on new data releases. Early nowcasting applications that use high dimensions and mixed frequencies in a state-space setting are Evans (2005), Giannone et al. (2008), and Angelini et al. (2010). Aastveit et al. (2014) extend these methods to compute density nowcasts (not just point nowcasts) of GDP growth. Bańbura, Giannone, Modugno, and Reichlin (2013) survey recent developments and technical issues in nowcasting.

[m] The index is calibrated to the smoothed component of GDP growth, specifically the reported index is the common component of Euro-area GDP, filtered to eliminate high-frequency variation.

3.2 Forecasting

The literature on forecasting with DFMs is very large and we do not attempt a comprehensive survey, instead we make some high-level comments. Eickmeier and Ziegler (2008) provide a survey and meta-analysis of work in the field through the mid-2000s. They find that factor forecasts tend to outperform small-model forecasts, and that factor forecasts tend to work better for US real activity than for US inflation. For more recent references, extensions of DFM forecasting methods, and comparisons to other high-dimensional methods, see Stock and Watson (2012b), D'Agostino and Giannone (2012), Clements (Forthcoming), and Cheng and Hansen (2015).

4. IDENTIFICATION OF SHOCKS IN STRUCTURAL VARs

This section provides a self-contained survey of contemporary methods for identification of structural VARs. The methods are presented in a unified way that allows them to be adapted directly to structural DFMs, as discussed in the next section.

A long-standing goal of empirical macroeconomics is to estimate the effect on the economy of unanticipated structural disturbances, commonly called shocks. Examples of shocks include an unanticipated rate hike by the central bank (a monetary policy shock), an unexpected jump in oil prices due to oil supply disruptions (oil supply shock), an unexpected improvement in productivity (productivity shock), and an unanticipated shift in aggregate demand (demand shock). These shocks induce unexpected changes in the values of economic variables, for example, a contractionary monetary policy shock increases the short-term interest rate. Because these shocks are autonomous, they are uncorrelated with other shocks. Because shocks are unanticipated, they are serially uncorrelated.[n]

If a time series of shocks were observed, it would be straightforward to estimate the effect of that shock, say ε_{1t}, on a macro variable y_t by regressing y_t on current and past values of ε_{1t}. Because the shock ε_{1t} is uncorrelated with the other shocks to the economy, that regression would have no omitted variable bias. The population coefficients of that regression would be the dynamic causal effect of that shock on the dependent variable, also called the structural impulse response function (SIRF). The cumulative sum of those population coefficients would be the cumulative causal effect of that shock over time, called the cumulative SIRF. Thus if the time series of shocks were observed, its dynamic effect could be estimated in a way that required no additional modeling assumptions. Unfortunately, a complete time series of shocks is rarely if ever observed—a constructed time series of shocks will have measurement error and/or miss some events—so that this ideal regression of y_t on current and past ε_{1t} typically is infeasible.

[n] See the chapter by Ramey (2016, this Handbook) for an extensive discussion of shocks in structural VARs.

Because direct observation of a complete series of shocks without measurement error typically is infeasible, a large number of methods have been developed to identify shocks in time series models with a minimum of additional assumptions. The dominant framework for this identification, due to Sims (1980), is structural vector autoregressions. The premise of SVARs is that the space of the innovations to a vector of time series variables Y_t—that is, the one step ahead forecast errors of Y_t based on a population projection of Y_t onto its past values—spans the space of the structural shocks. Said differently, in population the econometrician is assumed to be as good at one step ahead forecasting of the economy as an agent who directly observes the structural shocks in real time. The task of identifying the structural shock of interest thus reduces to the task of finding the linear combination of the innovations that is the structural shock. Sims (1980) originally proposed doing this construction using short-run "timing" restrictions. Subsequently, a host of other approaches for identifying structural shocks have been developed, including long-run restrictions based on the cumulative SIRFs, identification by heteroskedasticity, partial identification by sign restrictions on the SIRFs, and most recently by the use of external instruments.

This section has four themes. The first is the quest in the literature for increasingly credible identification schemes. This emphasis on identification parallels the identification revolution in microeconometrics, which stresses the importance of credible restrictions, typically in the form of isolating as-if random variation in the data, to identify a causal effect of interest.

Second, methods that identify a unique SIRF of interest (that is, identification schemes in which the SIRF is point identified) have natural interpretations in terms of instrumental variables or generalized method of moments (GMM) regression.

Third, we stress the importance of the choice of normalization of the shocks and make the case for what we call the unit effect normalization, which is different than the prevalent normalization that sets the shock variance to one. Although this normalization choice does not matter in population, it *does* matter in sample, and we argue that the unit effect normalization is the most natural in most applications. Moreover, the unit shock normalization makes the extension of SVAR methods to structural DFMs straightforward.

The fourth theme ties the previous three together: this quest for credible identification can push a research design to focus on exogenous movements that explain only a small fraction of the variation in the data, which in turn can affect inference. In the point-identified settings, we cast this potential pitfall in terms of weak instruments or weak identification. In the set-identified settings (eg, identification of SVARs by sign restrictions), these issues arise in the form of sensitivity of inference to Bayesian prior distributions, even if those priors are intended to be, in some sense, uninformative.

The focus of this section is explicating the normalization, identification schemes, and issues raised by weak identification. We provide references to, but spend little time on, conventional methods of inference, which is typically done using bootstrap methods (Kilian, 1998, 2001) or by computing a Bayesian posterior distribution (Sims and Zha, 1998, 1999). For textbook treatments of VARs, conventional asymptotics, and conventional inference, see Lütkepohl (2015). Kilian (2015) and Bjørnland and Thorsrud (2015b) provide complementary summaries of SVAR methods, with more details and examples than are given here but without the focus on our four themes.

This section is written to complement the chapter by Ramey (2016, this Handbook); while the broad coverage of material is similar, this section focuses more on methods and econometric issues, while Ramey's chapter focuses more on applications and assessing identification in practice.

Section 4.1 lays out the SVAR notation and assumptions, including the normalization condition in Section 4.1.3. Various methods for identifying the SIRFs are discussed in Sections 4.2–4.7.

4.1 Structural Vector Autoregressions

SVAR analysis undertakes to identify the structural impulse responses of observable variables to one or more shocks, which are linear combinations of the VAR innovations.

4.1.1 VARs, SVARs, and the Shock Identification Problem

4.1.1.1 The VAR

Let Y_t be a $n \times 1$ vector of stationary time series, assumed for convenience to have mean zero. A pth order VAR model represents Y_t as a linear function of its first p lagged values plus a serially uncorrelated disturbance η_t. This disturbance η_t, which is referred to as the innovation in Y_t, has conditional mean zero given past Y; thus η_t is the population one step ahead forecast error under squared-error loss. That is, the VAR(p) model of Y_t is,

$$Y_t = A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + \eta_t \quad \text{or} \quad A(\mathrm{L}) Y_t = \eta_t, \tag{18}$$

where $A(\mathrm{L}) = I - A_1 \mathrm{L} - \cdots - A_p \mathrm{L}^p$ and L is the lag operator, and where the disturbance η_t is a martingale difference sequence with covariance matrix Σ_η, so that η_t is serially uncorrelated.

In practice, Y_t will generally have nonzero mean and the VAR in (18) would include an intercept. The assumption of zero mean and no intercept in the VAR is made without loss of generality to simplify notation.

The VAR (18) is called the *reduced-form VAR*. The ith equation in (18) is the population regression of Y_{it} onto lagged values of Y_t. Because (18) is the population regression of Y_t onto its lags, its parameters $A(\mathrm{L})$ and Σ_η are identified.

The *innovation* in Y_{it} is the one step ahead forecast error, η_{it}, in the ith equation in (18).

The *vector moving average representation* of Y_t, which in general will be infinite order, expresses Y_t in terms of current and past values of the innovations:

$$Y_t = C(\mathrm{L})\eta_t, \quad \text{where } C(\mathrm{L}) = I + C_1\mathrm{L} + C_2\mathrm{L}^2 + \cdots = A(\mathrm{L})^{-1}. \tag{19}$$

4.1.1.2 The SVAR

A structural VAR model represents Y_t not in terms of its innovations η_t, but rather in terms of a vector of underlying *structural shocks* ε_t, where these structural shocks represent unexpected exogenous disturbances to structural economic relationships such as production functions (productivity shocks), central bank reaction functions (monetary policy shocks), or oil supply functions (oil supply shocks).[o] The SVAR assumes that the innovations are a linear combination of the unobserved structural shocks:

$$\eta_t = H\varepsilon_t. \tag{20}$$

The structural shocks are assumed to be uncorrelated[p]:

$$E\varepsilon_t\varepsilon_t' = \Sigma_\varepsilon = \begin{pmatrix} \sigma^2_{\varepsilon_1} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{\varepsilon_n} \end{pmatrix}. \tag{21}$$

Substituting (20) into (18) and (19) delivers the structural VAR and the structural moving average representation of the observable variables in terms of the structural shocks:

$$A(\mathrm{L})Y_t = H\varepsilon_t \text{ or } B(\mathrm{L})Y_t = \varepsilon_t, \quad \text{where } B(\mathrm{L}) = H^{-1}A(\mathrm{L}) \text{ (Structural VAR)} \tag{22}$$

$$Y_t = D(\mathrm{L})\varepsilon_t, \quad \text{where } D(\mathrm{L}) = C(\mathrm{L})H, \text{ (Structural MA)} \tag{23}$$

where the second expression in (22) holds if H^{-1} exists.

4.1.1.3 The SVAR Identification Problem

Because A(L) and Σ_η are identified from the projection of Y_t onto its past, the parameters of the structural VAR (22) and the structural MA (23) are identified if H and Σ_ε are identified. The problem of identifying H and Σ_ε is known as the SVAR identification problem. Strictly speaking, the concept of identification refers to nonrandom parameters or functions, but because D(L) is the projection of Y_t onto current and past shocks, the SVAR identification problem is also called the problem of identifying the structural shocks.

[o] Ramey (2016) characterizes structural shocks as having three characteristics: (1) they are exogenous and unforecastable, (2) they are uncorrelated with other shocks, and (3) they represent either unanticipated movements in exogenous variables or news about future movements in exogenous variables.

[p] This assumption that Σ_ε is diagonal is a natural part of the definition of an autonomous structural shock. For example, if one was to posit that two structural shocks were correlated, presumably there would be some structural reason or linkage, but if so then one of the shocks (or both) would be responding to the other endogenously in which case it would not be an exogenous structural shock. See Ramey (2016) for a discussion of this assumption.

4.1.1.4 SIRFs, Historical Decompositions, and Forecast Error Variance Decompositions

The structural MA (23) summarizes the dynamic causal effect of the shocks on current and future Y_t, and it directly delivers two key objects in SVAR analysis: the SIRF and the decomposition of Y_t into structural shocks. With the additional assumption (21) that the structural shocks are uncorrelated, the structural moving average representation also delivers the structural forecast error variance decomposition (FEVD).

The SIRF is the time path of the dynamic causal effect on variable Y_{it} of a unit increase in ε_{jt} at date 0. Let D_h denote the hth lag matrix of coefficients in $D(\mathrm{L})$. Then $D_{h,ij}$ is the causal effect on the ith variable of a unit increase in the jth shock after h periods, that is, $D_{h,ij}$ is the effect on Y_{it+h} of a unit increase in ε_{jt}. Thus the *structural impulse response function* ($SIRF_{ij}$) is the sequence of structural MA coefficients,

$$SIRF_{ij} = \left\{D_{h,ij}\right\}, \quad h = 0, 1, \ldots, \text{ where } D_h = C_h H, \tag{24}$$

where from (19) $C(\mathrm{L}) = A(\mathrm{L})^{-1}$. The contemporaneous effect, D_0, is called the impact effect; note that $D_0 = H$ because $C_0 = I$.

The *cumulative structural impulse response function* is the cumulative dynamic causal effect on Y_t of a unit shock at date 0. Expressed in terms of $D(\mathrm{L})$, the cumulative SIRF on variable i of shock j after h periods is $\sum_{k=0}^{h} D_{k,ij}$.

Because $D(\mathrm{L})\varepsilon_t$ is a linear function of current and lagged values of ε_t, (23) is the historical decomposition of the path of Y_t into the distinct contributions of each of the structural shocks; given $D(\mathrm{L})$, this decomposition is unique.

The FEVD$_{h,ij}$ measures how important the jth shock is in explaining the variation in Y_{it} by computing the relative contribution of that shock to the variance of the unexpected changes in Y_{it} over h periods, that is, to the variance of its h-step ahead forecast errors. The FEVD is,

$$FEVD_{h,ij} = \frac{\sum_{k=0}^{h} D_{k,ij}^2 \sigma_{\varepsilon_j}^2}{\operatorname{var}\left(Y_{it+h} \middle| Y_t, Y_{t-1}, \ldots\right)} = \frac{\sum_{k=0}^{h} D_{k,ij}^2 \sigma_{\varepsilon_j}^2}{\sum_{j=1}^{n} \sum_{k=0}^{h} D_{k,ij}^2 \sigma_{\varepsilon_j}^2}, \tag{25}$$

where $D(\mathrm{L}) = A(\mathrm{L})^{-1} H$.

4.1.1.5 System Identification

System identification entails identification of the full matrix H and thus the full matrix $D(\mathrm{L})$ of SIRFs. System identification makes the assumption that the space of innovations spans the space of structural shocks, so that H is invertible:

$$H^{-1} \text{ exists so that } \varepsilon_t = H^{-1} \eta_t. \tag{26}$$

Assumption (26) is equivalent to saying that the system SVAR representation (22) exists.

Eqs. (20) and (21) imply that

$$\Sigma_\eta = H\Sigma_\varepsilon H'. \tag{27}$$

The number of free parameters is $n(n+1)$ (n^2 in H and n in Σ_ε). Because covariance matrices are symmetric, the number of unique equations in $\Sigma_\eta = H\Sigma_\varepsilon H'$ is $n(n+1)/2$. Thus identification of H and Σ_ε requires $n(n+1)/2$ additional assumptions. Of these, n are obtained from normalizing the scale of the shocks, leaving $n(n-1)/2$ additional restrictions for identification of H.

When the shocks are i.i.d. Gaussian, the restrictions (27) are the only ones available for identification. If the shocks are not Gaussian then additional restrictions on higher moments can be available, and some research pursues the use of these restrictions. Typically these restrictions require strong additional assumptions, for example that the shocks are independently distributed (as opposed to simply uncorrelated) and in any event this approach does not enhance identification in the Gaussian case. We do not pursue further identification that exploits non-Gaussianity.

4.1.1.6 Single Shock Identification

In many applications, such as the application to the effect of oil supply shocks in Section 7, interest is in the effect of just one shock. Without loss of generality, let the shock of interest be the first shock, ε_{1t}. In general, the other shocks need not be identified to identify the SIRF for the first shock, and the innovations need not span the shocks other than ε_{1t} to identify the first SIRF. To stress this point, for single shock identification we rewrite (20) as,

$$\eta_t = H\begin{pmatrix}\varepsilon_{1t}\\ \tilde{\eta}_{\bullet t}\end{pmatrix} = [H_1 \quad H_\bullet]\begin{pmatrix}\varepsilon_{1t}\\ \tilde{\eta}_{\bullet t}\end{pmatrix} = \begin{pmatrix}H_{11} & H_{1\bullet}\\ H_{\bullet 1} & H_{\bullet\bullet}\end{pmatrix}\begin{pmatrix}\varepsilon_{1t}\\ \tilde{\eta}_{\bullet t}\end{pmatrix}, \tag{28}$$

where H_1 is the first column of H and $H_\bullet$ denotes the remaining columns and the final expression partitions these columns similarly, and where $\tilde{\eta}_{\bullet t}$ spans the space of η_t orthogonal to ε_{1t}. Because these other shocks are uncorrelated with $\varepsilon_{1t}, \mathrm{cov}(\varepsilon_{1t}, \tilde{\eta}_{\bullet t}) = 0$.

In single shock identification, the aim is to identify H_1. Given H_1, the structural moving average representation (23) can be written,

$$Y_t = C(\mathrm{L})\eta_t = C(\mathrm{L})H_1\varepsilon_{1t} + C(\mathrm{L})H_\bullet\tilde{\eta}_{\bullet t}, \quad \text{where } \mathrm{cov}(\varepsilon_{1t}, \tilde{\eta}_{\bullet t}) = 0. \tag{29}$$

Evidently, the SIRF for shock 1 is $C(\mathrm{L})H_1$ and the historical contribution of shock 1 to Y_t is $C(\mathrm{L})H_1\varepsilon_{1t}$.

If H in (28) is invertible, then ε_{1t} can be obtained as a linear combination of η_t. Denote the first row of H^{-1} by H^1. It follows from the partitioned inverse formula and the assumption (21) that the shocks are mutually uncorrelated that if H_1 is identified, then H^1 is identified up to scale. In turn, knowing H^1 up to scale allows construction of the shock ε_{1t} up to scale:

$$\varepsilon_{1t} = H^1\eta_t \propto \left[1 \quad \tilde{H}^{1\bullet}\right]\eta_t, \tag{30}$$

where $\widetilde{H}^{1\bullet}$ is a function of H_1 and Σ_η.[q] Thus identification of H_1 permits the construction of ε_{1t} up to scale. An implication of (30) is that identification of H_1 and identification of the shock are interchangeable.[r]

Note that (30) obtains without the additional assumption that the innovations span all the shocks or, for that matter, that they span any shock other than ε_{1t}.

4.1.2 Invertibility

The structural MA representation $Y_t = D(\mathrm{L})\varepsilon_t$ represents Y_t in terms of current and past values of the structural shocks ε_t. The moving average is said to be *invertible* if ε_t can be expressed as a distributed lag of current and past values of the observed data Y_t. SVARs typically assume $\varepsilon_t = H^{-1}\eta_t = H^{-1}A(\mathrm{L})Y_t$, so an SVAR typically imposes invertibility.[s] Yet, an economic model may give rise to a structural moving average process that is not invertible. If so the VAR innovations will not span the sapce of the structural shocks. Because identification of the shocks and identification of the SIRF are equivalent, if the true SIRF is not invertible, a SVAR constructed from the VAR innovations will not recover the true SIRF.

[q] Use the partitioning notation for H in the final expression in (28) and the partitioned matrix inverse formula to write, $H^1 = \left[H^{11} \quad -H^{11}H_{1\bullet}H_{\bullet\bullet}^{-1}\right] \propto \left[1 \quad -H_{1\bullet}H_{\bullet\bullet}^{-1}\right]$, where H^{11} is the scalar, $H^{11} = \left(H_{11} - H_{1\bullet}H_{\bullet\bullet}^{-1}H'_{\bullet 1}\right)^{-1}$. Because the goal is to identify ε_{1t} up to scale, the scale of ε_{1t} is arbitrary, so for convenience we adopt the normalization that $\Sigma_\varepsilon = I$; this is the unit standard deviation normalization of Section 4.1.3 and is made without loss of generality. Then (27) implies that $\Sigma_\eta = HH'$. Adopt partitioning notation for Σ_η conformable with that of H in (28). Then $\Sigma_\eta = HH'$ implies that $\Sigma_{\eta,1\bullet} = H_{11}H'_{\bullet 1} + H_{1\bullet}H'_{\bullet\bullet}$ and $\Sigma_{\eta,\bullet\bullet} = H_{\bullet 1}H'_{\bullet 1} + H_{\bullet\bullet}H'_{\bullet\bullet}$, which in turn implies $H_{1\bullet}H'_{\bullet\bullet} = \Sigma_{\eta,1\bullet} - H_{11}H'_{\bullet 1}$ and $H_{\bullet\bullet}H'_{\bullet\bullet} = \Sigma_{\eta,\bullet\bullet} - H_{\bullet 1}H'_{\bullet 1}$. Using these final two expressions and the fact that $H_{1\bullet}H'_{\bullet\bullet}\left(H_{\bullet\bullet}H'_{\bullet\bullet}\right)^{-1} = H_{1\bullet}H_{\bullet\bullet}^{-1}$ yields $H_{1\bullet}H_{\bullet\bullet}^{-1} = \left(\Sigma_{\eta,1\bullet} - H_{11}H'_{\bullet 1}\right)\left(\Sigma_{\eta,\bullet\bullet} - H_{\bullet 1}H'_{\bullet 1}\right)^{-1}$. Thus $H^1 \propto \left[1 \quad \widetilde{H}^{1\bullet}\right]$, where $\widetilde{H}^{1\bullet} = -\left(\Sigma_{\eta,1\bullet} - H_{11}H'_{\bullet 1}\right)\left(\Sigma_{\eta,\bullet\bullet} - H_{\bullet 1}H'_{\bullet 1}\right)^{-1}$. Because Σ_η is identified from the reduced form, knowledge of H_1 and the uncorrelated shock assumption therefore determines H^1, and thus the shock ε_{1t}, up to scale.

[r] Here is a second, perhaps more intuitive, method for constructing ε_{1t} from η_t given H_1, the assumption (21) that the shocks are mutually uncorrelated, and the invertibility of H. Let $H_1^\perp$ be any $n \times (n-1)$ matrix with linearly independent columns that are orthogonal to H_1. Then $H_1^{\prime\perp}\eta_t = H_1^{\prime\perp}H\varepsilon_t = H_1^{\prime\perp}[H_1 \quad H_\bullet]\varepsilon_t = \left[0 \quad H_1^{\prime\perp}H_\bullet\right]\varepsilon_t = H_1^{\prime\perp}H_\bullet\varepsilon_{\bullet t}$. If H is invertible, then $H_1^{\prime\perp}H_\bullet$ is invertible, so $\varepsilon_{\bullet t} = \left(H_1^{\prime\perp}H_\bullet\right)^{-1}H_1^{\prime\perp}\eta_t$. In addition, $H'_1\eta_t = H'_1H\varepsilon_t = H'_1H_1\varepsilon_{1t} + H'_1H_\bullet\varepsilon_{\bullet t}$. Because ε_{1t} and $\varepsilon_{\bullet t}$ are uncorrelated, $H'_1\eta_t - \mathrm{Proj}\left(H'_1\eta_t|\varepsilon_{\bullet t}\right) = H'_1H_1\varepsilon_{1t}$, where $\mathrm{Proj}(X|Y)$ is the population projection of X on Y. Because $\varepsilon_{\bullet t} = \left(H_1^{\prime\perp}H_\bullet\right)^{-1}H_1^{\prime\perp}\eta_t$, $\varepsilon_{1t} = \left(H'_1H_1\right)^{-1}\left[H'_1\eta_t - \mathrm{Proj}\left(H'_1\eta_t|\varepsilon_{\bullet t}\right)\right] = \left(H'_1H_1\right)^{-1}\left[H'_1\eta_t - \mathrm{Proj}\left(H'_1\eta_t|H_1^{\prime\perp}\eta_t\right)\right]$; this is an alternative representation of the linear combination of η_t given by $H^1\eta_t$ in (30).

[s] In linear filtering theory, a time series representation is called *fundamental* if the disturbances are a function of current and past values of the observable data. Accordingly, the invertibility assumption is also referred to as the assumption that the structural shocks are fundamental.

There are at least three reasons why the structural moving average might not be invertible. One is that there are too few variables in the VAR. For example, suppose that there are four shocks of interest (monetary policy, productivity, demand, oil supply) but only three variables (interest rates, GDP, the oil price) in the VAR. It is impossible to reconstruct the four shocks from current and lagged values of the three observed time series, so the structural moving average process is not invertible. Estimates from a SVAR constructed from the VAR innovations will therefore suffer from a form of omitted variable bias.

Second, some elements of Y may be measured with error, which effectively adds more shocks (the measurement error) to the model. Again, this makes it impossible to reconstruct the structural shocks from current and lagged values of Y. This source of noninvertibility can be thought of as errors-in-variables bias.

Third, noninvertibility can arise when shocks contain news about the future. To see the mechanics of the problem, consider the first-order moving average univariate model with a single lag: $Y_t = \varepsilon_t - d\varepsilon_{t-1}$. Solving for ε_t as a function of current and lagged values of Y_t yields $\varepsilon_t = \sum_{i=0}^{h-1} d^i Y_{t-i} + d^h \varepsilon_{t-h}$. If $|d| < 1$, then $d^h \approx 0$ for h large and $E\left(\varepsilon_t - \sum_{i=0}^{h-1} d^i Y_{t-i}\right)^2 \rightarrow 0$ as $h \rightarrow \infty$, so that ε_t can be recovered from current and lagged values of y and the process is invertible. In contrast, when $|d| > 1$, the initial value of ε_0 remains important, so the process is not invertible. In this case, however, ε_t can be recovered from current and future values of y_t: solving the moving average process forward yields the representation, $\varepsilon_t = -(1/d)\sum_{i=1}^{h}(1/d)^i Y_{t+i} + (1/d)^h \varepsilon_{t+h}$, where $E\left((1/d)^h \varepsilon_{t+h}\right)^2 \rightarrow 0$ when $|d| > 1$. In economic models, noninvertibility can arise, for example, because technological innovations (shocks) may have small initial effects on productivity and much larger effects on future productivity, so a technology shock today (an invention today) is actually observed in the data as a productivity increase in the future. As a second example, if the central bank announces that it will raise interest rates next month, the monetary policy shock occurs today but is not be observed in the overnight rate until next month. Like the case of omitted variables, news shocks are an example of economic agents knowing more about shocks than the econometrician can decipher from current and past data.

Unfortunately, statistics based on the second moments of the data—which include the parameters of the SVAR—cannot determine whether the true SIRF is invertible or not: each noninvertible moving average representation has an invertible moving average representation that is observationally equivalent based on the second moments of the data. To see this, consider the univariate first-order moving average example of the previous paragraph, $y_t = \varepsilon_t - d\varepsilon_{t-1}$. By direct calculation, $\text{var}(y_t) = \left(1 + d^2\right)\sigma_\varepsilon^2$, $\text{cov}(y_t, y_{t-1}) = -d\sigma_\varepsilon^2$, and $\text{cov}(y_t, y_{t-i}) = 0$, $|i| > 1$. It is readily verified that for any set of parameter values

$(d, \sigma_\varepsilon^2)$ with $|d| < 1$, the alternative parameter values $\left(\tilde{d}, \tilde{\sigma}_\varepsilon^2\right) = \left(d^{-1}, d^2\sigma_\varepsilon^2\right)$ produce the same autocovariances; that is, $(d, \sigma_\varepsilon^2)$ and $\left(d^{-1}, d^2\sigma_\varepsilon^2\right)$ are observationally equivalent values of the parameters based on the second moments of the data. If the data are Gaussian, then these two sets of parameter values are observationally equivalent based on the likelihood. Because these pairs have the same autocovariances, they produce the same reduced-form VAR, but they imply different SIRFs.

Noninvertibility is an important threat to the validity of SVAR analysis. Hansen and Sargent (1991) provide an early and important discussion, Sargent (1987) provides an illuminating example using the permanent income model of consumption, and Fernández-Villaverde et al. (2007) discuss the restrictions on linear economic models that give rise to invertibility. For more detailed discussion of the literature and references, see Forni et al. (2009), Leeper et al. (2013), Plagborg-Møller (2015), and Ramey (2016, this Handbook). As Forni et al. (2009) point out and as discussed in more detail in Section 5, SDFMs can resolve the problems of measurement error, omitted variables, and in some cases timing (news) through the use of large numbers of series.

4.1.3 Unit Effect Normalization

Because the structural shocks are unobserved, their sign and scale are arbitrary and must be normalized. There are two normalizations commonly used, the unit standard deviation normalization and the unit effect normalization.

The unit *standard deviation normalization* makes each shock have unit variance:

$$\Sigma_\varepsilon = I \quad \text{(unit standard deviation normalization)}. \tag{31}$$

The normalization (31) fixes the units of the shock, but not its sign. The sign must be fixed separately, for example by defining a positive monetary shock to increase the target rate on impact.

The *unit effect normalization* fixes the sign and scale of the jth shock so that a unit increase in ε_{jt} induces a contemporaneous unit increase in a specific observed variable, which we take to be Y_{jt}. Written in terms of the H matrix, the unit effect normalization sets

$$H_{jj} = 1 \quad \text{(unit effect normalization)}. \tag{32}$$

Equivalently, under the unit effect normalization a unit increase in ε_{jt} increases η_{jt} by one unit, which in turn increases Y_{jt} by one unit. For example, if the Federal Funds rate is measured in percentage points, then a unit monetary shock induces a one percentage point increase in the Federal Funds rate. A unit shock to productivity growth increases the growth rate of productivity by one percentage point, and so forth.

For system identification, both normalizations provide n additional restrictions on H, so that $n(n-1)/2$ additional restrictions are needed.

For single shock identification, both normalizations set the scale of ε_{1t}. Under the unit standard deviation assumption, $\sigma_{\varepsilon_1}^2 = 1$. Under the unit effect normalization,

$$H_1 = \begin{pmatrix} 1 \\ H_{1\bullet} \end{pmatrix}. \tag{33}$$

In both cases, $n-1$ additional restrictions are needed to identify H_1.

In population, these two normalizations are interchangeable. Nevertheless, the unit effect normalization is preferable for three reasons.

First, the unit effect normalization is in the units needed for policy analysis or real-world interpretation. A monetary policy maker needs to know the effect of a 25 basis point increase in the policy rate; providing the answer in standard deviation units does not fulfill that need. When oil prices fall by, say, 10%, because of an oil supply shock, the question is what the effect of that fall is on the economy; again, stating the SIRFs in standard deviation units does not answer that question.

Second, although the two formulations are equivalent in population, statistical inference about the SIRFs differs under the two normalizations. In particular, it is an inferential error to compute confidence intervals for SIRFs under the unit standard deviation normalization, then renormalize those bands so that they answer the questions relevant to policymakers. The inferential error is that this renormalization entails dividing by an estimator of H_{11}, which introduces additional sampling uncertainty. If, under the unit standard deviation normalization, H_{11} is close to zero, then this sampling variability can be considerable and renormalization introduces inference problems related to weak instruments.[t]

Third, as discussed in the next section, the unit effect normalization allows SVAR identification schemes to be extended directly to SDFMs.

For these reasons, we adopt the unit effect normalization throughout this chapter.

Finally, we note that the unit effect normalization could alternatively involve the normalization that shock j induces a unit increase in variable i. In this case, the normalization for shock j would be $H_{ij} = 1$ instead of $H_{jj} = 1$ as in (32). If each shock has a unit impact on a different VAR innovation, the distinction we are making here is trivial because the named shocks can always be ordered to align with the order of the variables in the VAR. For example, without loss of generality the Fed funds rate can be listed first, the monetary policy shock can be taken to be the first shock, and $H_{11} = 1$ is the unit effect normalization.

[t] Another way to state this problem is in the context of bootstrap draws of the IRFs. If the bootstrap uses the unit standard deviation normalization to compute confidence intervals, then multiplies the confidence intervals by a scaling coefficient which converts from standard deviation to native units, the resulting IRF confidence intervals do not incorporate the sampling uncertainty of that scaling coefficient. In contrast, if the bootstrap does that conversion for every draw, which is equivalent to using the unit effect normalization, then the IRF confidence intervals do incorporate the sampling uncertainty of the unit conversion step.

This distinction, however, becomes nontrivial when two distinct shocks are normalized to have unit effects on the same variable. For example, suppose one was interested in investigating the separate effects of an oil supply shock (ε_{1t}, say) and an oil inventory demand shock (ε_{2t}, say), and for the purpose of the investigation it was useful to fix the scales of the two shocks so that they each produced a one percentage point increase in the price of oil. Without loss of generality, let the oil price be the first variable so η_{1t} is the innovation in the oil price. Then this alternative unit effect normalization would be that $H_{11}=1$ and $H_{12}=1$. If the results will be presented using this normalization, then adopting this normalization from the outset ensures that confidence intervals will correctly incorporate the data-dependent transformations to impose this normalization.

Because the circumstance described in the previous paragraph is unusual, throughout this chapter we use the version of the unit effect normalization in (32).

4.1.4 Summary of SVAR Assumptions.

We now collect the assumptions underlying SVAR analysis:

(SVAR-1) The innovations in Y_t, η_t, span the space of the one or more structural shocks:

- **(a)** for system identification, $\eta_t = H\varepsilon_t$ as in (20) and H^{-1} exists and
- **(b)** for single shock identification, (28) holds and H^1 exists.

(SVAR-2) The structural shocks are uncorrelated as in (21).

(SVAR-3) The scale of the shocks is normalized using either the unit standard deviation normalization (31) or the unit effect normalization (32).

With one exception, these assumptions, which were discussed earlier, are needed for all the shock identification schemes discussed in this section. The exception is single shock identification based on direct measurement of the time series of structural shocks, which, because the shock is observed, requires only assumption SVAR-2.

For this chapter, we make the further assumptions:

(SVAR-4) The innovations η_t are the one step ahead forecast errors from the VAR(p) (18) with time-invariant parameters A(L) and Σ_η.

(SVAR-5) The VAR lag polynomial A(L) is invertible.

Assumptions SVAR-4 and SVAR-5 are technical assumptions made for convenience. For example, SVAR-4 can be relaxed to allow for breaks, or time variation can be introduced into the VAR parameters using the methods of, for example, Cogley and Sargent (2005) or Sims and Zha (2006). Assumption SVAR-5 presumes that the variables have been transformed to stationarity, typically using first differences or error correction terms. Alternatively the series could be modeled in levels in which case the SIRF would have the interpretation of a cumulative SIRF. Levels specifications are used in much of the literature. These relaxations of SVAR-4 and SVAR-5 do not materially affect any of the subsequent discussion and they are made here to streamline the discussion.

4.2 Contemporaneous (Short-Run) Restrictions

Contemporaneous restrictions rest on timing arguments about the effect of a given shock on a given variable within the period (monthly if monthly data, etc.). Typically these are zero restrictions, indicating that shock ε_{jt} does not affect Y_{it} (equivalently, does not affect η_{it}) within a period because of some sluggish or institutional feature of Y_{it}. These contemporaneous timing restrictions can identify all the shocks, or just some shocks.

4.2.1 System Identification

Sims's (1980) original suggestion for identifying the structural shocks was of this form, specifically he adopted an ordering for the variables in which the first innovation responds only to the first shock within a period, the second innovation responds only to the first and second shocks, etc. Under this recursive scheme, the shocks are simply linear regression residuals, where the first regression only controls for lagged observables, the second regression controls for lags and one contemporaneous variable, etc. For example, in many recursive monetary SVARs, the monetary policy shock is identified as the residual from an Taylor rule-type regression.

This recursive identification scheme is a Wold (1954) causal chain and corresponds to assuming that H is lower triangular. Because $\Sigma_\eta = H\Sigma_\varepsilon H'$, the lower-triangular assumption implies that $H\Sigma_\varepsilon^{1/2} = Chol\left(\Sigma_\eta\right)$, where $Chol$ denotes the Cholesky factorization. With the unit effect normalization, H is obtained as the renormalized Cholesky factorization, that is, $H = Chol\left(\Sigma_\eta\right)\Sigma_\varepsilon^{-1/2}$, where $\Sigma_\varepsilon = diag(\{[Chol(\Sigma_\eta)_{jj}]^2, j = 1, \ldots, n\})$. This lower-triangular assumption remains a common identification assumption used in SVAR empirical applications.

Nonrecursive restrictions also can provide the $n(n-1)/2$ contemporaneous restrictions for system identification. For example, some of the elements of H can be specified by drawing on application-specific information. An early example of this approach is Blanchard and Watson (1986), who used information about automatic stabilizers in the budget to determine the contemporaneous fiscal response to aggregate demand shocks which, along with zero restrictions based on timing arguments, identified H.

Blanchard and Watson (1986) also show how short-run restrictions on the coefficients can be reinterpreted from an instrumental variables perspective.

4.2.2 Single Shock Identification

Identification of a single shock requires fewer restrictions on H; here we give three examples. The first example is to suppose that a given variable (without loss of generality, Y_{1t}) responds within the period only to a single structural shock; if so, then $\varepsilon_{1t} = \eta_{1t}$ and no additional assumptions are needed to identify ε_{1t}. This first example corresponds to ordering the variable first in a Cholesky factorization, and no additional assumptions are needed about the ordering of the remaining variables (or in fact whether the remaining shocks are identifiable).

The second example makes the opposite assumption: that a given shock affects only one variable within a period, and that variable (and innovation) potentially responds to all other shocks as well. This second example corresponds to ordering the variable last in a Cholesky factorization.

The third example is the "Slow-*r*-Fast" identification scheme frequently used to identify monetary policy shocks, see, for example, Christiano, Eichenbaum, and Evans, (1999) and Bernanke et al. (2005). Under this scheme, so-called slow-moving variables Y_t^s such as output and prices do not respond to monetary policy or to movements in asset prices within the period; through monetary policy, the Fed funds rate r_t responds to shocks to the slow-moving variables within a period but not to asset price developments; and fast-moving variables Y_t^f, such as asset prices and expectational variables, respond to all shocks within the period. This delivers the block recursive scheme,

$$\begin{pmatrix} \eta_t^s \\ \eta_t^r \\ \eta_t^f \end{pmatrix} = \begin{pmatrix} H_{ss} & 0 & 0 \\ H_{rs} & H_{rr} & 0 \\ H_{fs} & H_{fr} & H_{ff} \end{pmatrix} \begin{pmatrix} \varepsilon_t^s \\ \varepsilon_t^r \\ \varepsilon_t^f \end{pmatrix} \text{ where } Y_t \text{ is partitioned } \begin{pmatrix} Y_t^s \\ r_t \\ Y_t^f \end{pmatrix}, \tag{34}$$

where H_{ss} is square. Under (34), η_t^s spans the space of ε_t^s, so the monetary policy shock ε_t^r is the residual in the population regression of the Fed funds rate innovation η_t^r on η_t^s. Equivalently, ε_t^r is identified as the residual in the regression of the monetary instrument on current values of slow-moving variables as well as lags of all the variables.

4.3 Long-Run Restrictions

Identification of the shocks, or of a single shock, can also be achieved by imposing restrictions on the long-run effect of a given shock on a given variable (Shapiro and Watson, 1988; Blanchard and Quah, 1989; King, Plosser, Stock, and Watson, 1991). Because Y_t is assumed to be stationary, the cumulative long-run effect of ε_t on future values of Y_t is the sum of the structural MA coefficients $D(1)$, where $D(1) = C(1)H = A(1)^{-1}H$, where $C(1)$ and $A(1)$ are, respectively, the sums of the reduced-form MA and VAR coefficients.

4.3.1 System Identification

Let Ω denote the long-run variance matrix of Y_t, that is, $\Omega = \text{var}\left(\sqrt{n}\bar{Y}\right) = 2\pi$ times the spectral density matrix of Y_t at frequency zero. Then

$$\Omega = A(1)^{-1}\Sigma_\eta A(1)^{-1'} = A(1)^{-1}H\Sigma_\varepsilon H' A(1)^{-1'} = D(1)\Sigma_\varepsilon D(1)'. \tag{35}$$

Imposing $n(n-1)/2$ restrictions on $D(1)$ permits identifying $D(1)$ and, because $A(1)^{-1}H = D(1)$, H is identified by $H = A(1)D(1)$.

A common approach is to adopt identifying assumptions that imply that $D(1)$ is lower triangular. For example, Blanchard and Quah (1989) identify a demand shock as having no long-run effect on the level of output. Let $Y_t =$ (GDP growth, unemployment rate) and let ε_{1t} be an aggregate supply shock and ε_{2t} be an aggregate demand shock. The assumption that ε_{2t} has no long-run effect on the *level* of output is equivalent to saying that its cumulative effect on output *growth* is zero. Thus the long-term effect of ε_{2t} (the demand shock) on Y_{1t} (output growth) is zero, that is, $D_{12}(1) = 0$, so $D(1)$ is lower triangular.

In another influential paper, Gali (1999) used long-run restrictions to identify a technology shock. Specifically, Gali (1999) uses a small aggregate structural model to argue that only the technology shock has a permanent effect on the level of labor productivity. Let $Y_t =$ (labor productivity growth, hours growth), ε_{1t} be a technology shock, and ε_{2t} be a non-technology shock. Gali's (1999) restriction that the nontechnology shock has zero long-run effect on the *level* of labor productivity implies that $D_{12}(1) = 0$, so that $D(1)$ is lower triangular.

Blanchard and Quah (1989), King, Plosser, Stock, and Watson (1991), and Gali (1999) use the unit standard deviation normalization, so that $\Sigma_\varepsilon = I$ and, by (35), $\boldsymbol{\Omega} = D(1)D(1)'$. The lower triangular factorization of $\boldsymbol{\Omega}$ is uniquely the Cholesky factorization, $D(1) = Chol(\boldsymbol{\Omega})$. Using the first expression in (35) and $H = A(1)D(1)$, the combination of the unit standard deviation normalization and the identifying restriction that $D(1)$ is lower triangular provides the closed-form expression for H,

$$H = A(1)Chol\left[A(1)^{-1}\Sigma_\eta A(1)^{-1}\right]. \tag{36}$$

In general, the sample estimate of H can be estimated by substituting sample counterparts for the reduced-form VAR, $\hat{A}(1)$ and $\hat{\Sigma}_{\hat{\eta}}$, for the population matrices, imposing the restrictions on $D(1)$, and solving (35). In the case that $D(1)$ is lower triangular and the unit standard deviation assumption is used, the estimator of H has the closed-form solution which is the sample version of (36).

4.3.2 Single Shock Identification

Long-run restrictions can also identify a single shock. The Blanchard and Quah (1989) and Gali (1999) examples have $n = 2$, but suppose that $n > 2$. Then the assumption that only ε_{1t} affects Y_{1t} in the long run imposes $n - 1$ zero restrictions on the first row of $D(1)$, and implies that ε_{1t} is proportional to $A(1)^1\eta_t$, where $A(1)^1$ is the first row of $A(1)^{-1}$. Thus this assumption identifies ε_{1t} up to scale, and the scale is then set using either the unit effect normalization or the unit standard deviation normalization.

4.3.3 IV Interpretation of Long-Run Restrictions

Shapiro and Watson (1988) provide an instrumental variables interpretation of identification by long-run restrictions. We illustrate this interpretation for a two-variable

VAR(1). Following (22), write the SVAR as $B(\mathrm{L})Y_t = \varepsilon_t$, where $B(\mathrm{L}) = H^{-1}A(\mathrm{L}) = B_0 + B_1\mathrm{L}$, where the final expression assumes the VAR lag length $p = 1$. Add and subtract $B_0\mathrm{L}$ so that $B(\mathrm{L})Y_t = (B_0 + B_1\mathrm{L})Y_t = B_0\Delta Y_t + B(1)Y_{t-1}$, and note that $B_0 = H^{-1}$ so that the SVAR can be written, $H^{-1}\Delta Y_t = -B(1)Y_{t-1} + \varepsilon_t$. Under the unit effect normalization, $H_{11} = H_{22} = 1$ so, using the formula for the inverse of a 2×2 matrix, the SVAR can be written,

$$\begin{aligned} \Delta Y_{1t} &= H_{12}\Delta Y_{2t} - \det(H)B(1)_{11}Y_{1t-1} - \det(H)B(1)_{12}Y_{2t-1} + \det(H)\varepsilon_{1t} \\ \Delta Y_{2t} &= H_{21}\Delta Y_{1t} - \det(H)B(1)_{21}Y_{1t-1} - \det(H)B(1)_{22}Y_{2t-1} + \det(H)\varepsilon_{2t}. \end{aligned} \tag{37}$$

The parameters H_{12} and H_{21} are unidentified without a further restriction on the simultaneous equations model (37), however, long-run restrictions on $D(1)$ provide such a restriction. Specifically, the assumption that $D(1)$ is lower triangular implies that $D(1)^{-1} = B(1)$ is lower triangular, so that $B(1)_{12} = 0$. Thus, the assumption that $D(1)$ is lower triangular implies that Y_{2t-1} is excluded from the first equation of (37), and thus is available as an instrument for ΔY_{2t} to estimate H_{12} in that equation. Because Y_{2t-1} is predetermined, $E(\varepsilon_{1t}Y_{2t-1}) = 0$ so Y_{2t-1} satisfies the exogeneity condition for a valid instrument.

As an example, consider the special case of the VAR(1) (37) where,

$$\begin{aligned} \Delta Y_{1t} &= H_{12}\Delta Y_{2t} + \det(H)\varepsilon_{1t} \\ \Delta Y_{2t} &= H_{21}\Delta Y_{1t} + (\alpha - 1)Y_{2t-1} + \det(H)\varepsilon_{2t}. \end{aligned} \tag{38}$$

Because ΔY_{2t} depends on ΔY_{1t}, (38) is a system of simultaneous equations and neither H_{12} nor H_{21} can be estimated consistently by OLS. However, because Y_{2t-1} does not appear in the first equation, it can be used as an instrument for ΔY_{2t} to estimate H_{12}. The instrumental variables estimator of H_{12} is,

$$\hat{H}_{12} = \frac{\sum_{t=2}^{T}\Delta Y_{1t}Y_{2t-1}}{\sum_{t=2}^{T}\Delta Y_{2t}Y_{2t-1}}. \tag{39}$$

This instrumental variables interpretation is noteworthy for two reasons. First, although standard estimation algorithms for long-run identification, such as the Cholesky factor expression (36), appear to be quite different from instrumental variables, when the system is exactly identified the two estimation approaches are equivalent. Thus, the "equation counting" identification approach to identification is the same as having a valid instrument for ΔY_{2t}.

Second, the IV interpretation links the inference problem under long-run restrictions to the well-studied topic of inference in IV regressions. Here, we focus on one aspect of inference in IV regression which turns out to be relevant for SVARs with long-run restrictions: inference when instruments are weak.

4.3.4 Digression: Inference in IV Regression with Weak Instruments

An instrument in IV regression is said to be weak if its correlation with the included endogenous regressor is small. Although a detailed discussion of weak instruments and weak identification is beyond the scope of this chapter, it is useful to lay out the central ideas here because they also arise in other SVAR identification schemes. For this digression only, we modify notation slightly to align with the standard regression model. With this temporary notation, the IV regression model is,

$$\begin{aligned} Y_{1t} &= \beta Y_{2t} + u_t \\ Y_{2t} &= \pi' Z_t + V_t \end{aligned} \tag{40}$$

where Y_{2t} is the single included endogenous variable, β is the coefficient of interest, and the second equation in (40) is the first-stage equation relating the included endogenous variable to the vector of k instruments, Z_t. The instruments are assumed to be exogenous in the sense that $E(Z_t u_t) = 0$. When there is a single instrument, the IV estimator is

$$\hat{\beta}_{IV} = \frac{\sum_{t=1}^{T} Y_{1t} Z_t}{\sum_{t=1}^{T} Y_{2t} Z_t}. \tag{41}$$

With multiple instruments, there are multiple estimators available, such as two-stage least squares.

The weak-instrument problem arises when the included endogenous variable Y_{2t} is weakly correlated with Z_t or, equivalently, when π in (40) is small. In this case, the sample covariance in the denominator of (41) can have a mean sufficiently close to zero that, in some samples, the denominator itself could be close to zero or even have a sign different from the population covariance. When the sampling distribution of the denominator includes small values, the result is bias in the IV estimator, heavy tails in its distribution, and substantial departures from normality of its associated t-statistic. These features are general and arise in time series, panel, and cross-sectional regression, with multiple instruments, multiple included endogenous regressors, and in GMM estimation (eg, Nelson and Startz, 1990a,b; Staiger and Stock, 1997; Stock and Wright, 2000).

In linear IV regression, the primary measure of strength of an instrument is the so-called concentration parameter, divided by the number of instruments. The concentration parameter is defined in the classical linear instrumental variables model with homoscedasticity and i.i.d. observations. The concentration parameter is $\mu^2 = \pi' Z' Z \pi / \sigma_v^2$, where σ_v^2 is the variance of the first-stage error. The quantity μ^2/k is the noncentrality parameter of the F-statistic testing the coefficient on the instrument in the first-stage regression. One rule of thumb is that weak-instrument problems are an important concern when this first-stage F-statistic is less than 10 (Staiger and Stock, 1997).[u]

4.3.5 Inference Under Long-Run Restrictions and Weak Instruments

A number of studies have pointed out that SVAR inference based on long-run restrictions can be delicate to seemingly minor changes, such as different sample periods or different number of VAR lags. In addition, in Monte Carlo simulations, IRFs based on long-run restrictions have been found to be biased and/or have confidence intervals that do not have the desired coverage probability; see, for example, Christiano et al. (2006). One interpretation of these problems, as put forth by Faust and Leeper (1997), is that they arise because it is difficult to estimate the long-run variance $\mathbf{\Omega}$, which entails estimating $A(1)^{-1}$. In our view, however, this interpretation, while not incorrect, is less useful than posing the problem in terms of the IV framework earlier. Viewing the problem as weak identification both explains the pathologies of the sampling distribution and points the way toward inference procedures that are robust to these problems.

We therefore focus on the IV interpretation of identification by long-run restrictions and weak-instrument issues, initially raised by Sarte (1997), Pagan and Robertson (1998), and Watson (2006). We focus on the special case (38) and the IV estimator (39), however as shown by these authors these comments apply generally to inference using long-run restrictions.

Comparison of the SVAR example (38) and (39) to the IV model and estimator (40) and (41) indicates that the instrument Y_{2t} will be weak when α is sufficiently close to one. Consider the special case $H_{21}=0$, so that the second equation in (38) is the first stage and the first-stage coefficient is $\alpha-1$. A direct calculation in this case shows that the concentration parameter is $T(\alpha-1)^2/(1-\alpha^2)$. For $T=100$, the concentration parameter is 5.3 for $\alpha=0.9$ and is 2.6 for $\alpha=0.95$. These are small concentration parameters, well below the rule-of-thumb cutoff of 10.

Gospodinov (2010) provides a more complete treatment of the distribution theory when the excluded variable is persistent and shows that in general standard inferences will be misleading when the instrument is weak (estimated IRFs are biased, confidence intervals do not have the advertised coverage rates).

Because the weak-instrument problems arise when roots are large, standard methods for inference in the presence of weak instruments under stationarity (eg, Stock and Wright, 2000) no longer apply directly. Chevillon et al. (2015) develop a method for constructing

[u] There is now a very large literature on weak instruments in IV regression and weak identification in generalized method of moments estimation. Andrews and Stock (2007) survey the early econometrics literature on weak instruments. For a recent survey of weak instruments in the context of estimation of the New Keynesian Phillips curve, see Mavroeidis et al. (2014).

confidence sets in this application that is robust to this weak-instruments problem, and they find that using weak-instruments procedures change conclusions in some classic long-run identification SVAR papers, including Blanchard and Quah (1989).

4.4 Direct Measurement of the Shock

Measuring ε_{1t} through direct observation solves the identification problem, and some papers undertake to do so.

One approach to direct measurement of shocks uses narrative sources to determine exogenous policy changes. This method was developed by Romer and Romer (1989) for the measurement of monetary policy shocks, and the same authors have used this approach to measure tax, financial distress, and monetary policy shocks (Romer and Romer, 2004, 2010, 2015). For example, Romer and Romer (2010) use textual data including presidential speeches and congressional reports to construct a series of exogenous tax changes. Ramey and Shapiro (1998) and Ramey (2011) use related methods to measure government spending shocks.

A series of papers take this approach to measuring monetary policy shocks by exploiting the expectations hypothesis of the term structure and/or high-frequency financial data. Early contributions include Rudebusch (1998), Kuttner (2001), Cochrane and Piazzesi (2002), Faust et al. (2003, 2004), Gürkaynak et al. (2005), and Bernanke and Kuttner (2005), and recent contributions (with references) are Campbell et al. (2012), Hanson and Stein (2015), and Nakamura and Steinsson (2015). For example, Kuttner (2001) measures the monetary policy shock as the change in the Fed Funds futures rate on the day that the Federal Open Market Committee (FOMC) announces a target rate change. Under the expectations hypothesis, any expected change in the target rate will be incorporated into the preannouncement rate, so the change in the Fed Funds futures rate on the announcement date measures its unexpected movement. Cochrane and Piazzesi (2002) take a similar approach, using changes in the Eurodollar rate around FOMC target change announcements. Upon aggregation to the monthly level, this yields a series of monetary policy shocks, which they use as a regressor to estimate SIRFs.

Another set of applications of this method is to the direct measurement of oil supply shocks. Hamilton (2003) and Kilian (2008a) develop an historical chronology of OPEC oil supply disruptions based on exogenous political events to construct numerical estimates of exogenous oil production shortfalls, that is, exogenous shocks to oil supply.

The approach of directly measuring shocks is ambitious and creative and often delivers new insights. This approach, however, has two challenges. The first is that there are inevitable questions about whether the constructed series measures only the exogenous shock of interest. For example, short-term interest rates can change at announcement dates because of an exogenous monetary shock resulting in a change in a target rate, or because the change in the target rate revealed inside knowledge that the Fed might have about the economy (that is, about the values of other shocks). Additionally, if the window around the announcement is too wide, then rate changes can reflect influences other than the monetary shock (Nakamura and Steinsson, 2015).

The second challenge is that these constructed shocks rarely measure the entirety of the structural shock. For example, some of the monetary shock could be revealed in speeches by Federal Reserve officials in the weeks leading up to a FOMC meeting, so that the change in short rates before and after the FOMC meeting understates the full shock. Whether this omission leads to bias in the estimator of the effect of the monetary policy shock depends on whether the measured shock is correlated with the unmeasured shock. If the measured and unmeasured components are correlated, then this measurement error produces bias in the SIRF estimated using the constructed shock.

The first of these problems, exogeneity, is intrinsic to the research design and does not have a econometric resolution. The second of these problems, errors-in-variables bias, can be solved using econometric methods, in particular by using the measured shock series as an external instrument as discussed in Section 4.7.

4.5 Identification by Heteroskedasticity

Identification can also be achieved by assuming that the H matrix remains fixed but the structural shocks are heteroskedastic. This heteroskedasticity can take the form of different heteroskedasticity regimes, or conditional heteroskedasticity.

4.5.1 Identification by Heteroskedasticity: Regimes

Rigobon (2003) and Rigobon and Sack (2003, 2004) showed that H can be identified by assuming it is constant across regimes in which the variance of the structural shocks change.

Suppose that H is constant over the full sample, but there are two variance regimes, one in which the structural shocks have diagonal variance matrix Σ_ε^1 and a second with diagonal variance matrix Σ_ε^2. Because $\eta_t = H\varepsilon_t$ in both regimes, the variance matrices of η_t in the two regimes, Σ_η^1 and Σ_η^2 satisfy,

$$\begin{aligned}\Sigma_\eta^1 &= H\Sigma_\varepsilon^1 H' \\ \Sigma_\eta^2 &= H\Sigma_\varepsilon^2 H'\end{aligned} \tag{42}$$

The first matrix equation in (42) (the first regime) delivers $n(n+1)/2$ distinct equations, as does the second, for a total of n^2+n equations. Under the unit effect normalization that the diagonal elements of H are 1, H has n^2-n unknown elements, and there are an additional $2n$ unknown diagonal elements of Σ_ε^1 and Σ_ε^2, for a total of n^2+n unknowns. Thus the number of equations equals the number of unknowns.

For these equations to solve uniquely for the unknown parameters, they must provide independent information (satisfy a "rank" condition). For example, proportional heteroskedasticity $\Sigma_\varepsilon^2 = a\Sigma_\varepsilon^1$ provides no additional information because then $\Sigma_\eta^2 = a\Sigma_\eta^1$ and the equations from the second regime are the same as those from the first regime. In practice, it is difficult to check the "rank" condition because Σ_η^1 and Σ_η^2 must be estimated. For example, in the previous example $\Sigma_\eta^2 = a\Sigma_\eta^1$ in population, but the sample estimates of Σ_η^1 and Σ_η^2 would not be proportional because of sampling variability.

Economic reasoning or case-specific knowledge is used in identification by heteroskedasticity in one and, in some applications, two places. The first is to make the case that H does not vary across heteroskedasticity regimes, that is, that H is time-invariant even though the variances of the structural shocks are time varying. The second arises when some of the shocks are not naturally associated with a specific observable variable. For example, Rigobon (2003) works through a bivariate example of supply and demand in which the variance of the supply disturbance is posited to increase, relative to the variance of the demand disturbance, and he shows that this increase identifies the slope of the demand curve, however this identification requires a priori knowledge about the nature of the change in the relative shock variances. Similarly, Rigobon and Sack (2004) and Wright (2012) exploit the institutional fact that monetary policy shocks arguably have a much larger variance at announcement dates than otherwise, while plausibly their effect (H_1) is the same on announcement dates and otherwise. This heteroskedasticity around announcement dates provides a variant of the approach discussed in Section 4.3 in which the shock itself is measured as changes in some market rate around the announcement.

For additional references and discussion of regime-shift heteroskedasticity, see Lütkepohl and Netšunajev (2015) and Kilian (2015).

4.5.2 Identification by Heteroskedasticity: Conditional Heteroskedasticity

The idea of identification by conditional heteroskedasticity is similar to that of identification by regime-shift heteroskedasticity. Suppose that the structural shocks are conditionally heteroskedastic but H is constant. Then $\eta_t = H\varepsilon_t$ implies the conditional moment matching equations,

$$E\left(\eta_t\eta_t' | Y_{t-1}, Y_{t-2}, \ldots\right) = HE\left(\varepsilon_t\varepsilon_t' | Y_{t-1}, Y_{t-2}, \ldots\right)H'. \tag{43}$$

The conditional covariance matrix of ε_t is diagonal. If those variances evolve according to a GARCH process, then they imply a conditionally heteroskedastic process for η_t. Sentana and Fiorentini (2001) and Normandin and Phaneuf (2004) show that a GARCH process for ε_t combined with (43) can identify H. Lanne et al. (2010) extend this reasoning from GARCH models to Markov switching models. These are similar to the regime-shift model in Section 4.5.1, however the regime-shift indicator is latent; see Hamilton (2016). For further discussion, see Lütkepohl and Netšunajev (2015).

4.5.3 Instrumental Variables Interpretation and Potential Weak Identification

As pointed out by Rigobon (2003) and Rigobon and Sack (2003), identification by heteroskedasticity regimes has an instrumental variables interpretation, and this interpretation illustrates the potential inference challenges when the change in the variance provides only limited identification power either because the change is small, or because there are few observations in one of the regimes.

To illustrate the instrumental variables interpretation of identification by heteroskedasticity, let $n=2$ and suppose that the variance of the first shock varies between the two regimes while the variance of the other shock does not. This is the assumption used by Rigobon and Sack (2004) and Wright (2012) with high-frequency data, in which the variance of the monetary policy shock (ε_{1t}) is elevated around FOMC announcement dates while the variance of the other shocks does not change around announcement dates. Then under the unit effect normalization, (42) becomes,

$$\begin{pmatrix} \Sigma^j_{\eta_1\eta_1} & \Sigma^j_{\eta_1\eta_2} \\ \Sigma^j_{\eta_2\eta_1} & \Sigma^j_{\eta_2\eta_2} \end{pmatrix} = \begin{pmatrix} 1 & H_{12} \\ H_{21} & 1 \end{pmatrix} \begin{pmatrix} \sigma^2_{\varepsilon_1,j} & 0 \\ 0 & \sigma^2_{\varepsilon_2} \end{pmatrix} \begin{pmatrix} 1 & H_{21} \\ H_{12} & 1 \end{pmatrix}, \quad j=1,2, \tag{44}$$

where $\sigma^2_{\varepsilon_1}$ varies across regimes (announcement dates, or not) while $\sigma^2_{\varepsilon_2}$ does not.

Writing out the equations in (44) and solving shows that H_{21} is identified as the change in the covariance between η_{1t} and η_{2t}, relative to the change in the variance of η_{1t}:

$$H_{21} = \frac{\Sigma^2_{\eta_1\eta_2} - \Sigma^1_{\eta_1\eta_2}}{\Sigma^2_{\eta_1\eta_1} - \Sigma^1_{\eta_1\eta_1}}. \tag{45}$$

This suggests the estimator,

$$\hat{H}_{21} = \frac{\sum_{t=1}^{T} \hat{\eta}_{2t} Z_t}{\sum_{t=1}^{T} \hat{\eta}_{1t} Z_t}, \tag{46}$$

where $Z_t = D_t\hat{\eta}_{1t}$, where $D_t = -1/T_1$ in the first regime and $D_t = 1/T_2$ in the second regime, where T_1 and T_2 are the number of observation in each regime, and where $\hat{\eta}_t$ are the innovations estimated by full-sample OLS or weighted least squares.

The estimator in (46) is the instrumental variables estimator in the regression of $\hat{\eta}_{2t}$ on $\hat{\eta}_{1t}$, using Z_t as an instrument. Note the similarity of this IV interpretation to the IV interpretation in (39) arising from the very different identifying assumption that the cumulative IRF is lower triangular, so that H_{21} is estimated by the instrumental variables estimator using Y_{2t-1} as an instrument for ΔY_{2t}.

The IV expression (46) connects inference in the SVAR identification by heteroskedasticity to inference in instrumental variables regression, and in particular to inference when instruments might be weak. In (46), a weak instrument corresponds to the case that

Z_t is weakly correlated with $\hat{\eta}_{1t}$, that is, when the population change in the variance of η_{1t}, which appears in the denominator of (45), is small. Using the weak-instrument asymptotic nesting of Staiger and Stock (1997), one can show that, under standard moment conditions, $\hat{H}_{21} \xrightarrow{d} z_2/z_1$, where z_1 and z_2 are jointly normally distributed variables and where the mean of z_1 is $T^{1/2}\left(\Sigma^2_{\eta_1\eta_1} - \Sigma^1_{\eta_1\eta_1}\right)$. If the variability in z_1 is sizeable compared with this mean, then the estimator will in general have a nonnormal and potentially bimodal distribution with heavy tails, and inference based on conventional bootstrap confidence intervals will be misleading.

These weak-instrument problems can arise if the regimes each have many observations, but the difference between the regime variances is small, or if the differences between the variances is large across regimes but one of the regimes has only a small number of observations. In either case, what matters for the distribution of $\hat{H}_{21}$ is the precision of the estimate of the change in the variance of η_{1t}, relative to the true change.

Work on weak-identification robust inference in SVARs identified by heteroskedasticity is in its early stages. Magnusson and Mavroeidis (2014) lay out a general approach to construction of weak-identification robust confidence sets, and Nakamura and Steinsson (2015) implement weak-identification robust inference in their application to differential monetary policy shock heteroskedasticity around FOMC announcement dates.

4.6 Inequality (Sign) Restrictions

The identification schemes discussed so far use a priori information to identify the parameters of H, or the parameters of the first column of H in the case of single shock identification. The sense in which these parameters are identified is the conventional one: different values of the parameter induce different distributions of the data, so that the parameters of H (or H_1) are identified up to a single point. But achieving point identification can entail strong and, in many cases, controversial assumptions. As a result, in two seminal papers, Faust (1998) and Uhlig (2005) argued that instead identification could be achieved more convincingly by imposing restrictions on the signs of the impulse responses. They argued that such an approach connects directly with broad economic theories, for example a broad range of monetary theories suggest that monetary stimulus will have a nonnegative effect on economic activity over a horizon of, say, 1 year. This alternative approach to identification, in which the restriction takes the form of inequality restrictions on the IRF, does not produce point identification, however it does limit the possible values of H (or H_1) to a set. That is, under inequality restrictions, H (or H_1) is set identified.

Set identification introduces new econometric issues for both computation and inference. The standard approach to set identification in SVARs is to use Bayesian methods, which are numerically convenient. This section therefore begins by reviewing the mechanics of Bayesian inequality restriction methods, then turns to inferential issues arising from set identification with a focus on Bayesian sign-identified SVARs. The section concludes with some new research suggesting alternative ways to address these inferential issues.

4.6.1 Inequality Restrictions and Computing an Estimate of the Identified Set

In some applications, economic theory or institutional circumstances might provide a strong argument about the sign of the effect of a given shock on some variable. For example, in a supply and demand example with price and quantity as data, economic theory strongly suggests that the supply elasticity is positive, the demand elasticity is negative, so a positive supply shock increases quantity and decreases price while a positive demand shock increases both quantity and price. More generally, theory might suggest the sign of the effect of a given positive shock on one or more of the variables in the VAR over a certain number of horizons, that is, theory might suggest sign restrictions on elements of the SIRF.

As shown by Faust (1998) and Uhlig (2005) and surveyed by Fry and Pagan (2011), sign restrictions, or more generally inequality restrictions on the SIRF, can be used to help identify the shocks. In general, inequality restrictions provide set, but not point, identification of H, that is, they serve to identify a set of H matrices which contains the unique true H. The econometric problem, then, is how to estimate H and how to perform inference about H given that it is set identified.

The dominant approach in the literature is Bayesian, following Uhlig (2005). The Bayesian inference problem is to compute the posterior distribution of the SIRFs $D(\mathrm{L})$ given the data and a prior. With abuse of notation, we denote this posterior by $f(D|Y)$.

Computing $f(D|Y)$ requires a prior distribution for $D(\mathrm{L})$. Because $D(\mathrm{L})=A(\mathrm{L})^{-1}H$, developing a prior for $D(\mathrm{L})$ in turn entails developing a prior for $A(\mathrm{L})$ and H. Uhlig's (2005) algorithm adopts the unit standard deviation normalization (31), so that $\Sigma_\eta=HH'$. Thus any H can be written as $\Sigma_\eta^{1/2}Q$, where $\Sigma_\eta^{1/2}$ is the Cholesky decomposition of Σ_η and Q is an orthonormal matrix. Thus, under the unit standard deviation normalization, $D(\mathrm{L})=A(\mathrm{L})^{-1}\Sigma_\eta^{1/2}Q$. This expression has substantial computational advantages: $A(\mathrm{L})$ and Σ_η are reduced-form parameters which have conjugate priors under the standard assumption of normally distributed errors, and the only nonstandard part of the prior is Q. Moreover, the dimension for the prior over Q is substantially reduced because $QQ'=\mathrm{I}_n$. Let $\mathfrak{D}$ denote the set of IRFs satisfying the sign restriction, so the prior imposing the sign restrictions is proportional to $\mathbf{1}[D(L)\in\mathfrak{D}]$.

Continuing to abuse notation, and adopting the convention that the priors over $A(\mathrm{L})$, Σ_η, and Q are independent conditional on $D(L)\in\mathfrak{D}$, we can therefore write the posterior $f(D|Y)$ as

$$\begin{aligned} f(D|Y) &\propto f\left(Y|A(L),\Sigma_\eta,Q\right)\pi(A)\pi\left(\Sigma_\eta\right)\pi(Q)\mathbf{1}[D(L)\in\mathfrak{D}] \\ &\propto f\left(A(L),\Sigma_\eta|Y\right)\pi(Q)\mathbf{1}[D(L)\in\mathfrak{D}] \end{aligned} \tag{47}$$

where $f(Y|A(\mathrm{L}),\Sigma_\eta,Q)$ is the Gaussian likelihood for the SVAR $A(\mathrm{L})Y_t=\Sigma_\eta^{1/2}Q\varepsilon_t$, with $\Sigma_\varepsilon=\mathrm{I}_n$, and $f(A(\mathrm{L}),\Sigma_\eta|Y)$ is the posterior of the reduced-form VAR, where the second line in (47) follows because the likelihood does not depend on Q. Uhlig's (2005) algorithm uses conjugate Normal–Wishart priors for A(L) and Σ_{η}^{-1}, so computation of (or drawing from) $f(A(\mathrm{L}),\Sigma_{\eta}|Y)$ is straightforward.

The sign restrictions are imposed using the following algorithm.

(1) Draw a candidate $\widetilde{Q}$ from $\pi(Q)$, and $(\tilde{A}(L), \widetilde{\Sigma}_{\eta})$ from the posterior $f(A(\mathrm{L}),\Sigma_{\eta}|Y)$.

(2) Compute the implied SIRF, $\widetilde{D}(L)=\widetilde{A}(L)^{-1}\widetilde{\Sigma}_{\eta}^{1/2}\widetilde{Q}$.

(3) Retain $\widetilde{D}(L)$ if it satisfies the inequality restrictions.

(4) Repeat Steps (1)–(3) many times to obtain draws from $f(D|Y)$.

This algorithm uses a prior distribution $\pi(Q)$ over the space of orthonormal matrices. In the two-dimensional case all orthonormal matrices can be written as,

$$\widetilde{Q}=\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{48}$$

Thus drawing from $\pi(Q)$ reduces to drawing from a prior over θ, $0\leq\theta\leq 2\pi$. Following Uhlig (2005), it is conventional to use (48) with $\theta\sim \mathrm{U}[0,2\pi]$.

For $n>2$, the restrictions are more complicated. For reasons of computational speed, Rubio-Ramírez et al. (2010) recommend using the QR or householder transformation method for drawing $\widetilde{Q}$, also see Arias et al. (2014). The QR method for constructing a draw of $\widetilde{Q}$ in step (1) proceeds by first drawing a $n\times n$ matrix $\widetilde{W}$, with elements that are independent standard normals, then using the QR decomposition to write $\widetilde{W}=\widetilde{Q}\widetilde{R}$, where $\widetilde{Q}$ is orthonormal and $\widetilde{R}$ is upper triangular.

The choice of prior $\pi(Q)$—in the $n=2$ case, the prior distribution for θ in (48)—is consequential and ends up being informative for the posterior, and we return to this issue in the next section.

4.6.1.1 Single Shock Identification

The discussion here has focused on system identification, however it can also be implemented for identification of a single shock. Specifically, if the inequality restrictions only involve one shock ε_{1t}, then those restrictions only involve the first column of $\widetilde{Q}$, $\widetilde{Q}_1$, and the resulting draw of H_1 is $\Sigma_{\eta}^{1/2}\widetilde{Q}_1$.

4.6.2 Inference When H *Is Set Identified*

The statistical problem is to provide a meaningful characterization of what the data tell us about the true value of H (and thus the true SIRFs) when H is only set identified. As pointed out by Fry and Pagan (2011), Moon and Schorfheide (2012), Moon et al. (2013), and Baumeister and Hamilton (2015a), the standard treatment of uncertainty using the posterior computed according to the algorithm in the preceding subsection raises a number of conceptual and technical problems. Central to these problems is that, because the SIRF is a nonlinear transformation of the parameter over which the prior is placed—in the $n=2$ case, over θ in (48)—a seemingly flat prior over Q ends up being highly informative for inference. Thus inference about the SIRFs is driven by assumptions unrelated to the economic issues at hand (priors over the space of orthonormal matrices) and which have opaque but impactful implications.

We focus on two inferential problems. To illustrate the issues, we consider a stripped-down two-variable SVAR.[v] The researcher is interested in constructing SIRFs and makes the sign restriction that the effect of shock 1 on both variables 1 and 2 is nonnegative on impact and for the first four periods; that is, $D_{h,11} \geq 0$ and $D_{h,21} \geq 0$, $h = 0, \ldots, 4$, where $D(\mathrm{L}) = A(\mathrm{L})^{-1}H$ is the SIRF.

To keep the example as simple as possible, suppose that the reduced-form VAR is first order, that $A(\mathrm{L})$ is diagonal, and that the innovations have identity innovation variance. That is,

$$A(\mathrm{L}) = \begin{pmatrix} 1 - \alpha_1 L & 0 \\ 0 & 1 - \alpha_2 L \end{pmatrix} \text{ where } \alpha_1, \alpha_2 > 0 \text{ and } \Sigma_\eta = I \tag{49}$$

so $Chol(\Sigma_\eta) = I$. Further suppose that the sample size is sufficiently large that these reduced-form parameters can be treated as known; thus the only SVAR uncertainty arises from Q or, because $n = 2$, from θ in (48). The researcher draws candidate orthonormal matrices $\widetilde{\mathrm{Q}}$ using (48), where $\theta \sim \mathrm{U}[0, 2\pi]$.[w] What is the resulting inference on the SIRF $D(\mathrm{L})$?

Under these assumptions, both the identified set for the SIRF for the first shock and the posterior distribution can be computed analytically. In large samples, for a particular draw $\widetilde{\mathrm{Q}}$, the candidate IRF is,

$$\widetilde{D}(L) = \widetilde{A}(L)^{-1}\Sigma_\eta^{1/2}\widetilde{\mathrm{Q}} = \begin{pmatrix} (1 - \alpha_1 L)^{-1}\cos\theta & -(1 - \alpha_1 L)^{-1}\sin\theta \\ (1 - \alpha_2 L)^{-1}\sin\theta & (1 - \alpha_2 L)^{-1}\cos\theta \end{pmatrix}, \tag{50}$$

where the equality uses the large-sample assumption that there is no sampling variability associated with estimation of $A(\mathrm{L})$ or Σ_η, so that the posterior draws $\left(\widetilde{A}(L), \widetilde{\Sigma}_\eta\right) = \left(A(\mathrm{L}), \Sigma_\eta\right)$. Applying the sign restrictions to the first column of (50) implies that $\widetilde{D}(L)$ satisfies the sign restrictions if $\cos\theta \geq 0$ and $\sin\theta \geq 0$, that is, if $0 \leq \theta \leq \pi/2$. Thus the identified set for $D_{21}(\mathrm{L})$ is $0 \leq D_{21}(\mathrm{L}) \leq (1 - \alpha_2 \mathrm{L})^{-1}$, so the identified set for the hth lag of the IRF is $[0, \alpha_2^h]$.

Because $D_{21}(\mathrm{L}) = (1 - \alpha_2 \mathrm{L})^{-1}\sin\theta$, the posterior distribution of the h-period SIRF of shock 2 on variable 1, $D_{h,21}$, is the posterior distribution of $\alpha_2^h \sin\theta$, where $\theta \sim \mathrm{U}[0, \pi/2]$. The mean of this posterior is $E[D_{h,21}] = E\left(\alpha_2^h \sin\theta\right) = 2\alpha_2^h/\pi \approx 0.637\alpha_2^h$ and the posterior median is $0.707\alpha_2^h$. By a change of variables, the posterior density of $D_{h,21}$ is $p_{\hat{D}_{21,i}|Y}(x) \propto 2\alpha_2^h/\pi\sqrt{1 - x^2}$, and the equal-tailed 68% posterior coverage region is $[0.259\alpha_2^h, 0.966\alpha_2^h]$.

[v] This example is similar to the $n = 2$ example in Baumeister and Hamilton (2015a), but further simplified.

[w] In the case $n = 2$, this is equivalent to drawing $\widetilde{\mathrm{Q}}$ using the QR algorithm discussed in Section 4.6.1.

This example illustrates two issues with sign-identified Bayesian inference. First, the posterior coverage interval concentrates strictly within the identified set. As pointed out by Moon and Schorfheide (2012), this result is generic to set-identified Bayesian econometrics in large samples. From a frequentist perspective, this is troubling. In standard parametric settings, in large samples Bayesian 95% posterior intervals coincide with frequentist 95% confidence intervals so, from a frequentist perspective, Bayes confidence sets contain the true parameter value in 95% of all realizations of the sample for all values of the true parameter. This is not the case in this sign-identified setting, however: over repeated samples, the Bayesian interval contains the true parameter value all of the time for some values of the parameter, and none of the time for others.[x]

Second, although the sign restrictions provide no a priori knowledge over the identified region, the "flat" prior on θ induces an informative posterior over the identified set, and in this example places most of the mass on large values of $D_{h,21}$. Although this effect is transparent in this simple example, Baumeister and Hamilton (2015a) show that the implied posteriors over the identified set can have highly informative and unintuitive shapes in more complicated models and in higher dimensions. The presence of sampling uncertainty in A(L) and Σ_η, which this example assumes away, further complicates the problem of knowing how inference is affected by the prior distribution.

In practice there is additional sampling variability in the reduced-form parameters A(L) and Σ_η. In the Bayesian context, this variability is handled by additionally integrating over the priors for those parameters, and with sampling variability the Moon and Schorfheide (2012) result that the posterior coverage set is strictly contained in the identified set need not hold. The lesson of the example, however, is that Bayesian posterior inference depends on the arbitrary prior over the space of orthonormal matrices. In short, conventional Bayesian methods can be justified from a subjectivist Bayes perspective, but doing so results in inferences that a frequentist would find unacceptable.[y]

[x] The asymptotic coincidence of Bayesian and frequentist confidence sets in standard parametric models, and of the posterior mean and the maximum likelihood estimator, is generally known as the Bernstein–von Mises theorem. Freedman (1999) provides an introduction to the theorem and examples of the breakdown of the theorem other than set-identified inference here. Also see Moon and Schorfheide (2012).

[y] A technical issue with Bayesian sign-identified SVARs is that it is conventional to examine impulse responses pointwise, as we did in the example by examining the posterior for $D_{h,21}$ for a given h rather than as a function of h. Thus the values of the VAR parameters corresponding to the posterior mode at one horizon will in general differ from the value at another horizon. See Sims and Zha (1999) for a discussion. Inoue and Kilian (2013) suggest a way to handle this problem and compute most likely IRFs pathways not pointwise.

4.6.2.1 Implications of the Unit Standard Deviation Normalization

The use of the unit standard deviation normalization in conventional Bayesian algorithms means that the SIRFs are all in standard deviation units. For questions posed in native units (what is the effect of a +25 basis point monetary policy shock to the Federal Funds rate?), it is necessary to rescale by the standard deviation of the shock. As Fry and Pagan (2011) point out, in the set-identified context, this rescaling raises additional inferential problems beyond those in the point-identified setting. Specifically, the conversion to the unit effect normalization must be done for each draw, not at a final step, because there is no consistent estimator for H under this method.

4.6.2.2 New Approaches to Inference in Set-Identified SVARs

These inferential problems are difficult and research is ongoing. Here, we briefly describe five new approaches.

The first two approaches are frequentist. A great deal of econometric research over the past decade has tackled frequentist approaches to set-identified inference in general. Inference when the parameter is identified by moment inequalities is nonstandard and—as in the SVAR application—can have the additional problem that the number of moment inequalities can be large but that only one or a few inequalities might be binding for a given value of the parameters. Including many non-binding inequalities for inference typically widens confidence intervals. The two approaches proposed to date for frequentist inference in set-identified SVARs differ in how to handle the problem of many inequalities. Moon et al. (2013) start with all the inequalities, then use a modification of Andrews and Soares's (2010) moment selection procedure to tighten the confidence intervals. Alternatively, Gafarov and Montiel Olea (2015) use only inequality constraints on H (ie, impact effects), which yield substantial computational simplifications. Their results suggest that, despite using fewer restrictions, confidence intervals can be tighter in some applications than if all the inequalities are used.

The remaining approaches are Bayesian. Baumeister and Hamilton (2015a) suggest replacing the prior on Q (on θ in the two-dimensional case) with a prior directly on the impact multiplier, that is, on H_{21}. That prior could be flat, truncated (for sign restrictions) or otherwise informative. This approach addresses the problem in the example earlier that the "flat" prior $\pi(Q)$ on the space of orthonormal matrices induces an informative posterior for the IRF even in large samples. However, this approach remains subject to the Moon and Schorfheide (2012) critique that the Bayesian posterior set asymptotically falls strictly within the identified set.

Giacomini and Kitagawa (2014) propose instead to use robust Bayes inference. This entails sweeping through the set of possible priors over Q, computing posterior regions for each, and reporting the posterior region that is the union of the prior-specific regions, and range of posterior means which is the range of the prior-specific posterior means.

They provide conditions under which the robust credible set converges to the identified set if the sample is large (thereby avoiding the Moon and Schorfheide (2012) critique).

Plagborg-Møller (2015) takes a very different approach and treats the SIRF as the primitive over which the prior is placed; in contrast to Baumeister and Hamilton (2015a,b) who place priors on the impact effect (H), Plagborg-Møller (2015) places a joint prior over the entire IRF. By directly parameterizing the structural MA representation he also handles the problem of noninvertible representations, where the prior serves to distinguish observationally equivalent SVARs.

4.7 Method of External Instruments

Instrumental variables estimation uses some quantifiable exogenous variation in an endogenous variable to estimate the causal effect of the endogenous variable. If a variable measuring such exogenous variation is available for a given shock, but that variable is not included in the VAR, it can be used to estimate the SIRF using a vector extension of instrumental variable regression. This method, which is due to Stock (2008), has been used in a small but increasing number of recent papers including Stock and Watson (2012a), Mertens and Ravn (2013), and Gertler and Karadi (2015). This method is also called the "proxy VAR" method, but we find the "method of external instruments" more descriptive.

Consider identification of the single shock ε_{1t}. Suppose that there is a vector of variables Z_t that satisfies:

$$\text{(i)}\quad E\left(\varepsilon_{1t}Z_t'\right)=\alpha'\neq 0 \tag{51}$$

$$\text{(ii)}\quad E\left(\varepsilon_{jt}Z_t'\right)=0,\quad j=2,\ldots,n. \tag{52}$$

The variable Z_t is called an *external instrument*: external because it is not an element of Y_t in the VAR, and an instrument because it can be used to estimate H_1 by instrumental variables.

Condition (i) corresponds to the usual relevance condition in instrumental variables regression and requires that the instrument be correlated with the endogenous variable of interest, ε_{1t}. Condition (ii) corresponds to the usual condition for instrument exogeneity and requires that the instrument be uncorrelated with the other structural shocks.

Conditions (i) and (ii), combined with the assumption (21) that the shocks are uncorrelated and the unit effect normalization (32), serve to identify H_1 and thus the structural shock. To see this, use $\eta_t=H\varepsilon_t$ along with (i) and (ii) and the partitioning notation (28) to write,

$$\begin{pmatrix} E\left(\eta_{1t}Z_t'\right) \\ E\left(\eta_{\bullet t}Z_t'\right)\end{pmatrix}=E\left(\eta_tZ_t'\right)=E\left(H\varepsilon_tZ_t'\right)=[H_1\quad H_\bullet]\begin{pmatrix} E\left(\varepsilon_{1t}Z_t'\right) \\ E\left(\varepsilon_{\bullet t}Z_t'\right)\end{pmatrix}=H_1\alpha'=\begin{pmatrix}\alpha' \\ H_{1\bullet}\alpha'\end{pmatrix}, \tag{53}$$

where $\eta_{\bullet t}$ denotes the final $n-1$ rows of η_t, the second equality uses $\eta_t = H\varepsilon_t$, the third equality uses the partitioning notation (28), the fourth equality uses (i) and (ii), and the final equality uses the unit effect normalization $H_{11}=1$ in (33).

Equating the first and the final expressions in (53) show that $H_{1\bullet}$, and thus H_1 and ε_{1t}, are identified. In the case of a single instrument, one obtains the expression,

$$H_{1\bullet} = \frac{E\eta_{\bullet t}Z_t}{E\eta_{1t}Z_t}. \tag{54}$$

This expression has a natural instrumental variables interpretation: the effect of ε_{1t} on η_{jt}, that is, the jth element of H, is identified as the coefficient in the population IV regression of η_{jt} onto η_{1t} using the instrument Z_t.

As with standard instrumental variables regression, the success of the method of external instruments depends on having at least one instrument that is strong and credibly exogenous. Although the literature on SVAR estimation using external instruments is young, at least in some circumstances such instruments are plausibly available. For example, the Cochrane and Piazzesi (2002) measure of the monetary shock discussed in Section 4.2 is not in fact the monetary shock: as they note, even if it successfully captures that part of the shock that was learned as an immediate result of FOMC meetings, it is possible that speeches of FOMC members and other Fed actions could provide signals of rate movements before the actual FOMC meeting. Thus, the Cochrane and Piazzesi (2002) measure is better thought of as an instrumental variable for the shock, not the shock itself; that is, it is plausibly correlated with the monetary policy shock and, because it is measured in a window around the FOMC meeting, it is plausibly exogenous. Viewed in this light, many of the series constructed as measures of shocks discussed in Section 4.4 are not in fact the actual shock series but rather are instruments for the shock series. Accordingly, SVARs that include these measures of shocks as a variable are not actually measuring the SIRF with respect to those shocks, but rather are measuring a reduced-form IRF with respect to this instrument for the shocks. In contrast, the method of external instruments identifies the IRF with respect to the structural shock.

As with IV regression more generally, if the instrument is weak then conventional asymptotic inference is unreliable. The details of external instruments in SVARs are sufficiently different from IV regression that the methods for inference under weak identification do not apply directly in the SVAR application. Work on inference with potentially weak external instruments in SVARs is currently under way (Montiel Olea et al., 2016).

5. STRUCTURAL DFMs AND FAVARs

Structural DFMs hold the possibility of solving three recognized shortcomings of SVARs. First, including many variables increases the ability of the innovations to span the space of structural shocks, thereby addressing the omitted variables problem discussed in Section 4.1.2. Second, because the shocks are shocks to the common factors, DFMs provide a natural framework for allowing for measurement error or idiosyncratic variation in individual series, thereby addressing the errors-in-variables problem in Section 4.1.2. Third, high-dimensional structural DFMs make it possible to estimate SIRFs, historical decompositions, and FEVDs that are consistent across arbitrarily many observed variables. Although these goals can be achieved using high-dimensional VARs, because the number of VAR parameters increases with n^2, those large-n VARs require adopting informative priors which typically are statistical in nature. In contrast, because in DFMs the number of parameters increases proportionately to n, DFMs do not require strong restrictions, beyond the testable restrictions of the factor structure, to estimate the parameters.

This section describes how SVAR methods extend directly to DFMs, resulting in a SDFM. In a SDFM, all the factors are unobserved. With a minor modification, one or more of the factors can be treated as observed, in which case the SDFM becomes a *FAVAR*. The key to meshing SVAR identification straight forwardly with DFMs is two normalizations: the "named factor" normalization in Section 2.1.3 for DFMs and the unit effect normalization described in Section 4.1.3 for SVARs. The named factor normalization ascribes the name of, say, the first variable, to the first factor, so that the innovation in the first factor equals the innovation in the common component of the first variable. The unit effect normalization says that the structural shock of interest, say the first shock, has a unit effect on the innovation to the first factor.

Taken together, these normalizations link an innovation in a factor to the innovation in a common component in a variable (naming) and set the scale of the structural shock (unit effect). For example, a one percentage point positive monetary supply shock increases the innovation in the Fed funds factor by one percentage point, which increases the innovation to the common component of the Federal funds rate by one percentage point, which increases the Federal funds rate by one percentage point. These normalizations do not identify the monetary policy shock, but any scheme that would identify the monetary policy shock in a SVAR can now be used to identify the monetary policy shock from the factor innovations.

This section works through the details of the previous paragraph. The section first considers SDFMs in the case of no additional restrictions on the factor loading matrix Λ, next turns to SDFMs in which Λ has additional restrictions and concludes with the extension of SVAR identification methods to FAVARs. This section provides a unified treatment that clarifies the link between SVARs, SDFMs, and FAVARs, including extensions to overidentified cases.

The literature has taken a number of approaches to extending SVARs to structural DFMs, and this section unifies and extends those approaches. The original FAVAR structure is due to Bernanke et al. (2005). Stock and Watson (2005) propose an approach with different normalizations and the treatment here streamlines theirs. The treatment of exactly identified SDFMs here is the same as in Stock and Watson (2012a). The other closest treatments in the literature are Forni and Gambetti (2010), Bai and Ng (2013), Bai and Wang (2014), and Bjørnland and Thorsrud (forthcoming).

5.1 Structural Shocks in DFMs and the Unit Effect Normalization

The structural DFM posits that the innovations in the factors are linear combinations of underlying structural shocks ε_t.

5.1.1 The SDFM

The SDFM augments the static DFM (6) and (7) with the assumption (20) that the factor innovations η_t are linear combinations of the structural shocks ε_t:

$$\overset{n\times1}{X_t} = \overset{n\times r}{\Lambda}\,\overset{r\times1}{F_t} + \overset{n\times1}{e_t} \tag{55}$$

$$\overset{r\times r}{\Phi}(\mathrm{L})\,\overset{r\times1}{F_t} = \overset{r\times q}{G}\,\overset{q\times1}{\eta_t} \quad \text{where } \Phi(\mathrm{L}) = I - \Phi_1\mathrm{L} - \cdots - \Phi_p\mathrm{L}^p, \tag{56}$$

$$\overset{q\times1}{\eta_t} = \overset{q\times q}{H}\,\overset{q\times1}{\varepsilon_t} \tag{57}$$

where following (7), there are r static factors and q dynamic factors, with $r \geq q$. In this system, the q structural shocks ε_t impact the common factors but not the idiosyncratic terms. Additionally, we assume that (SVAR-1)—(SVAR-3) in Section 4.1.4 hold, that the $q \times q$ matrix H is invertible (so the structural shocks can be recovered from the factor innovations), and that the shocks are mutually uncorrelated, that is, Σ_ε is diagonal as in (21).

The SIRF is obtained by substituting (57) into (56) and the result into (55) to obtain,

$$X_t = \Lambda\Phi(\mathrm{L})^{-1}GH\varepsilon_t + e_t. \tag{58}$$

The dynamic causal effect on all n variables of a unit increase in ε_t is the SIRF, which is $\Lambda\Phi(\mathrm{L})^{-1}GH$. Equivalently, the first term on the right-hand side of (58) is the moving average representation of the common component of X_t in terms of the structural shocks.

If interest is only in one shock, say the first shock, then the SIRF for that shock is $\Lambda\Phi(\mathrm{L})^{-1}GH_1$.

The SDFM generalizes the SVAR by allowing for more variables than structural shocks, and by allowing each variable to have idiosyncratic dynamics and/or measurement error. In the special case that there is no idiosyncratic error term (so $e_t = 0$), $r = q = n$, $\Lambda = I$, and $G = I$, the SDFM (58) is simply the structural MA representation (23), where $\Phi(\mathrm{L}) = A(\mathrm{L})$.

5.1.2 Combining the Unit Effect and Named Factor Normalizations

The SDFM (55)–(57) requires three normalizations: Λ, G, and H. We first consider the case $r = q$, so that the static factors have a full-rank covariance matrix, then turn to the case of $r \geq q$.

5.1.2.1 Normalization with $r=q$

In this case, set $G=I$, so that η_t are the innovations to the factors. We use the named factor normalization (12) for Λ and the unit effect normalization (32) for H. Using these two normalizations provides SIRFs in the native units of the variables and ensure that inference about SIRFs will not err by neglecting the data-dependent rescaling needed to convert from standard deviation units (if the unit standard deviation normalization is used) to native units.

As discussed in Section 2.1.3, the named factor normalization associates a factor innovation (and thus a factor) with the innovation to the common component of the naming variable. Without loss of generality, place the naming variables first, so that the first factor adopts the name of the first variable and so forth up to all r factors. Then $\Lambda_{1:r}=I_r$ where, as in (12), $\Lambda_{1:r}$ denotes rows 1 through r of Λ. If there are no overidentifying restrictions on Λ, Λ and F_t can first be estimated by principal components, then transformed as discussed following (12). That is, letting PC denote the principal components estimators,

$$\hat{\Lambda}=\begin{bmatrix} I_r \\ \hat{\Lambda}^{PC}_{r+1:n}\left(\hat{\Lambda}^{PC}_{1:r}\right)^{-1}\end{bmatrix} \quad \text{and} \quad \hat{F}_t=\hat{\Lambda}^{PC}_{1:r}\hat{F}^{PC}_t. \tag{59}$$

Together, the named factor normalization and the unit effect normalization set the scale of the structural shocks. For example, if the oil price and oil price supply shock are ordered first, a unit oil price supply shock induces a unit innovation in the first factor, which is the innovation in the common component of the oil price, which increases the oil price by one native unit (for example, by one percentage point if the oil price is in percent). Restated in terms of the notation in (58) and (59), the impact effect of ε_{1t} on X_{1t} is $\Lambda_1 H_1$, where Λ_1 is the first row of Λ. Because $\Lambda_1=(1\ 0\ \ldots\ 0)$ and the unit effect normalization sets $H_{11}=1$, $\Lambda_1' H_1=1$. Thus a unit increase in ε_{1t} increases X_{1t} by one (native) unit.

This approach extends to overidentifying restrictions on Λ using the methods of Section 2.3.1. To be concrete, in Section 7 we consider an empirical application to identifying an oil supply shock. Our dataset has four different oil prices (the US producer price index for crude petroleum, Brent, West Texas Intermediate (WTI), and the US refiners' acquisition cost of imported oil estimated by the US Energy Information Administration, all in logs). These series, which are available over different time spans, generally move together but their spreads vary because of local conditions and differences in crude oil grades. All four variables are measures of oil prices that have been, used in the oil-macro literature. We therefore model the real oil price factor innovation as impinging on all four real prices with a unit coefficient. The named factor normalization thus is,

$$\begin{bmatrix} p_t^{PPI-Oil} \\ p_t^{Brent} \\ p_t^{WTI} \\ p_t^{RAC} \\ X_{5:n,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ & \Lambda_{5:n} & & \end{bmatrix} \begin{bmatrix} F_t^{oil} \\ F_{2:r,t} \end{bmatrix} + e_t, \tag{60}$$

where $p_t^{PPI-Oil}$ is the logarithm of the real price of crude oil from the producer price index, etc. Strictly speaking, any one of the first four rows of (60) is a naming normalization; the remaining rows are additional restrictions that treat the other three oil prices as additional indicators of the single oil price shock. At this point the number of static factors r is left unspecified; in the empirical application of Section 7, we use $r=8$.

Given the restricted Λ in (60), the static factors can be estimated by restricted principal components as described in Section 2.3.1 (by numerical minimization of the least-squares objective function (14) subject to the restrictions on Λ shown in (60)). The first factor computed from this minimization problem is the oil factor.

5.1.2.2 Normalization with $r > q$

If the empirical analysis of the DFM discussed in Section 2.4.2 indicates that the number of dynamic factors q is less than the number of static factors r, then an additional step is needed to estimate G. This step also needs to be consistent with the unit effect normalization. Accordingly, we normalize G so that

$$G = \begin{bmatrix} I_q \\ G_{q+1:r} \end{bmatrix}, \tag{61}$$

where $G_{q+1:r}$ is an unrestricted $(q-r) \times q$ matrix.

In population, G satisfying (61) can be constructed by first obtaining the innovations a_t to the factors, so that $\Phi(\mathrm{L})F_t = a_t$. Because $r > q$, $\Sigma_a = Ea_t a_t'$ has rank q. Partition $a_t = \left(a_{1t}'\, a_{2t}'\right)'$, where a_{1t} is $q \times 1$ and a_{2t} is $(r-q) \times 1$, and similarly partition Σ_a. Assuming that the upper $q \times q$ block of Σ_a is full rank, we can set $\eta_t = a_{1t}$ and $G_{q+1:r} = \Sigma_{a,21}\Sigma_{a,11}^{-1}$. This construction results in the normalization (61).

In sample, these population objects can be replaced by sample objects. That is, let $\hat{a}_t$ be the residuals from a regression of $\hat{F}_t$ onto p lags of $\hat{F}_t$, let $\hat{\eta}_t = \hat{a}_{1t}$ and let $\hat{\Sigma}_a$ denote the sample covariance matrix of $\hat{a}_t$. Then $\hat{G}_{q+1:r} = \hat{\Sigma}_{a,21}\hat{\Sigma}_{a,11}^{-1}$ is the matrix of coefficients in the regression of $\hat{a}_{2t}$ onto $\hat{\eta}_t$.[z]

[z] This algorithm assumes that the sample inverse $\hat{\Sigma}_{a,11}^{-1}$ is well behaved.

5.1.2.3 Estimation Given an Identification Scheme

With the normalization set, the identification schemes discussed in Section 4 carry over directly. The innovation η_t in Section 4 is now the innovation to the factors, however, the factors (or the subset that are needed) have now been named, and the scale has been set on the structural shocks, so all that remains is to implement the identification scheme. The formulas in Section 4 carry over with the notational modification of setting $A(\mathrm{L})$ in Section 4 to $\Phi(\mathrm{L})$. Section 6 illustrates two contemporaneous restriction identification schemes for oil prices.

5.1.3 Standard Errors for SIRFs

There are various ways to compute standard errors for the SIRFs and for other statistics of interest such as FEVDs. The method used in this chapter is the parametric bootstrap, which (like other standard bootstrap methods) applies only when there is strong identification.

The parametric bootstrap used here proceeds as follows.

1. Estimate Λ, F_t, $\Phi(\mathrm{L})$, G, and Σ_η, and compute the idiosyncratic residual $\hat{e}_t = X_t - \hat{\Lambda}\hat{F}_t$.
2. Estimate univariate autoregressive processes for $\hat{e}_t$, $\hat{e}_{it} = d_i(L)\hat{e}_{it-1} + \zeta_{it}$ (this chapter uses an AR(4)).
3. Generate a bootstrap draw of the data by (a) independently drawing $\widetilde{\eta}_t \sim N\left(0, \hat{\Sigma}_\eta\right)$ and $\zeta_{it} \sim N\left(0, \hat{\sigma}^2_{\zeta_i}\right)$; (b) using the draws of ζ_{it} and the autoregression coefficients $\hat{d}_i(L)$ to generate idiosyncratic errors $\tilde{e}_t$; (c) using $\hat{\Phi}(L)$, $\hat{G}$, and $\widetilde{\eta}_t$ to generate factors $\widetilde{F}_t$; and (d) generating bootstrap data as $\widetilde{X}_t = \hat{\Lambda}\widetilde{F}_t + \widetilde{e}_t$.
4. Using the bootstrap data, estimate Λ, F_t, $\Phi(\mathrm{L})$, G, and H to obtain a bootstrap estimate of the SIRF $\Lambda\Phi(\mathrm{L})GH$. For identification of a subset of shocks, replace H with the columns of H corresponding to the identified shock(s).
5. Repeat Steps 3 and 4 for the desired number of bootstrap draws, then construct bootstrap standard errors, confidence intervals, and/or tests.

Variations on this approach are possible, for example the normal errors drawn in Step 3 could be replaced by block bootstrap resampling of the residuals from the factor VAR and the idiosyncratic autoregression.

There is ongoing work on improving inference in DFMs, SDFMs, and FAVARs using the bootstrap. For example, Yamamoto (2012) develops a bootstrap procedure for FAVARs under the unit standard deviation normalization. Corradi and Swanson (2014) consider the bootstrap for tests of the stability of the factor loadings and factor-augmented regression coefficients. Gonçalves and Perron (2015) establish the asymptotic validity of the bootstrap for the parameters in factor-augmented regressions. Gonçalves et al. (forthcoming) develop bootstrap prediction intervals for DFMs for h-period ahead forecasts. Going into detail on these developments is beyond the scope of this paper.

5.2 Factor-Augmented Vector Autoregressions

Originally developed by Bernanke et al. (2005), FAVARs model some of the factors as observed variables while the remaining factors are unobserved. The FAVAR thus imposes restrictions on the DFM, specifically, that one or more of the factors is measured without error by one or more of the observable variables. Accordingly, SVAR identification methods with the unit effect normalization carry over directly to FAVARs.

The FAVAR model can be represented in two ways. The first is as a DFM with parametric restrictions imposed. For simplicity, consider the case of a single observed factor $\widetilde{F}_t$ which is measured without error by the variable Y_t, r unobserved factors F_t, and order the variable observing $\widetilde{F}_t$ first. Then the structural FAVAR model is,

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} 1 & 0_{1\times r} \\ & \Lambda \end{pmatrix} \begin{pmatrix} \widetilde{F}_t \\ F_t \end{pmatrix} + \begin{pmatrix} 0 \\ u_t \end{pmatrix}, \tag{62}$$

$$F_t^+ = \Phi(\mathrm{L})F_{t-1}^+ + G\eta_t \quad \text{where } F_t^+ = \begin{pmatrix} \widetilde{F}_t \\ F_t \end{pmatrix}, \tag{63}$$

$$\eta_t = H\varepsilon_t. \tag{64}$$

Thus, the FAVAR model combines the unit effect normalization on the factor loadings in (12) with the assumption that there is no idiosyncratic component for the variable observing $\widetilde{F}_t$.

The second, more common representation of the FAVAR model makes the substitution $Y_t = \widetilde{F}_t$ (from the first line of (62)), so that Y_t is included as a factor directly:

$$X_t = \Lambda \begin{pmatrix} Y_t \\ F_t \end{pmatrix} + u_t \tag{65}$$

$$F_t^+ = \Phi(\mathrm{L})F_{t-1}^+ + G\eta_t, \quad \text{where } F_t^+ = \begin{pmatrix} Y_t \\ F_t \end{pmatrix}, \tag{66}$$

$$\eta_t = H\varepsilon_t. \tag{67}$$

With this substitution, the SDFM identification problem becomes the SVAR identification problem, where the VAR is now in terms of $(Y_t \;\; F_t')$. The factors and factor loadings can be estimated by least squares; if there are overidentifying restrictions on Λ, they can be imposed using restricted least squares as in Section 2.3.1[aa].

As an illustration, consider Bernanke et al.'s (2005) FAVAR application of the "slow-R-fast" identification scheme for monetary policy shocks. This original FAVAR application achieves two goals. First, by including a large number of variables, it addresses

[aa] Additional details about implementing this restricted least squares approach are provided in the discussion of the empirical application in Section 7.3.

the omitted variable problem of low-dimensional VARs and in particular aims to resolve the so-called "price puzzle" of monetary VARs (see Ramey, 2016, this Handbook). Second, the joint modeling of these many variables permits estimating internally consistent SIRFs for an arbitrarily large list of variables of interest.

In the slow-*R*-fast scheme, monetary policy shocks or news/financial shocks are assumed not to affect slow-moving variables like output, employment, and price indices within a period, monetary policy responds within a period to shocks to slow-moving variables but not to news or financial shocks, and fast-moving variables (like asset prices) respond to all shocks, including news/financial shocks that are reflected only in those variables.[bb] Let "*s*" and "*f*" denote slow/fast-moving variables, innovations, and shocks, order the slow-moving variables first in X_t, and (departing from the convention earlier) order the slow-moving innovations and factors first, followed by the observable factor ($Y_t = R_t$, the Fed funds rate), then the fast-moving factors and innovations. Then the Bernanke et al. (2005) implementation of the slow-*R*-fast identification scheme is,

$$\begin{pmatrix} X_t^s \\ X_t^f \end{pmatrix} = \begin{pmatrix} \Lambda_{ss} & 0 & 0 \\ \Lambda_{fs} & \Lambda_{fr} & \Lambda_{ff} \end{pmatrix} \begin{pmatrix} F_t^s \\ r_t \\ F_t^f \end{pmatrix} + e_t \qquad (68)$$

$$\boldsymbol{\Phi}(\mathrm{L}) \begin{pmatrix} F_t^s \\ r_t \\ F_t^f \end{pmatrix} = \begin{pmatrix} \eta_t^S \\ \eta_t^r \\ \eta_t^f \end{pmatrix}, \quad \text{and} \qquad (69)$$

$$\begin{pmatrix} \eta_t^S \\ \eta_t^r \\ \eta_t^f \end{pmatrix} = \begin{pmatrix} H_{ss} & 0 & 0 \\ H_{rs} & 1 & 0 \\ H_{fs} & H_{fr} & H_{ff} \end{pmatrix} \begin{pmatrix} \varepsilon_t^S \\ \varepsilon_t^r \\ \varepsilon_t^f \end{pmatrix}. \qquad (70)$$

This scheme imposes overidentifying restrictions on Λ in (68), and those restrictions can be imposed by restricted principal components as in Section 2.3.1.

6. A QUARTERLY 200+ VARIABLE DFM FOR THE UNITED STATES

Sections 6 and 7 illustrate the methods in the previous section using a 207-variable DFM estimated using quarterly data, primarily for the US economy. This section describes the reduced-form DFM: the number of factors, its fit, and its stability. Section 7 uses the reduced-form DFM to estimate structural DFMs that estimate the effect of oil market shocks on the economy under various identification schemes.

[bb] For additional discussion of the slow-*R*-fast scheme, see Christiano et al. (1999).

6.1 Data and Preliminary Transformations

The data are quarterly observations on 207 time series, consisting of real activity variables, prices, productivity and earnings, interest rates and spreads, money and credit, asset and wealth variables, oil market variables, and variables representing international activity. The series are listed by category in Table 1, and a full list is given in the Data Appendix. Data originally available monthly were converted to quarterly by temporal averaging. Real activity variables and several other variables are seasonally adjusted. The dataset updates and extends the dataset used in Stock and Watson (2012a); the main extension is that the dataset used here includes Kilian's (2009) international activity measure and data on oil market, which are used in the analysis in the next section of the effects of oil market shocks on the economy. The full span of the dataset is 1959Q1–2014Q4. Only 145 of the 207 series are available for this full period.

From this full dataset, a subset was formed using the 86 real activity variables in the first four categories in Table 1; this dataset will be referred to as the "real activity dataset." Of the real activity variables, 75 are available over the full sample.

The dataset is described in detail in the Data Appendix.

6.1.1 *Preliminary Transformations and Detrending*

The data were subject to four preliminary transformations. First, the DFM framework summarized in Section 2 and the associated theory assumes that the variables are second-order stationary. For this reason, each series was transformed to be approximately

Table 1 Quarterly time series in the full dataset

	Category	Number of series	Number of series used for factor estimation
(1)	NIPA	20	12
(2)	Industrial production	11	7
(3)	Employment and unemployment	45	30
(4)	Orders, inventories, and sales	10	9
(5)	Housing starts and permits	8	6
(6)	Prices	37	24
(7)	Productivity and labor earnings	10	5
(8)	Interest rates	18	10
(9)	Money and credit	12	6
(10)	International	9	9
(11)	Asset prices, wealth, and household balance sheets	15	10
(12)	Other	2	2
(13)	Oil market variables	10	9
	Total	207	139

Notes: The real activity dataset consists of the variables in the categories 1–4.

integrated of order zero, for example real activity variables were transformed to growth rates, interest rates were transformed to first differences, and prices were transformed to first differences of rates of inflation. The decisions about these transformations were guided by unit root tests combined with judgment, and all similar series within a category were subject to the same transformation (for example, all measures of employment were transformed to growth rates). Selected cointegrating relations were imposed by including error correction terms. Specifically, interest rate spreads are modeled as integrated of order zero.

Second, a small number of outliers were removed. Third, following Stock and Watson (2012a), the long-term mean of each series was removed using a biweight filter with bandwidth of 100 quarters. This step is nonstandard and is discussed in the next subsection. Fourth, after these transformations, the series were standardized to have unit standard deviation.

The Data Appendix provides more details on these steps, including the preliminary transformation of each series.

6.1.1.1 Removing Low-Frequency Trends

Recent research has documented that there has been a long-term slowdown in the mean growth rate of GDP over the postwar period, see Stock and Watson (1996, 2012a), Council of Economic Advisers (2013), and Gordon (2014, 2016). Although there is debate over the cause or causes of this slowdown, it is clear that long-term demographic shifts play an important role. The entry of women into the US labor force during the 1970–90s increased the growth rate of the labor force, and thus increased the growth rate of full-employment GDP, and the aging and retirement of the workforce are now decreasing the labor force participation rate (Aaronson et al., 2014 and references therein). The net effect of these demographic shifts is a reduction in the annual growth rate of GDP due to supply side demographics of approximately one percentage point from the early 1980s to the present. This long-term slowdown is present in many NIPA aggregates and in theory could appear in long-term trends in other series as well, such as interest rates.

These long-term trends, while important in their own right, are relevant to the exercise here for reasons that are technical but nonetheless important. These trends pose two specific problems. First, if the trends are ignored and the series, say employment growth and GDP growth, are modeled as stationary, then because these persistent components are small, the empirically estimated model will be mean reverting. However, the underlying causes of the trends, such as demographics, do not suggest mean reversion. Thus ignoring these long-term trends introduces misspecification errors into forecasts and other reduced-form exercises. Second, structural analysis that aims to quantify the response of macroeconomic variables to specific shocks generally focus on shocks that have transitory effects on GDP growth, such as monetary shocks, demand shocks, or oil supply shocks. Ignoring long-term trends by modeling growth rates as mean reverting introduces specification error in the dynamics of VARs and DFMs: the reduced-form IRFs confound the responses to these transitory shocks with the slowly unfolding trends arising from other sources.

In principal one could model these long-term trends simultaneously with the other factors, for example by adopting a random walk drift term as a factor appearing in the growth rate of some series. This approach has the advantage of explicitly estimating the low-frequency trends simultaneously with the rest of the DFM, however it has the disadvantage of requiring time series models for these trends, thereby introducing the possibility of parametric specification error. Because the purpose of the DFM analysis in this and the next section—and more generally in the vast bulk of the VAR and DFM literature—is analysis and forecasting over short- to medium-horizons (say, up to 4 years), a simpler and arguably more robust approach is simply to remove the low-frequency trends and to estimate the time series model using detrended growth rates.

For these reasons, we detrend all the series prior to estimating the DFM. Although the decline in these growth rates has been persistent, neither the underlying reasons for the declines nor visual inspection of the trends (eg, as displayed in Stock and Watson, 2012a; Gordon, 2014) suggest that they follow a linear trend, so that linear detrending is not appropriate.

The specific detrending method used here follows Stock and Watson (2012a). First, the series is transformed to being approximately integrated of order zero as discussed earlier, for example employment is transformed to employment growth. Second, the trend of each transformed series (for example, employment growth) is estimated nonparametrically using a biweight low-pass filter, with a bandwidth of 100 quarters.[cc]

Fig. 2 compares the biweight filter to three other filters that could be used to estimate the low-frequency trend: an equal-weighted moving average filter with 40 leads and lags (ie, an 81-quarter centered moving average), the Hodrick and Prescott (1997) filter with the conventional quarterly tuning parameter (1600), and the Baxter and King (1999) lowpass bandpass filter with a passband of 200 quarters, truncated to ±100 lags. Each of these filters is linear, so that the estimated trend is $w(\mathrm{L})x_t$ where x_t is the original series (eg, employment growth) and where $w(\mathrm{L})$ generically denotes the filter. Fig. 2A plots the weights of these filters in the time domain and Fig. 2B plots the spectral gain of these filters.[dd]

As can be seen in these figures, the biweight filter is very similar to the Baxter–King lowpass filter. It is also comparable to the equal-weight moving average filter of ±40 quarters, however the biweight filter avoids the noise induced by the sharp cutoff of the moving average filter (these higher frequency components in the moving average filter are evident in the ripples at higher frequencies in the plot of its gain in Fig. 2B). In contrast, all three of these filters focus on much lower frequencies than the Hodrick and

[cc] Tukey's biweight filter $w(\mathrm{L})$ is two sided with $w_j = c(1-(j/B)^2)^2$ for $|j| \leq B$ and $|j| = 0$ otherwise, where B is the bandwidth and c is a normalization constant such that $w(1) = 1$.

[dd] For filter $w(\mathrm{L})$, the estimated trend is $w(\mathrm{L})x_t$ and the detrended series is $x_t - w(\mathrm{L})x_t$. The spectral gain of the filter $w(\mathrm{L})$ is $\|w(e^{i\omega})\|$, where $\|\cdot\|$ is the complex norm.

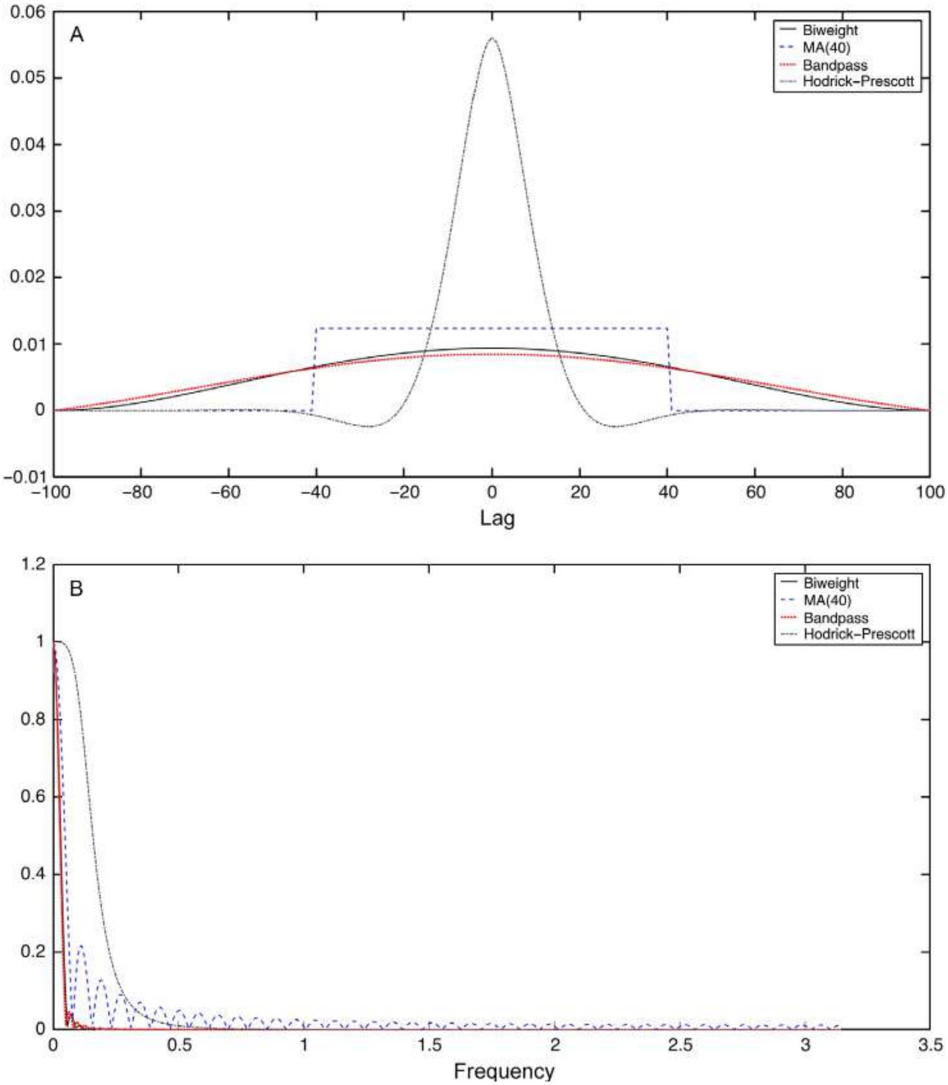


Fig. 2 Lag weights and spectral gain of trend filters. *Notes:* The biweight filter uses a bandwidth (truncation parameter) of 100 quarters. The bandpass filter is a 200-quarter low-pass filter truncated after 100 leads and lags (Baxter and King, 1999). The moving average is equal-weighted with 40 leads and lags. The Hodrick and Prescott (1997) filter uses 1600 as its tuning parameter.

Prescott filter, which places most of its weight on lags of ±15 quarters. The biweight filter estimates trends at multidecadal frequencies, whereas the Hodrick and Prescott trend places considerable weight on fluctuations with periods less than a decade.

The biweight filter needs to be modified for observations near the beginning and end of the sample. One approach would be to estimate a time series model for each series, use forecasts from that model to pad the series at end points, and to apply the filter to this padded series. This approach corresponds to estimating the conditional expectation of the filtered series at the endpoints, given the available data. However, doing so requires estimating a model which raises the problems discussed earlier, which our approach to trend removal aims to avoid: if the trends are ignored when the model is estimated, then the long-term forecasts revert to the mean and this mean reversion potentially introduces misspecification into the trend estimation, but alternatively specifying the trends as part of the model introduces potential parametric misspecification. Instead, the approach used here is to truncate the filter, renormalize, and apply the modified filter directly to the available data for observations within a bandwidth of the ends of the sample.[ee]

6.1.2 *Subset of Series Used to Estimate the Factors*

The data consist of series at multiple levels of aggregation and as a result some of the series equal, or nearly equal, the sum of disaggregated component series. Although the aggregation identity does not hold in logarithms, in the context of the DFM, the idiosyncratic term of the logarithm of higher-level aggregates is highly correlated with the share weighted average of the idiosyncratic term of the logarithms of its disaggregated components. For this reason, when the disaggregated components series are available, the disaggregated components are used to estimate the factors but the higher-level aggregate series are not used.

For example, the dataset contains total IP, IP of final products, IP of consumer goods, and seven sectoral IP measures. The first three series are constructed from the seven sectoral IP series in the dataset, so the idiosyncratic terms of the three aggregates are collinear with those of the seven disaggregated components. Consequently, only the seven disaggregated sectoral IP series are used to estimate the factors.

The aggregates not used for estimating the factors include GDP, total consumption, total employment and, as just stated, total IP. In all, the elimination of aggregates leaves 139 series in the full dataset for estimation of the factors. For the real activity dataset, eliminating aggregates leave 58 disaggregate series for estimating the factor. Table 1 provides the number of series used to estimate the factors by category.

6.2 Real Activity Dataset and Single-Index Model

The first step is to determine the number of static factors in the real activity dataset. Fig. 3 shows three scree plots computed using the 58 disaggregate series in the real activity dataset: using the full dataset and using subsamples split in 1984, a commonly used estimate of the Great Moderation break date. Table 2 (panel A) summarizes statistics related to the number of factors: the marginal R^2 of the factors (that is, the numerical values of the first bar in Fig. 3), the Bai and Ng (2002) IC_{p2} information criterion, and the Ahn and Horenstein (2013) eigenvalue ratio.

[ee] For example, suppose observation t is $m < B$ periods from the end of the sample, where B is the bandwidth. Then the estimated trend at date t is $\sum_{i=-B}^{m} w_i x_{t+i} \Big/ \sum_{i=-B}^{m} w_i$, where w_i is the weight at lag i of the unadjusted two-sided filter.

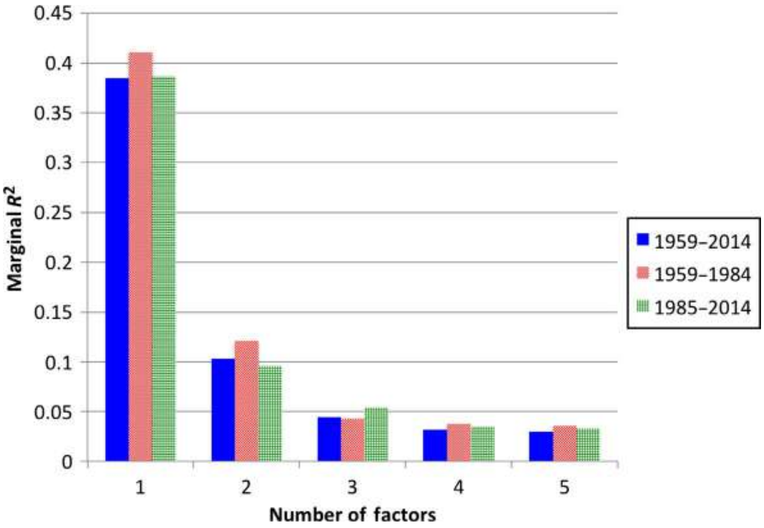


Fig. 3 Scree plot for real activity dataset: full sample, pre-1984, and post-1984.

First consider the full-sample estimates. As seen in Fig. 3, the dominant contribution to the trace R^2 of the 58 subaggregates comes from the first factor which explains fully 38.5% of the variance of the 58 series. Still, there are potentially meaningful contributions to the trace R^2 by the second and possibly higher factors: the marginal R^2 for the second factor over the full sample is 10.3%, for the third is 4.4%, and the total R^2 for the first five is 59.4%, a large increase over the 38.5% explained by the first factor alone. This suggests at least one, but possibly more, factors in the real activity dataset. The Bai and Ng (2002) IC_{p2} criterion estimates three factors, while the Ahn–Horenstein ratio estimates one factor. Unfortunately, such ambiguity is typical, and in such cases judgment must be exercised, and that judgment depends on the purpose to which the DFM is used.

Fig. 1 (shown in Section 1) plots the four-quarter growth rate of GDP, IP, nonfarm employment, and manufacturing and trade sales along with their common components estimated using the single static factor.[ff] Of these, only manufacturing and trade sales were used to estimate the factors, the remaining series being aggregates for which component disaggregated series are in the dataset. Evidently, the full-sample single factor explains the variation of these series at annual through business cycle frequencies.

Fig. 4 presents estimates of the four-quarter growth in GDP and its common components computed using the full sample with 1, 3, and 5 factors (the single-factor common component also appears in Fig. 1). The common component of GDP has an R^2 of 0.73 with a single factor, which increases to 0.88 for five factors. Inspection

[ff] The common component of four-quarter growth is the four-quarter growth of the common component of the series. For the ith series, this common component is $\hat{\Lambda}_i\left(\hat{F}_t + \hat{F}_{t-1} + \hat{F}_{t-2} + \hat{F}_{t-3}\right)$, where $\hat{F}_t$ and $\hat{\Lambda}_i$ are, respectively, the principal components estimator of the factors and the ith row of the estimated factor loadings.

Table 2 Statistics for estimating the number of static factors

(A) Real activity dataset ($N = 58$ disaggregates used for estimating factors)

Number of static factors	Trace R^2	Marginal trace R^2	BN-IC_{p2}	AH-ER
1	0.385	0.385	−0.398	**3.739**
2	0.489	0.103	−0.493	2.338
3	0.533	0.044	**−0.494**	1.384
4	0.565	0.032	−0.475	1.059
5	0.595	0.030	−0.458	1.082

(B) Full dataset ($N = 139$ disaggregates used for estimating factors)

Number of static factors	Trace R^2	Marginal trace R^2	BN-IC_{p2}	AH-ER
1	0.215	0.215	−0.183	**2.662**
2	0.296	0.081	−0.233	1.313
3	0.358	0.062	−0.266	1.540
4	0.398	0.040	**−0.271**	1.368
5	0.427	0.029	−0.262	1.127
6	0.453	0.026	−0.249	1.064
7	0.478	0.024	−0.235	1.035
8	0.501	0.024	−0.223	1.151
9	0.522	0.021	−0.205	1.123
10	0.540	0.018	−0.185	1.057

(C) Amenguel-Watson estimate of number of dynamic factors: BN-IC_{pi} values, full dataset ($N = 139$)

No. of dynamic factors	Number of static factors									
	1	2	3	4	5	6	7	8	9	10
1	−0.098	−0.071	−0.072	−0.068	−0.069	−0.065	−0.064	−0.064	−0.064	−0.060
2		**−0.085**	−0.089	−0.087	−0.089	−0.084	−0.084	−0.084	−0.085	−0.080
3			**−0.090**	**−0.088**	**−0.091**	**−0.088**	**−0.088**	**−0.086**	**−0.086**	**−0.084**
4				−0.077	−0.080	−0.075	−0.075	−0.073	−0.072	−0.069
5					−0.064	−0.060	−0.062	−0.057	−0.055	−0.052
6						−0.045	−0.043	−0.040	−0.037	−0.036
7							−0.024	−0.022	−0.020	−0.018
8								−0.002	0.000	0.003
9									0.021	0.023
10										0.044

Notes: BN-IC_{p2} denotes the Bai and Ng (2002) IC_{p2} information criterion. AH-ER denotes the Ahn and Horenstein (2013) ratio of $(i+1)$th to ith eigenvalues. The minimal BN-IC_{p2} entry in each column, and the maximal Ahn–Horenstein ratio entry in each column, is the respective estimate of the number of factors and is shown in bold. In panel C, the BN-IC_{p2} values are computed using the covariance matrix of the residuals from the regression of the variables onto lagged values of the column number of static factors, estimated by principal components.

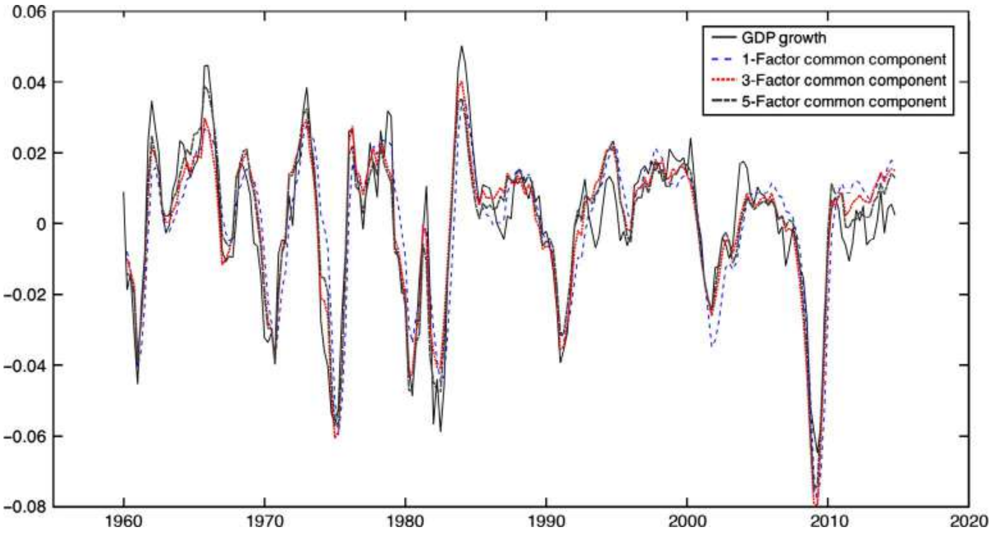


Fig. 4 Four-quarter GDP growth (*black*) and its common component based on 1, 3, and 5 static factors: real activity dataset.

of the fits for all series suggests that the factors beyond the first serve mainly to explain movements in some of the disaggregate series.

In principle, there are at least three possible reasons why there might be more than one factor among these real activity series.

The first possible reason is that there could be a single dynamic factor that manifests as multiple static factors; in the terminology of Section 2, perhaps $q=1$, $r>1$, and G in (7) has fewer rows than columns. As discussed in Section 2, it is possible to estimate the number of dynamic factors given the number of static factors, and applying the Amengual and Watson (2007) test to the real activity dataset, with three static factors, estimates that there is a single dynamic factor. That said, the contribution to the trace R^2 of possible additional dynamic factors remains large in an economic sense, so the estimate of a single dynamic factor is suggestive but not conclusive.

The second possible reason is that these series move in response to multiple structural shocks, and that their responses to those shocks are sufficiently different that the innovations to their common components span the space of more than one aggregated shock.

The third reason, discussed in Section 2, is that structural instability could lead to spuriously large numbers of static factors; for example, if there is a single factor in both the first and second subsamples but a large break in the factor loadings, then the full-sample PC would find two factors, one estimating the first-subsample factor (and being noise in the second subsample), the other estimating the second-subsample factor.

The three scree plots in Fig. 3 does not, however, show evidence of such instability. The scree plots are remarkably stable over the two subsamples and in particular the trace R^2 of the first factor is essentially the same whether the factor is computed over the full sample (38.5%), the pre-1984 subsample (41.1%), or the post-1984 subsample (38.7%). Consistent with this stability, the Bai and Ng (2002) criterion estimates two factors in the first subsample, three in the second, and three in the combined sample.

Fig. 5 provides additional evidence on this stability by plotting the four-quarter growth of the first estimated factor (the first principal component) computed over the full dataset and computed over the pre- and post-1984 subsamples. These series are nearly indistinguishable visually and the correlations between the full-sample estimate and the pre- and post-1984 estimates are high (both exceed 0.99). Thus Figs. 3–5 point to stability of the single-factor model. We defer formal tests for stability to the analysis of the larger DFM based on the full dataset.

Taken together, these results suggest that the first estimated factor (first principal component) based on the full dataset is a good candidate for an index of quarterly real economic activity.

Of course, other variables, such as financial variables, are useful for forecasting and nowcasting real activity. Moreover, while multiple macro shocks plausibly affect the movements of these real variables, the series in the real activity dataset provide only responses to those shocks, not more direct measures, so for an analysis of structural shocks one would want to expand the dataset so that the space of factor innovations more

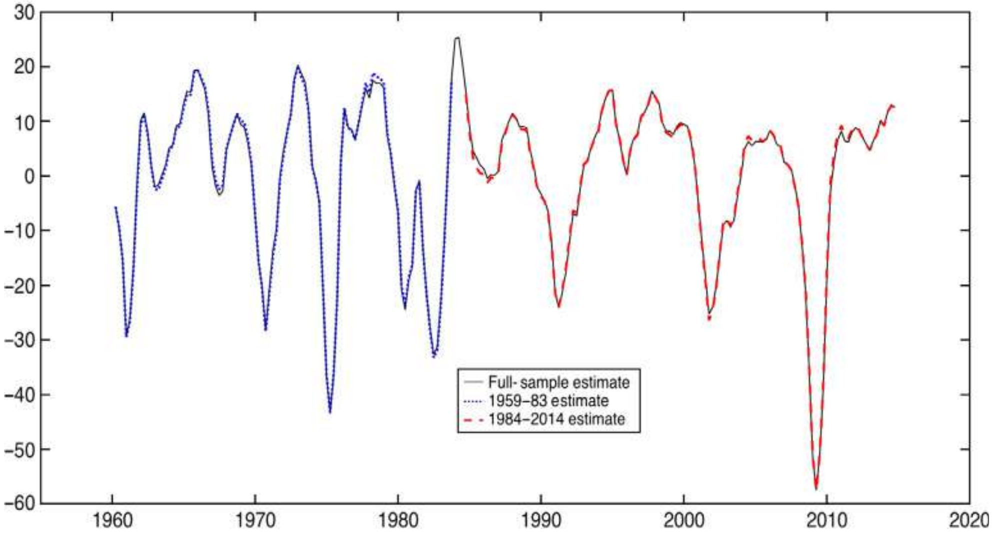


Fig. 5 First factor, real activity dataset: full sample, 1959–84, and 1984–2014.

plausibly spans the space of structural shocks. For example, one would want to include interest rates, which are responsive to monetary policy shocks, measures of oil prices and oil production, which are responsive to oil supply shocks, and measures of inflation, which would respond to both cost and demand shocks.

6.3 The Full Dataset and Multiple-Factor Model

6.3.1 Estimating the Factors and Number of Factors

Fig. 6A is the scree plot for the full dataset with up to 10 factors, and Table 2 (panel B) reports statistics related to estimating the number of factors. The Bai and Ng (2002) criterion chooses four factors, while the Ahn–Horenstein criterion chooses one factor. Compared to the real activity dataset, the first factor explains less of the variation and the decline in higher factors is not as sharp: the marginal R^2 of the fourth factor is 0.040, dropping only to 0.024 for the eighth factor. Under the assumption of anywhere between three and eight static factors, the Amengual and Watson (2007) test selects three dynamic factors (Table 2, panel C), only one less than the four static factors chosen by the Bai and Ng (2002) criterion. As is the case for the static factors, the decline in the marginal R^2 for the dynamic factors is gradual so the evidence on the number of dynamic factors is not clear cut.

Table 3 presents two different measures of the importance of the factors in explaining movements in various series. The first statistic, in columns A, is the R^2 of the common component for the models with 1, 4, and 8 factors; this statistic measures the variation in the series due to contemporaneous variation in the factor. According to the contemporaneous measure in columns A, the first factor explains large fractions of the variation in the growth of GDP and employment, but only small fractions of the variation in prices and financial variables. The second through fourth factors explain the variation in headline inflation, oil prices, housing starts, and some financial variables. The fifth through eighth factors explain much of the variation in labor productivity, hourly compensation, the term spread, and exchange rates. Thus, the additional factors that would be chosen by the Bai and Ng criterion explain substantial fractions of the variation in important classes of series.

Columns B of Table 3 presents a related measure: the fraction of the four quarters ahead forecast error variance due to the dynamic factors, for 1, 4, and 8 dynamic factors, computed under the assumption of eight static factors.[gg] For some series, including housing starts, the Ted spread, and stock prices, the fifth through eighth dynamic factors explain substantial fractions of their variation at the four-quarter horizon. Thus both

[gg] Use (6) and (7) to write $X_t = \Lambda\Phi(\mathrm{L})^{-1}G\eta_t + e_t$. Then the h-period ahead forecast error is $\mathrm{var}\left(\Lambda\sum_{i=0}^{h-1}\Phi_i G\eta_{t-i}\right) + \mathrm{var}(e_t \,|\, e_{t-h}, e_{t-h-1}, \ldots)$, and the fraction of the h-step forecast error variance explained by the dynamic factors is the ratio of the first term in this expression to the total. The term $\mathrm{var}(e_t|e_{t-h}, e_{t-h-1}, \ldots)$ is computed using an AR(4).

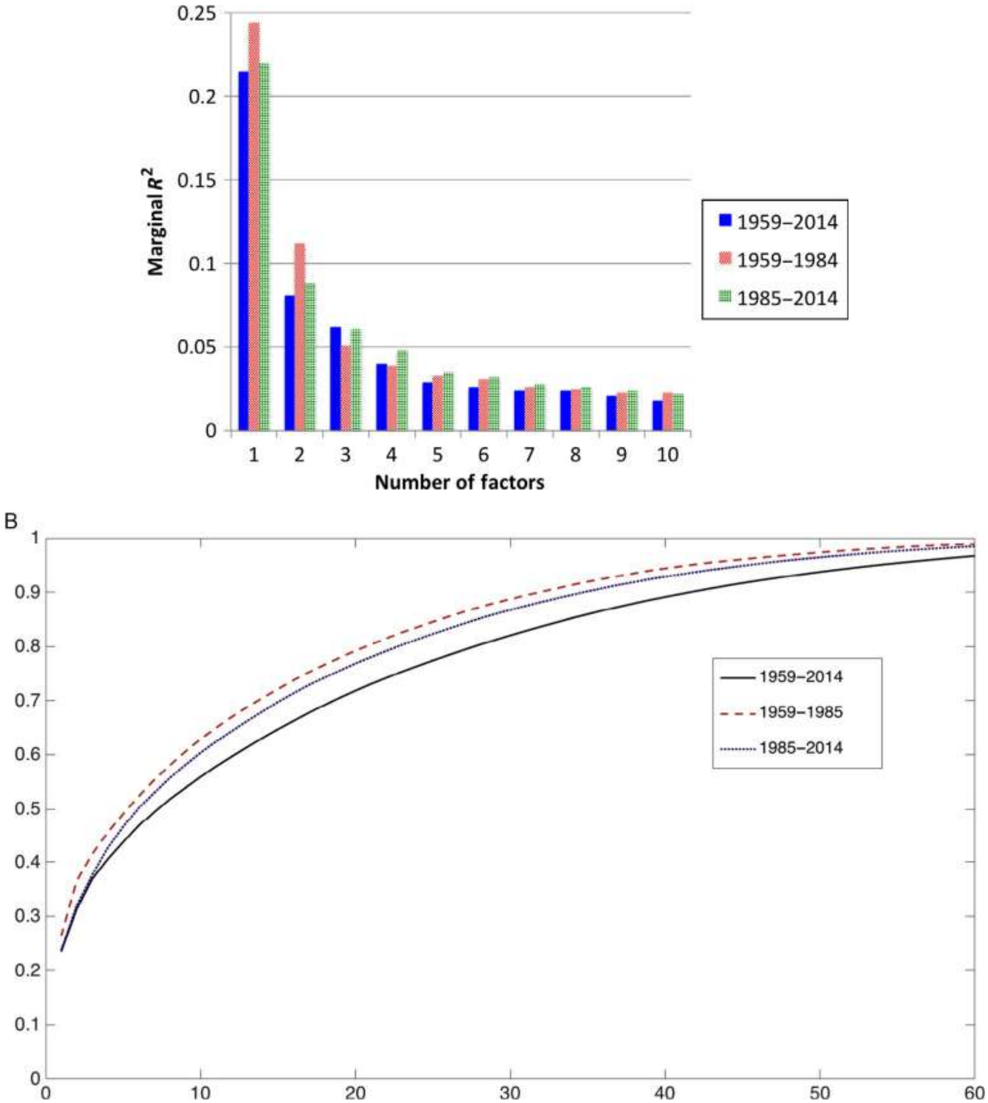


Fig. 6 (A) Scree plot for full dataset: full sample, pre-1984, and post-1984. (B) Cumulative R^2 as a function of the number of factors, 94-variable balanced panel.

blocks of Table 3 suggest that these higher factors, both static and dynamic, capture common innovations that are important for explaining some categories of series.

The scree plot in Fig. 6A and the statistics in Tables 2 and 3 point to a relatively small number of factors—between 4 and 8 factors—describing a large amount of the variation in these series. This said, a substantial amount of the variation remains, and it is germane to ask whether that remaining variation is from idiosyncratic disturbances or whether

Table 3 Importance of factors for selected series for various numbers of static and dynamic factors: full dataset DFM

	A. R^2 of common component			**B. Fraction of four quarters ahead forecast error variance due to common component**		
	Number of static factors *r*			**Number of dynamic factors *q* with *r* = 8 static factors**		
Series	**1**	**4**	**8**	**1**	**4**	**8**
Real GDP	0.54	0.65	0.81	0.39	0.77	0.83
Employment	0.84	0.92	0.93	0.79	0.86	0.90
Housing starts	0.00	0.52	0.67	0.49	0.51	0.75
Inflation (PCE)	0.05	0.51	0.64	0.34	0.66	0.67
Inflation (core PCE)	0.02	0.13	0.17	0.24	0.34	0.41
Labor productivity (NFB)	0.02	0.30	0.59	0.12	0.46	0.54
Real hourly labor compensation (NFB)	0.00	0.25	0.70	0.19	0.67	0.71
Federal funds rate	0.25	0.41	0.54	0.52	0.54	0.62
Ted-spread	0.26	0.59	0.61	0.18	0.33	0.59
Term spread (10 year–3 month)	0.00	0.36	0.72	0.32	0.38	0.63
Exchange rates	0.01	0.22	0.70	0.05	0.60	0.68
Stock prices (SP500)	0.06	0.49	0.73	0.14	0.29	0.79
Real money supply (MZ)	0.00	0.25	0.34	0.15	0.24	0.29
Business loans	0.11	0.49	0.51	0.13	0.16	0.23
Real oil prices	0.04	0.68	0.70	0.40	0.66	0.71
Oil production	0.09	0.10	0.12	0.01	0.04	0.12

there are small remaining correlations across series that could be the result of small, higher factors. Fig. 6B shows the how the trace R^2 increases with the number of principal components, for up to 60 principal components. The key question is whether these higher factors represent common but small fluctuations or, alternatively, are simply the consequence of estimation error, idiosyncratic disturbances, or correlated survey sampling noise because multiple series are derived in part from the same survey instrument. There is a small amount of work investigating the information content in the higher factors. De Mol et al. (2008) find that Bayesian shrinkage methods applied to a large number of series closely approximate principal components forecasts using a small number of factors. Similarly, Stock and Watson (2012b) use empirical Bayes methods to incorporate information in higher factors and find that for many series forecasts using this information do not improve on forecasts using a small number of factors. Carrasco and Rossi (forthcoming) use shrinkage methods to examine whether the higher factors improve forecasts. Onatski (2009, 2010) develops theory for factor models with many weak factors. Although the vast bulk of the literature is consistent with the interpretation that variation in macroeconomic data are associated with a small number of factors, the question of the information content of higher factors remains open and merits additional research.

The choice of the number of factors depends on the application at hand. For forecasting real activity, the sampling error associated with additional factors could outweigh their predictive contribution. In contrast, for the structural DFM analysis in Section 7 we will use eight factors because it is important that the factor innovations span the space of the structural shocks and the higher factors capture variation.

6.3.2 Stability

Table 4 summarizes various statistics related to the subsample stability of the four- and eight-factor models estimated on the full dataset. Table 4 (panel A) summarizes results for equation-by-equation tests of stability. The Chow test is the Wald statistic testing the hypothesis that the factor loadings are constant in a given equation, against the alternative that they have different values before and after the Great Moderation break date of 1984q4 (Stock and Watson, 2009; Breitung and Eickmeier, 2011, Section 3). The Quandt likelihood ratio (QLR) version allows for an unknown break date and is the maximum value of the Chow statistic (the sup-Wald statistic) for potential breaks in the central 70% of the sample, see Breitung and Eickmeier (2011) for additional discussion. In both the Chow and QLR tests, the full-sample estimate of the factors is used as regressors. The table reports the fraction of the series that rejects stability at the 1%, 5%, and 10% significance levels.[hh] Table 4 (panel B) reports a measure of the magnitude of the break, the correlation between the common component computed over a subsample and over the full sample, where the two subsamples considered are the pre- and post-1984 periods. Table 4 (panel C) breaks down the results in Table 4 (panels A and B) by category of series.

The statistics in Table 4 all point to a substantial amount of instability in the factor loadings. More than half the series reject stability at the 5% level for a break in 1984 in the four-factor model, and nearly two-thirds reject in the eight-factor model. As seen in Table 4 (panel C), the finding of a break in the factor loadings in 1984 is widespread across categories of series. Rejection rates are even higher for the QLR test of stability of the factor loadings.

A reasonable worry is that these rejection rates are overstated because the tests are oversized, and Monte Carlo evidence in Breitung and Tenhofen (2011) suggests that the size distortions could be large if the idiosyncratic disturbances are highly serially correlated. For this reason, it is also useful to check if the instability is large in an economic sense.

One such measure of the magnitude of the instability is whether the common component estimated over a subsample is similar to the full-sample common component. As shown in Table 4 (panel B), for at least half the series, the common components estimated

[hh] Results are reported for the 176 of the 207 series with at least 80 quarterly observations in both the pre- and post-1984 subsamples.

Table 4 Stability tests for the four- and eight-factor full dataset DFMs
(A) Fraction of rejections of stability null hypothesis

Level of test	Chow test (1984q4 break)	QLR test
(i) Four factors		
1%	0.39	0.62
5%	0.54	0.77
10%	0.63	0.83
(ii) Eight factors		
1%	0.55	0.94
5%	0.65	0.98
10%	0.72	0.98

(B) Distribution of correlations between full- and split-sample common components

	Percentile of distribution				
	5%	25%	50%	75%	5%
(i) Four factors					
1959–84	0.65	0.89	0.96	0.99	1.00
1985–2014	0.45	0.83	0.95	0.97	0.99
(ii) Eight factors					
1959–84	0.57	0.83	0.92	0.97	0.99
1985–2014	0.43	0.80	0.94	0.97	0.99

(C) Results by category (four factors)

			Median correlation between full- and split-sample common components	
Category	Number of series	Fraction of Chow test rejections for 5% test	1959–84	1985–2014
NIPA	20	0.50	0.98	0.96
Industrial production	10	0.50	0.98	0.97
Employment and unemployment	40	0.40	0.99	0.99
Orders, inventories, and sales	10	0.80	0.98	0.96
Housing starts and permits	8	0.75	0.96	0.91
Prices	35	0.49	0.88	0.90
Productivity and labor earnings	10	0.80	0.92	0.67
Interest rates	12	0.33	0.98	0.94
Money and credit	9	0.89	0.93	0.89
International	3	0.00	0.97	0.97
Asset prices, wealth, and household balance sheets	12	0.58	0.95	0.92
Other	1	1.00	0.95	0.91
Oil market variables	6	0.83	0.79	0.79

Notes: These results are based on the 176 series with data available for at least 80 quarters in both the pre- and post-84 samples. The Chow tests in (A) and (C) test for a break in 1984q4.

using the two subsample factor loadings are highly correlated. For a substantial portion of the series, however, there is a considerable difference between the full-sample and subsample estimates of the common components. Indeed, for 5% of the series, the correlation between the common component estimated post-1984 and the common component estimated over the full sample is less than 50% for both the four- and eight-factor models.

Interestingly, when broken down by category, for some categories, most of the subsample and full-sample common components are highly correlated (Table 4 (panel C), final two columns). This is particularly true for the real activity variables, a finding consistent with the stability of the common component shown in Fig. 5 for the single factor from the real activity dataset. However, for some categories the subsample and full-sample common components are quite different, with median within-category correlations of less than 0.9 in at least one subsample for prices, productivity, money and credit, and oil market variables.

On net, Table 4 points to substantial instability in the DFM. One model of this instability, consistent with the results in the table, is that there was a break around 1984, consistent with empirical results in Stock and Watson (2009), Breitung and Eickmeier (2011), and Chen et al. (2014). However, the results in Table 4 could also be consistent with more complicated models of time variation.

6.4 Can the Eight-Factor DFM Be Approximated by a Low-Dimensional VAR?

A key motivation for DFMs is that using many variables improves the ability of the model to span the space of the structural shocks. But is it possible to approximate the DFM by a small VAR[ii]? If so, those few variables could take the place of the factors for forecasting, and SVAR methods could be used directly to identify structural shocks without needing the SDFM apparatus: in effect, the unobserved factors could be replaced by observed factors in the form of this small number of variables. An approximation to the factors by observable variables could take two forms. The strong version would be for a small number of variables to span the space of the factors. A weaker version would be for a small number of variables to have VAR innovations that span the space of the factor innovations.[ij] Bai and Ng (2006b) develop tests for whether observable variables span the space of the unobserved factors and apply those tests to the Fama-French facots in portfolio analysis. Following Bai and Ng (2006b), we use canonical correlations to examine this possibility in our macro data application.

Table 5 examines the ability of four different VARs to approximate the DFM with eight static factors. The first two VARs are representative of small VARs used in empirical work: a four-variable system (VAR-A) with GDP, total employment, personal consumption expenditure (PCE) inflation, and the Fed funds rate, and an eight-variable system (VAR-B) that

[ii] We thank Chris Sims for raising this question.

[ij] If the observable variables are an invertible contemporaneous linear combination of the factors then the VAR and the factors will have the same innovations, but having the same innovations do not imply that the observable variables are linear combinations of contemporaneous values of the factors.

Table 5 Approximating the eight-factor DFM by a eight-variable VAR

	Canonical correlation							
	1	2	3	4	5	6	7	8
(A) Innovations								
VAR-A	0.76	0.64	0.6	0.49				
VAR-B	0.83	0.67	0.59	0.56	0.37	0.33	0.18	0.01
VAR-C	0.86	0.81	0.78	0.76	0.73	0.58	0.43	0.35
VAR-O	0.83	0.80	0.69	0.56	0.50	0.26	0.16	0.02
(B) Variables and factors								
VAR-A	0.97	0.85	0.79	0.57				
VAR-B	0.97	0.95	0.89	0.83	0.61	0.43	0.26	0.10
VAR-C	0.98	0.93	0.90	0.87	0.79	0.78	0.57	0.41
VAR-O	0.98	0.96	0.88	0.84	0.72	0.39	0.18	0.02

Notes: All VARs contain four lags of all variables. The canonical correlations in panel A are between the VAR residuals and the residuals of a VAR estimated for the eight static factors.

VAR-A was chosen to be typical of four-variable VARs seen in empirical applications. Variables: GDP, total employment, PCE inflation, and Fed funds rate.

VAR-B was chosen to be typical of eight-variable VARs seen in empirical applications. Variables: GDP, total employment, PCE inflation, Fed funds, ISM manufacturing index, real oil prices (PPI-oil), corporate paper-90-day treasury spread, and 10 year–3 month treasury spread.

VAR-C variables were chosen by stepwise maximization of the canonical correlations between the VAR innovations and the static factor innovations. Variables: industrial commodities PPI, stock returns (SP500), unit labor cost (NFB), exchange rates, industrial production, Fed funds, labor compensation per hour (business), and total employment (private).

VAR-O variables: real oil prices (PPI-oil), global oil production, global commodity shipment index, GDP, total employment (private), PCE inflation, Fed funds rate, and trade-weighted US exchange rate index.

Entries are canonical correlations between (A) factor innovations and VAR residuals and (B) factors and observable variables.

additionally has the ISM manufacturing index, the oil price PPI, the corporate paper-90-day treasury spread, and the 3 month–10 year treasury term spread. The eight variables in the third VAR (VAR-C) were selected using a stepwise procedure to produce a high fit between VAR residuals and the innovations in the eight static factors (ie, the residuals in the VAR with the eight static factors). This procedure led to the VAR-C variables being the index of IP, real personal consumption expenditures, government spending, the PPI for industrial commodities, unit labor costs for business, the S&P500, the 6 month–3 month term spread, and a trade-weighted index of exchange rates.[kk] The final VAR, VAR-O, is used for the SVAR analysis of the effect of oil shocks in Section 7 and is discussed there.

[kk] The variables in VAR-C were chosen from the 207 variables so that the *i*th variable maximizes the *i*th canonical correlation between the residuals from the *i*-variable VAR and the residuals from the eight-factor VAR. In the first step, the variable yielding the highest canonical correlation between its autoregressive residual and the factor VAR residuals was chosen. In the second step, the variable that maximized the second canonical correlation among all 206 two-variable VAR residuals (given the first VAR variable) and the factor VAR residuals was chosen. These steps continued until eight variables were chosen.

Table 5 (panel A) examines whether the VAR innovations are linear combinations of the eight innovations in the static factors by reporting the canonical correlations between the two sets of residuals. For the four-variable VAR, the first canonical correlation is large, as are the first several canonical correlations in the eight-variable VARs, indicating that some linear combinations of the DFM innovations can be constructed from linear combinations of the VAR innovations. But the canonical correlations drop off substantially. For the eight-variable VAR-B, the final four canonical correlations are less than 0.40, indicating that the innovation space of this typical VAR differs substantially from the innovation space of the factors. Even for VAR-C, for which the variables were chosen to maximize the stepwise canonical correlations of the innovations, the final three canonical correlations are less than 0.60, indicating that there is substantial variation in the factor innovations that is not captured by the VAR innovations.

Table 5 (panel B) examine whether the observable variables span the space of the factors, without leads and lags, by reporting the canonical correlations between the observable variables and the factors for the three VARs. For the four-variable VAR, the canonical correlations measure the extent to which the observable variables are linear combinations of the factors; for the eight-variable VARs, the canonical correlations measure whether the spaces spanned by the observable variables and the factors are the same, so that the eight latent factors estimated from the full dataset could be replaced by the eight observable variables. The canonical correlations in panel B indicate that the observable variables are not good approximations to the factors. In VAR-B, three of the canonical correlations are less than 0.50, and even in VAR-C two of the canonical correlations are less than 0.6.

These results have several caveats. Because the factors are estimated, the sample canonical correlations will be less than one even if in population they equal one, and no measure of sampling variability is provided. Also, VAR-C was chosen by a stepwise procedure, and presumably a better approximation would obtain were it possible to choose the approximating VAR out of all possible eight-variable VARs.[ll]

Still, these results suggest that while typical VARs capture important aspects of the variation in the factors, they fail to span the space of the factors and their innovations fail to span the space of the factor innovations. Overall, these results suggest that the DFM, by summarizing information from a large number of series and reducing the effect of measurement error and idiosyncratic variation, produces factor innovations that contain information not contained in small VARs.

[ll] Other methods for selecting variables, for example stepwise maximization of the *i*th canonical correlation between the variable and the factor (instead of between the VAR innovations and the factor innovations) yielded similar results to those for VAR-C in Table 5.

7. MACROECONOMIC EFFECTS OF OIL SUPPLY SHOCKS

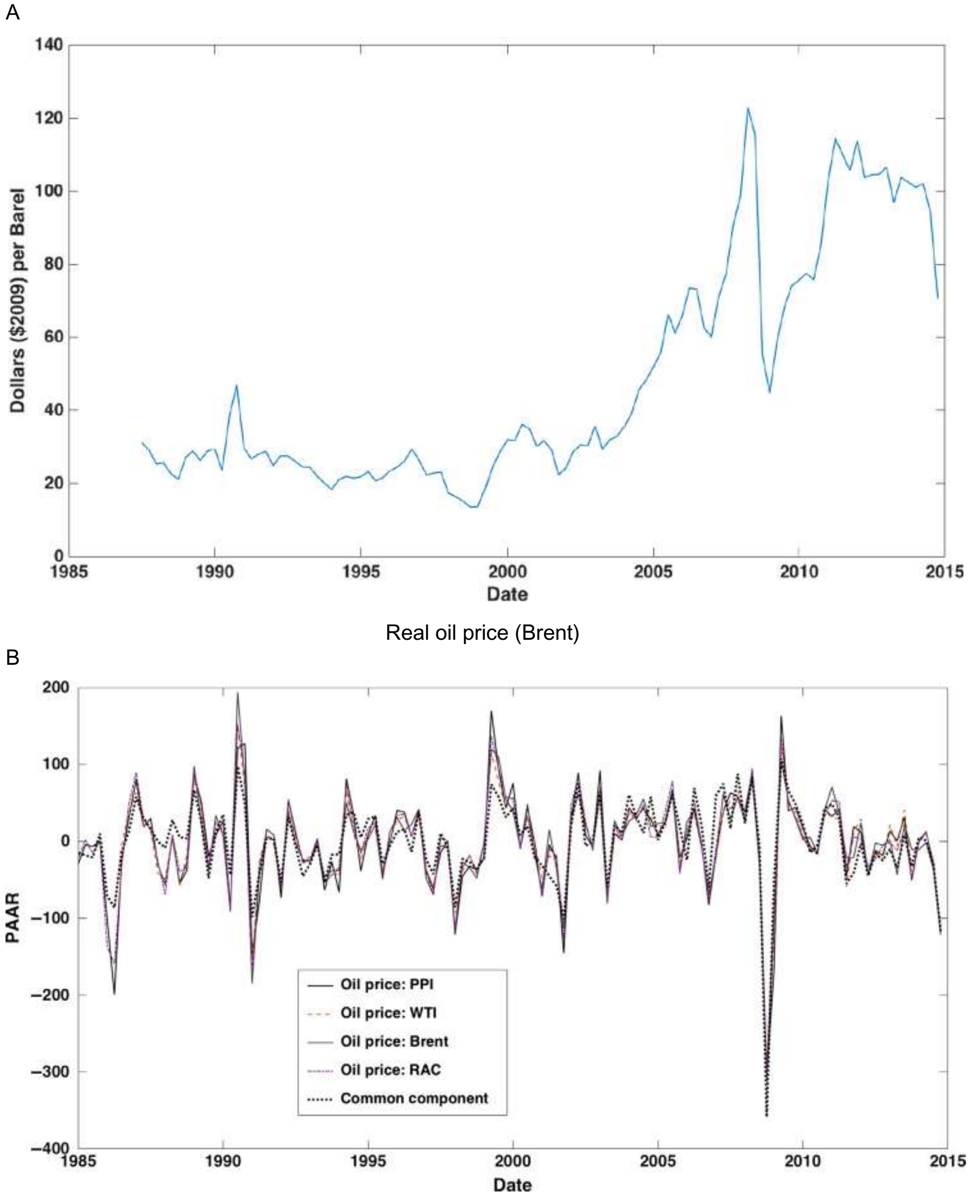
This section works through an empirical example that extends SVAR identification schemes to SDFMs. The application is to estimating the macroeconomic effects of oil market shocks, using identification schemes taken from the literature on oil and the macroeconomy. For comparison purposes, results are provided using a 207-variable SDFM with eight factors, a 207-variable FAVAR in which one or more of factors are treated as observed, and an eight-variable SVAR.

7.1 Oil Prices and the Macroeconomy: Old Questions, New Answers

Oil plays a central role in developed economies, and for much of the past half century the price of oil has been highly volatile. The oil price increases of the 1970s were closely linked to events such as the 1973–74 OPEC oil embargo and wars in the Middle East, as well as to developments in international oil markets (Hamilton, 2013; Baumeister and Kilian, 2016). The late 1980s through early 2000s were a period of relative quiescence, interrupted mainly by the spike in oil prices during the Iraqi invasion of Kuwait. Since the early 2000s oil prices have again been volatile. The nominal price of Brent oil, an international benchmark, rose from under $30/barrel in 2002 to a peak of approximately $140/barrel in June 2008. Oil prices collapsed during the financial crisis and ensuing recession, but by the spring of 2011 recovered to just over $100/barrel. Then, beginning the summer of 2014, oil prices fell sharply and Brent went below $30 in early 2016, a decline that was widely seen as stemming in part from the sharp increase in unconventional oil production (hydraulic fracturing). The real oil price over the last three decades is plotted in Fig. 7A.

Fig. 7B shows four measures of the quarterly percentage change in oil prices, along with its common component estimated using the eight factors from the 207-variable DFM of Section 6. Fig. 7B reminds us that there is no single price of oil, rather oil is a heterogeneous commodity differentiated by grade and extraction location. The four measures of real oil prices (Brent, WTI, US refiners' acquisition cost of imported oil and the PPI for oil, all deflated by the core PCE price index) move closely together but are not identical. As discussed later, in this section these series are restricted to have the same common component, which (as can be seen in Fig. 7B) captures the common movements in these four price indices.

Economists have attempted to quantify the effect of oil supply shocks on the US economy ever since the oil supply disruptions of the 1970s. In seminal work, Hamilton (1983) found that oil price jumps presaged US recessions; see Hamilton (2003, 2009) for updated extensive discussions. Given the historical context of the 1970s, the first wave of analysis of the effect of oil supply shocks on the economy generally treated unexpected changes in oil prices as exogenous and as equivalent to oil supply shocks. In the context of SVAR analysis,



Real oil price (Brent)

Quarterly percent change in real oil price: four oil price series and the common component

Fig. 7 Real oil price (2009 dollars) and its quarterly percent change.

this equivalent allows treating the innovation in the oil price equation as an exogenous shock, which in turn corresponds to ordering oil first in a Cholesky decomposition.[mm]

Recent research, however, has apended this early view that unexpected oil price movements are solely the result of exogenous oil supply shocks and has argued instead that much or most movements in oil prices are in fact due to shocks to global demand or perhaps to demand shocks that are specific to oil (inventory demand). For example, this view accords with the broad perception that the long climb of oil prices in the mid-2000s was associated with increasing global demand, including demand from China, in the face of conventional supply that was growing slowly or even declining before the boom in unconventional oil production began in the late 2000s and early 2010s.

The potential importance of aggregate demand shocks for determining oil prices was proposed in the academic literature by Barsky and Kilian (2002) and has been influentially promoted by Kilian (2008a,b, 2009). Econometric attempts to distinguish oil supply shocks from demand shocks generally do so using SVARs, broadly relying on three identification schemes. The first relies on timing restrictions to impose zeros in the H matrix of Eq. (20). The logic here, due to Kilian (2009), starts by noting that it is difficult to adjust oil production quickly in response to price changes, so that innovations in the quantity of oil produced are unresponsive to demand shocks during a sufficiently short period of time. As is discussed later in more detail, this timing restriction can be used to identify oil supply shocks.

The second identification scheme uses inequality restrictions: standard supply and demand reasoning suggest that a positive shock to the supply of oil will push down oil prices and increase oil consumption, whereas a positive shock to aggregate demand would push up both oil prices and consumption. This sign restriction approach has been applied by Peersman and Van Robays (2009), Lippi and Nobili (2012), Kilian and Murphy (2012, 2014), Baumeister and Peersman (2013), Lütkepohl and Netšunajev (2014), and Baumeister and Hamilton (2015b) among others.

The third identification approach identifies the response to supply shocks using instrumental variables. Hamilton (2003) used a list of exogenous oil supply disruptions, such as the Iraqi invasion of Kuwait, as an instrument in a single-equation estimation of the effect of oil supply shocks on GDP which Kilian (2008b) extended, also in a single-equation context. Stock and Watson (2012a) used the method of external instruments in a SDFM to estimate the impulse responses to oil supply shocks using various instruments, including (like Hamilton, 2003) a list of oil supply disruptions.

Broadly speaking, a common finding from this second wave of research is that oil supply shocks account for a small amount of the variation both in oil prices and in aggregate economic activity, at least since the 1970s. Moreover, this research finds that much or most of

[mm] Papers adopting this approach include Shapiro and Watson (1988) and Blanchard and Galí (2010).

the variation in oil prices (at least through 2014) arises from shifts in demand, mainly aggregate demand or demand more specifically for oil.

This section shows how this recent research on oil supply shocks can be extended from SVARs to FAVARs and SDFMs. For simplicity, this illustration is restricted to two contemporaneous identification schemes. The papers closest to the treatment in this section are Aastveit (2014), who uses a FAVAR with timing restrictions similar to the ones used here, Charnavoki and Dolado (2014) and Juvenal and Petrella (2015), who use sign restrictions in a SDFM, and Aastveit et al. (2015), who use a combination of sign and timing restrictions in a FAVAR. The results of this section are confirmatory of these papers and more generally of the modern literature that stresses the importance of demand shocks for determining oil prices, and the small role that oil supply shocks have played in determining oil production since the early 1980s. Although the purpose of this section is to illustrate these methods, the work here does contain some novel features and new results.

7.2 Identification Schemes

We consider two identification schemes based on the contemporaneous zero restrictions in the H matrix, that is, schemes of the form discussed in Section 4.2. The first identification scheme, which was used in the early oil shocks literature, treats oil prices as exogenous with oil price innovations assumed to be oil price supply shocks. The second identification scheme follows Kilian (2009) and distinguishes oil supply shocks from demand shocks by assuming that oil production responds to demand shocks only with a lag.[nn] The literature continues to evolve, for example Kilian and Murphy (2014) include inventory data and use sign restrictions to help to identify oil-specific demand shocks. The treatment in this section does not aim to push the frontier on this empirical issue, but rather to illustrate SDFM, FAVAR, and SVAR methods in a simple setting that is still sufficiently rich to highlight methods and modeling choices.

[nn] Kilian's (2009) treatment used monthly data, whereas here we use quarterly data. The timing restrictions, for example the sluggish response of production to demand, are more appropriate at the monthly than at the quarterly level. Güntner (2014) used sign restrictions in an oil-macro SVAR to identify demand shocks and find that oil producers respond negligibly to demand shocks within the month, and that most producers respond negligibly within a quarter, although Saudi Arabia is estimated to respond after a delay of 2 months. The recent development of fracking and horizontal drilling technology also could undercut the validity of the timing restriction, especially at the quarterly level, because new wells are drilled and fracked relatively quickly (in some cases in a matter of weeks). In addition, because well productivity declines much more rapidly than for conventional wells, nonconventional production can respond more quickly to price than can most conventional production. If the restrictions are valid at the monthly frequency but not quarterly, our estimated supply shocks would potentially include demand shocks, biasing our SIRFs. Despite these caveats, however, the results here are similar to those in Kilian's (2009) and Aastveit's (2014) monthly treatments with the same exclusion restrictions.

The "oil exogenous" identification scheme is implemented in three related models: a 207-variable SDFM with eight unobserved factors, a 207-variable FAVAR (that is, a SDFM in which some of the eight factors are treated as observed), and an eight-variable SVAR. The Kilian (2009) identification scheme is examined in a eight-variable VAR, in a 207-variable FAVAR with three observed and five unobserved factors, and a 207-variable FAVAR with one observed and seven unobserved factors. As is discussed later, this final FAVAR is used instead of a SDFM with all factors unobserved because the oil production innovation plays such a small macroeconomic role that it appears not to be spanned (or is weakly spanned) by the space of innovations to the macro factors.

For the SVAR, identification requires sufficient restrictions on H to identify the column of H associated with the oil supply shock and, for the second assumption, the columns associated with the aggregate demand and oil-specific demand shocks.

For the FAVARs in which the relevant factors (oil prices in the "oil price exogenous" case, and oil production, aggregate demand, and oil prices in the Kilian (2009) case) are all modeled as observed, no additional identifying restrictions are needed beyond the SVAR identifying restrictions.

For the SDFM and for the FAVAR with only one of the three factors observed, identification also entails normalizations on the factor loadings Λ and on the matrix G relating the dynamic factor innovations to the static factor innovations.

The SDFM and FAVAR models require determining the number of dynamic factors. Although Table 2 (panel C) can be interpreted as suggesting fewer dynamic than static factors, we err on the side of over-specifying the space of innovations so that they span the space of the reduced number of shocks of interest, and therefore set the number of dynamic factors equal to the number of static factors, so in turn the dimension of η_t (the factor innovations) is eight. Thus we adopt the normalization that G is the identity matrix.

7.2.1 Identification by Treating Oil Prices Innovations as Exogenous

The historical starting point of the oil shock literature holds that any unexpected change in oil prices is exogenous to developments in the US economy. One motivation for this assumption is that if unexpected changes in oil prices arise from unexpected developments in supply—either supply disruptions from geopolitical developments or unexpected upticks in production—then those changes are specific to oil supply, and thus can be thought of as oil supply shocks. A weaker interpretation is that oil prices are determined in the world market for oil so that unexpected changes in oil prices reflect international developments in the oil market, and thus are exogenous shocks (although they could be either oil supply or demand shocks). In either case, an unexpected increase in the real price of oil is interpreted as an exogenous oil price shock. Because the oil price shock is identified as the innovation in the (log) price of oil, it is possible to estimate structural impulse responses with respect to this shock.

7.2.1.1 SVAR and FAVAR

Without loss of generality, order the oil price first in the list of variables. The assumption that the oil price shock ε_t^{oil} is exogenous, combined with the unit effect normalization, implies that $\eta_{1t} = \varepsilon_t^{oil}$. Thus the relation between η_t and ε_t in (28) can be written,

$$\eta_t = \begin{pmatrix} 1 & 0 \\ H_{\bullet 1} & H_{\bullet\bullet} \end{pmatrix} \begin{pmatrix} \varepsilon_t^{oil} \\ \widetilde{\eta}_{\bullet t} \end{pmatrix}, \tag{71}$$

where $\widetilde{\eta}_{\bullet t}$ spans the space of η_t orthogonal to η_{1t}. The vector $H_{\bullet 1}$ is identified as the coefficient in the (population) regression of $\eta_{\bullet t}$ on η_{1t}.

In practice, this identification scheme is conveniently implemented by ordering oil first in a Cholesky decomposition; the ordering of the remaining variables does not matter for the purpose of identifying and estimating the SIRFs with respect to the oil shock.

7.2.1.2 SDFM

In addition to the identification of H in (71), identification in the SDFM requires normalization restrictions on the factor loadings Λ and on G. Because the number of static and dynamic factors is the same, we follow Section 5.1.2 and set G to the identity matrix.

If the dataset had a single oil price, then the named factor normalization would equate the innovation in the first factor with the innovations in the common component of oil. Accordingly, with a single oil price measure ordered first among the DFM variables, the first row of Λ would be $\Lambda_1 = (1\ 0\ \dots\ 0)$. The normalization of the next seven rows (there are eight static factors) is arbitrary, although some care must be taken so that the innovations of the common components of those seven variables, plus oil prices, spans the space of the eight factor innovations.

The 207-variable dataset, however, contains not one but four different measures of oil prices: Brent, WTI, refiners' acquisition cost, and the producer price of oil. All four series, specified as percentage changes in price, are used as indicators that measure the percentage change in the common (unobserved) price of oil, which is identified as the first factor by applying the named factor normalization to all four series. This approach entails using the specification of Λ in (60).[oo]

Because G is set to the identity matrix, the innovation to the oil price factor is the oil price innovation.

[oo] Figure 7 suggests that real oil prices are I(1), and we use oil price growth rates in the empirical analysis, ignoring cointegration restrictions. This is the second approach to handling cointegration discussed in Section 2.1.4. In a fully parametric DFM (Section 2.3.2), imposing cointegration improves efficiency of the estimates, but the constraint may lead to less efficient estimates in nonparametric (principal components) models. This treatment also allows all four oil prices to be used to estimate the loading on the first factor and therefore to name (identify) the oil price factor.

7.2.2 Kilian (2009) Identification

Following Kilian (2009), this scheme separately identifies an oil supply shock, an aggregate world commodity demand shock, and an oil-specific demand shock. This is accomplished by augmenting the system with a measure of oil production (barrels pumped during the quarter) and a measure of global real economic activity. The measure of global economic activity we use here is Kilian's (2009) global index of bulk dry goods shipments.

7.2.2.1 SVAR and FAVAR

The justification for the exclusion restrictions in the H matrix is as follows. (i) Because of technological delays in the ability to adjust production at existing wells, to shut down wells, and to bring new wells on line, crude oil production responds with a delay to demand shocks or to any other macro or global shocks. Thus, within a period, an unexpected change in oil production is exogenous and is therefore an exogenous supply shocks (ε_t^{OS}). Thus the innovation to oil production equals the oil supply shock. (ii) Global economic activity can respond immediately to oil supply shocks and responds to global aggregate demand shocks (ε_t^{GD}), but otherwise is sluggish and responds to no other shocks within the period. (iii) Real oil prices respond to oil supply shocks and aggregate demand shocks within the period, and to other oil price-specific shocks as well, but to no other macro or global shocks. Kilian interprets the other oil price-specific shocks (ε_t^{OD}) as shocks to oil demand that are distinct from aggregate demand shocks; examples are oil inventory demand shocks, perhaps driven by anticipated oil supply shocks, or speculative demand shocks.

The foregoing logic imparts an upper triangular structure to H and a Cholesky ordering to the shocks:

$$\begin{pmatrix} \eta_t^{oilproduction} \\ \eta_t^{globalactivity} \\ \eta_t^{oilprice} \\ \eta_{\bullet t} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ H_{12} & 1 & 0 & 0 \\ H_{13} & H_{23} & 1 & 0 \\ H_{1\bullet} & H_{2\bullet} & H_{3\bullet} & H_{\bullet\bullet} \end{pmatrix} \begin{pmatrix} \varepsilon_t^{OS} \\ \varepsilon_t^{GD} \\ \varepsilon_t^{OD} \\ \tilde{\eta}_{\bullet t} \end{pmatrix}, \tag{72}$$

where the unit coefficients on the diagonal impose the unit effect normalization and the variables are ordered such that the innovations are to global oil production, global aggregate demand, the price of oil, and the remaining series. The first three rows of H identify the three shocks of interest, and the remaining elements of the first, second, and third rows of H are identified as the population regression coefficients of the innovations on the shocks.

For convenience, the identification scheme (72) can be implemented by ordering the first three variables in the order of (72) and adopting a lower triangular ordering (Cholesky factorization) for the remaining variables, renormalized so that the diagonal elements of H equal 1. Only the first three shocks are identified, and the SIRFs with respect to those shocks do not depend on the ordering of the remaining variables.

7.2.2.2 SDFM

The SDFM is identified by the restrictions on H in (72), the named factor normalization for Λ, and setting G to be the identity matrix.

As mentioned earlier, the SDFM implementation treats the oil production factor as observed and the remaining seven factors as unobserved. Of these seven unobserved factors, we are interested in two linear combinations of the factor innovations that correspond to the global activity innovation and the oil price innovation. The combination of one observed factor, two identified unobserved factors, and five unidentified unobserved factors gives a hybrid FAVAR-SDFM. In this hybrid, the named factor normalization is,

$$\begin{bmatrix} Oil\ production_t \\ Global\ activity_t \\ p_t^{PPI-Oil} \\ p_t^{Brent} \\ p_t^{WTI} \\ p_t^{RAC} \\ X_{7:n,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \Lambda_{7:n} & & & & & \end{bmatrix} \begin{bmatrix} F_t^{Oil\ production} \\ F_t^{Global\ activity} \\ F_t^{oil\ price} \\ F_{4:r,t} \end{bmatrix} + \begin{bmatrix} 0 \\ e_{2t} \\ e_{3t} \\ e_{4t} \\ e_{5t} \\ e_{6t} \\ e_{7:n,t} \end{bmatrix} \tag{73}$$

where the first variable is $OilProduction_t$, which is treated as an observed factor, the second variable is the global activity (commodity shipment) index, and the next four variables are the four oil price measures. The first factor is the observed oil production factor. The next two factors, which are unobserved, are the global activity factor and the oil price factor. The identity matrix normalization of G associates the innovations with these factors, so that those innovations align with the first three innovations in (72).

7.3 Comparison SVAR and Estimation Details

7.3.1 Comparison SVAR

Because the SDFM is specified with eight static and dynamic factors, the comparison SVAR was chosen to have eight variables. Of the eight variables in the SVAR, three are those in Kilian's (2009) three-variable SVAR: the real oil price (PPI-oil), global oil production, and Kilian's (2009) global activity index (bulk dry shipping activity). The remaining five variables were chosen to represent different aspects of US aggregate activity, inflation, and financial markets: GDP, total employment, PCE inflation, the Federal funds rate, and a trade-weighted index of exchange rates.

Canonical correlations between the factor innovations and the VAR innovations are summarized in the "VAR-O" row of Table 5 (panel A). While the first few canonical correlations are large, the final four are 0.50 or less. Evidently, the VAR and factor innovations span substantially different spaces.

7.3.2 Summary of SDFM Estimation Steps

7.3.2.1 Summary of Steps

We now summarize the steps entailed in estimating the SIRF for the SDFM of Section 7.2.2 with one observed factor and three identified shocks. From (58), the SIRF with respect to the ith shock is,

$$SIRF_i = \Lambda \Phi(\mathrm{L})^{-1} GH_i, \tag{74}$$

where H_i is the ith column of H and $i = 1, 2, 3$. This *SIRF* is estimated in the following steps.

1. Order the variables as in (73) and, using the restricted Λ in (73), estimate the seven unobserved static factors by restricted least-squares minimization of (13) as discussed in Section 2.3.1.[pp] Augment these seven factors with $OilProduction_t$ so that the vector of eight factors has one observed factor (ordered first) and the seven estimated factors. The next five variables in the named factor normalization can be chosen arbitrarily so long as they are not linearly dependent. This step yields the normalized factors $\hat{F}_t$ and factor loadings $\hat{\Lambda}$.
2. Use $\hat{F}_t$ to estimate the VAR, $\hat{F}_t = \Phi(\mathrm{L})\hat{F}_{t-1} + \eta_t$, where the normalization $G = I$ is used and the number of innovations equals the number of factors.[qq]
3. Use the VAR residuals $\hat{\eta}_t$ to estimate H using the identifying restrictions in (72). Because of the lower triangular structure of H, this can be done using the Cholesky factorization of the covariance matrix of $\hat{\eta}_t$, renormalized so that the diagonal elements of H equal one.

7.3.2.2 Additional Estimation Details

Because of the evidence discussed in Section 6 that there is a break in the DFM parameters, possibly associated with the Great moderation break data of 1984, all models were estimated over 1985q1–2014q4.

Standard errors are computed by parametric bootstrap as discussed in Section 5.1.3.

7.4 Results: "Oil Price Exogenous" Identification

The focus of this and the next section is on understanding the differences and similarities among the SDFM, FAVAR, and SVAR results. We begin in this section with the results for the "oil price exogenous" identification scheme of Section 7.2.1.

Fig. 8 presents SIRFs for selected variables with respect to the oil price shock computed using the SDFM, the FAVAR in which oil is treated as an observed factor, and the

[pp] If there were only one oil price series then Λ and the factors could be estimated as the renormalized principal components estimates in (59).

[qq] If the number of innovations were less than the number of factors, the named factor normalization of G would be the upper diagonal normalization in (61) and the reduced number of innovations could be estimated as discussed following (61).

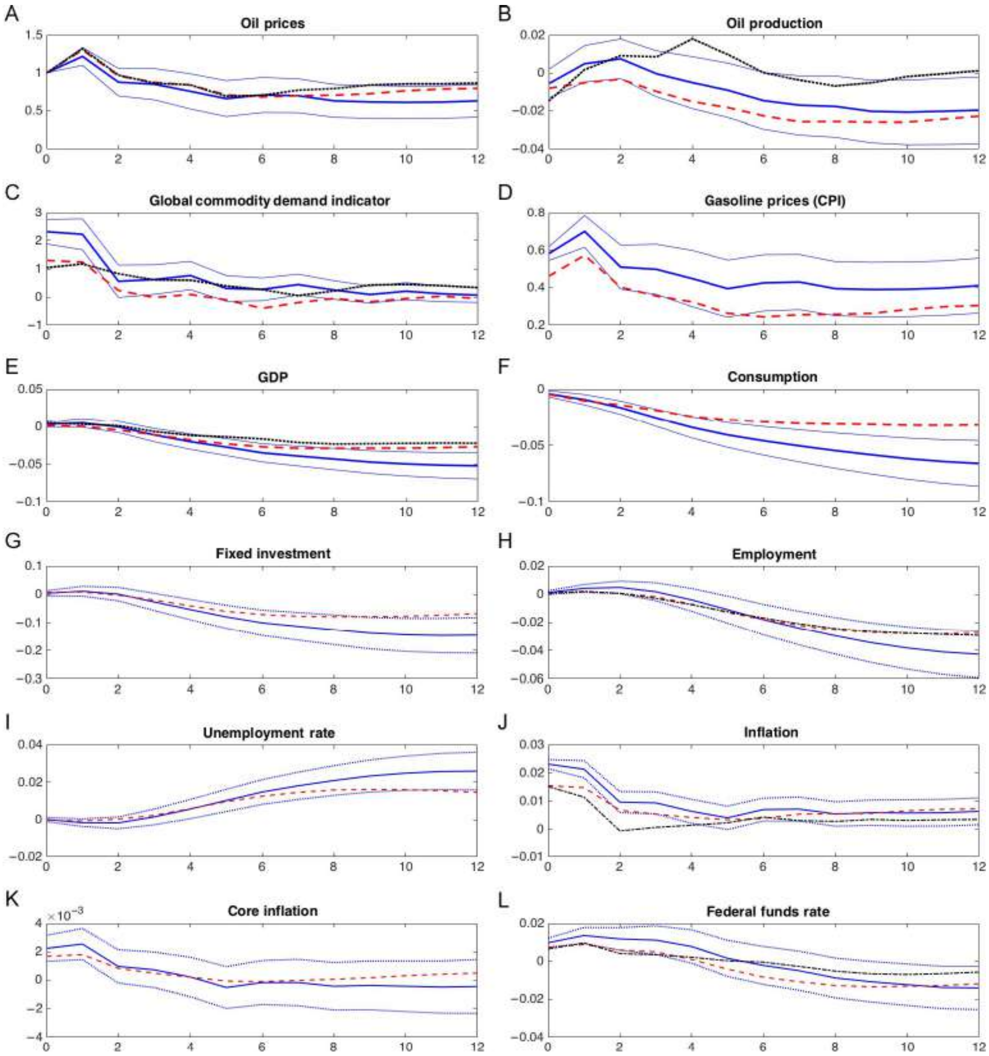


Fig. 8 Structural IRFs from the SDFM (*blue* (*dark gray* in the print version) *solid* with ±1 standard error bands), FAVAR (*red* (*gray* in the print version) *dashed*), and SVAR (*black dots*) for selected variables with respect to an oil price shock: "oil prices exogenous" identification. Units: standard deviations for Global Commodity Demand and percentage points for all other variables.

SVAR. The SVAR SIRFs are available only for the eight variables in the SVAR. The figure shows SIRFs in the log levels of the indicated variables. For example, according to the SDFM SIRFs in the *upper left panel* of Fig. 8, a unit oil price shock increases the level of oil prices by 1% on impact (this is the unit effect normalization), by additional 0.3% after one quarter, then the price of oil reverts partially and after four quarters is approximately

0.8% above its level before the shock. Equivalently, these SIRFs are cumulative SIRFs in the first differences of the variables.

The most striking feature of Fig. 8 is that all three sets of SIRFs are quite close, especially at horizons less than eight quarters. There are two main reasons for this. First, as can be seen in Fig. 7B (and in Table 3), a large fraction of the variance of the change in oil prices is explained by its common component, so the innovation in the common component in the unobserved factor DFM is similar to the innovation in the observed factor FAVAR. Second, the forecast errors for one quarter ahead changes in oil prices are similar whether they are generated using the factors or the eight-variable VAR (changes in oil prices are difficult to predict). Putting these two facts together, the innovations in oil prices (or the oil price factor) are quite similar in all three models and, under the oil price exogenous identification scheme, so are the shocks. Indeed, as shown in Table 8, the oil price shocks in the three models are similar (the smallest correlation is 0.72). In brief, the innovations in oil prices are spanned by the space of the factor price innovations.

This said, to the extent that the SDFM, FAVAR, and SVAR SIRFs differ, the FAVAR and SVAR SIRFs tend to be attenuated relative to the SDFM, that is, the effect of the oil shock in the SDFM is typically larger. This is consistent with the single observed factor in the FAVAR being measured with error in the FAVAR and SVAR models, which use a single oil price, however this effect is minor.

Concerning substantive interpretation, for the SDFM, FAVAR, and SVAR, two of the SIRFs are puzzling: the oil shock that increases oil prices is estimated to have a small effect on oil production that is statistically insignificant (negative on impact, slightly positive after one and two quarters), and a statistically significant *positive* immediate impact on global shipping activity. These two puzzling SIRFs raise the question of whether the oil price shock identified in the oil price exogenous scheme is in fact an oil supply shock, which (one would think) should be associated with a decline in oil production and either a neutral or negative impact effect on global shipping activity. These puzzling SIRFs suggest that it is important to distinguish oil price increases that arise from demand from those that stem from a shock to oil supply.

Table 7 presents six quarters ahead FEVDs for the identified shock; the results for the "oil price exogenous" identification are given in columns A for the FAVAR and SDFM. For most series, the FAVAR and SDFM decompositions are very similar, consistent with the similarity of the FAVAR and SDFM SIRFs in Fig. 8 over six quarters. The results indicate that, over the six-quarter horizon, the identified oil shocks explain no more than 10% of the variation in US GDP, fixed investment, employment, the unemployment rate, and core inflation. Curiously, the oil price shock explains a negligible fraction of the forecast errors in oil production. The series for which the FAVAR and SDFM FEVDs differ the most is the real oil price: not surprisingly, treating the oil price as the observed factor, so the innovation to the oil price is the oil shock, explains much more of the oil price forecast error than does treating the oil price factor as latent.

7.5 Results: Kilian (2009) Identification

As discussed in Section 7.2, the Kilian (2009) identification scheme identifies an oil supply shock, a global aggregate demand shock, and an oil-specific demand shock. Because there are eight innovations total in all the models examined here, this leaves five unidentified shocks (or, more precisely, a five-dimensional subspace of the innovations on which no identifying restrictions are imposed).

7.5.1 Hybrid FAVAR-SDFM

As indicated in Table 6, the innovations in the first eight principal components explain a very small fraction of the one step ahead forecast error of oil production, that is, the innovation in oil production is nearly not spanned by the space of factor innovations. Under the Kilian (2009) identification scheme, the innovation in oil production is the oil supply shock; but this oil supply shock is effectively not in the space of the eight shocks that explain the variation in the macro variables. This raises a practical problem for the SDFM because the identification scheme is asking it to identify a shock from the macro factor innovations, which is arguably not in the space of those innovations, or nearly is not in that space. In the extreme case that the common component of oil production is zero, the estimated innovation to that common component will simply be noise.

For this reason, we modify the SDFM to have a single observed factor, which is the oil production factor. The global demand shock and the oil-specific demand shock are, however, identified from the factor innovations. Thus this hybrid FAVAR–SDFM has one identified observed factor, two identified unobserved factors, and five unidentified unobserved factors.

As discussed in Section 7.2, the FAVAR treats the oil price (PPI-oil), global oil production, and the global activity index as observed factors, with five latent factors.

Table 6 Fraction of the variance explained by the eight factors at horizons $h=1$ and $h=6$ for selected variables: 1985:Q1–2014:Q4

Variable	$h=1$	$h=6$
GDP	0.60	0.80
Consumption	0.37	0.76
Fixed investment	0.38	0.76
Employment (non-ag)	0.56	0.94
Unemployment rate	0.44	0.90
PCE inflation	0.70	0.63
PCE inflation—core	0.10	0.34
Fed funds rate	0.48	0.71
Real oil price	0.74	0.78
Oil production	0.06	0.27
Global commodity shipment index	0.39	0.51
Real gasoline price	0.72	0.80

7.5.2 Results

Figs. 9–11 present SIRFs for the three identified shocks and Table 7, columns B, presents variance decompositions for six quarters ahead forecast errors. It is useful to discuss these results one shock at a time.

First consider the oil supply shock (Fig. 9). All three models identify the oil supply shock in the same way, as the one step ahead forecast error for oil supply. This variable is hard to

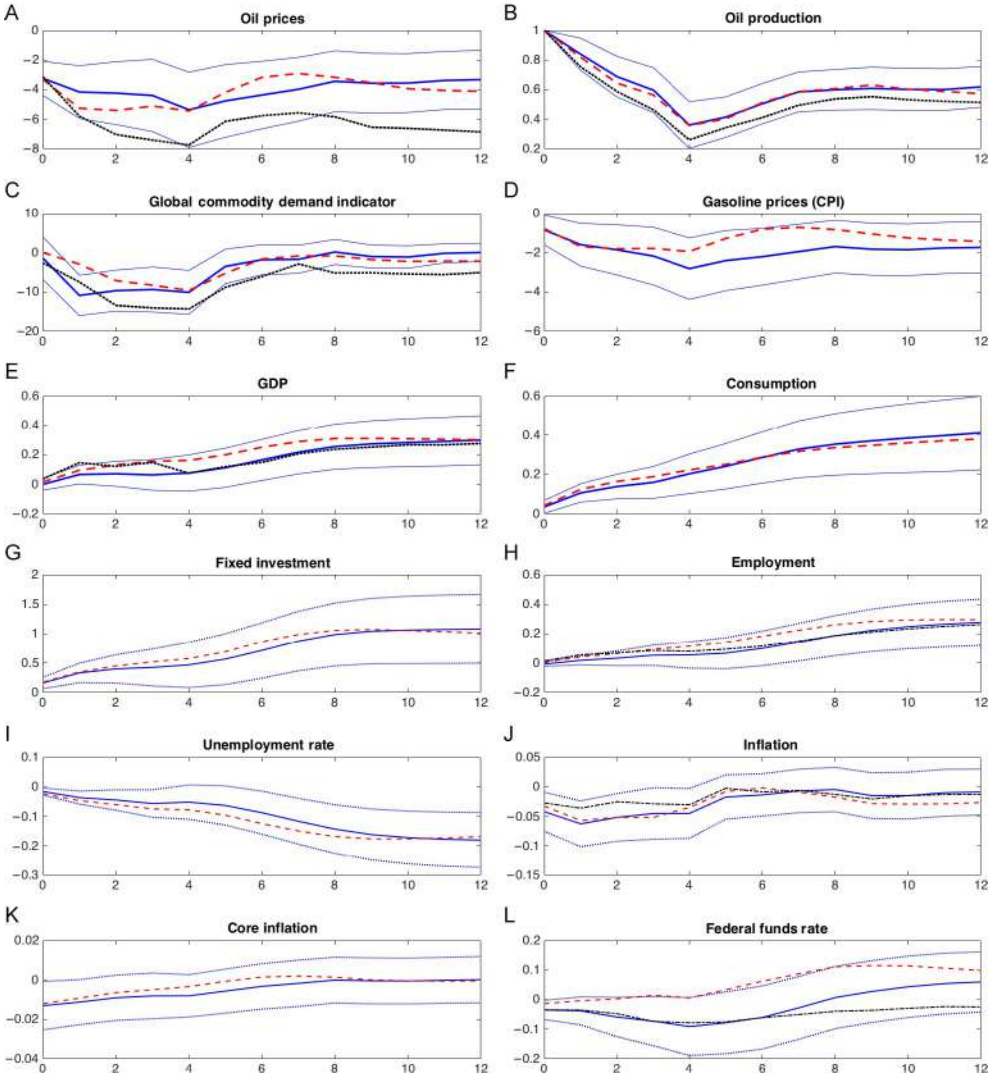


Fig. 9 Structural IRFs from the SDFM (*blue* (*dark gray* in the print version) *solid* with ±1 standard error bands), FAVAR (*red* (*gray* in the print version) *dashed*), and SVAR (*black dots*) for selected variables with respect to an oil supply shock: Kilian (2009) identification. Units: standard deviations for Global Commodity Demand and percentage points for all other variables.

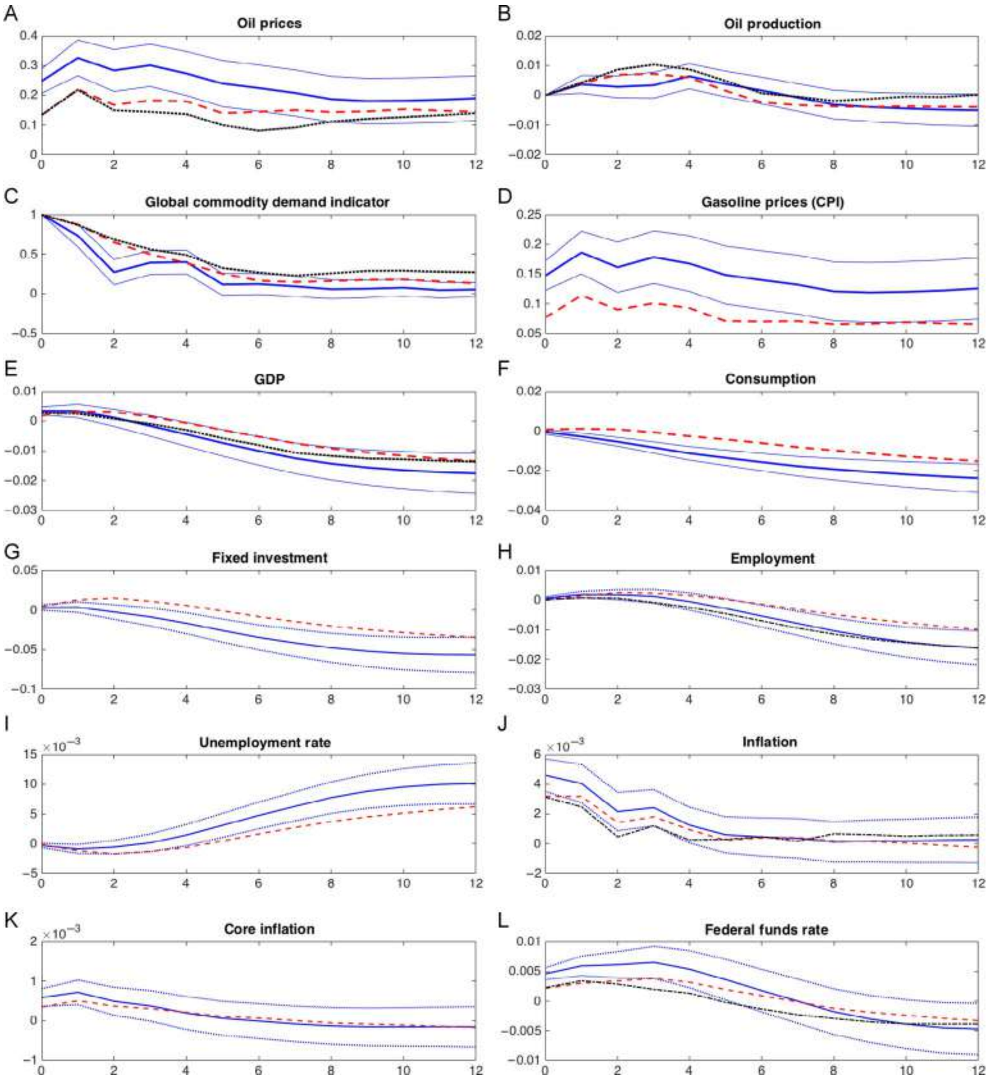


Fig. 10 Structural IRFs from the SDFM (*blue* (*dark gray* in the print version) *solid* with ±1 standard error bands), FAVAR (*red* (*gray* in the print version) *dashed*), and SVAR (*black dots*) for selected variables with respect to a global demand shock: Kilian (2009) identification. Units: standard deviations for Global Commodity Demand and percentage points for all other variables.

forecast and the forecasts, and thus forecast errors, do not substantially depend on the choice of conditioning set (lags of observed variables in the SVAR vs lags of factors in the FAVAR and SDFM). Thus the identified shocks are highly correlated (Table 8) and the SIRFs are quite similar across the three models. On a substantive note, the fraction of the variance of major macroeconomic variables explained by oil supply shocks is quite small (Table 7).

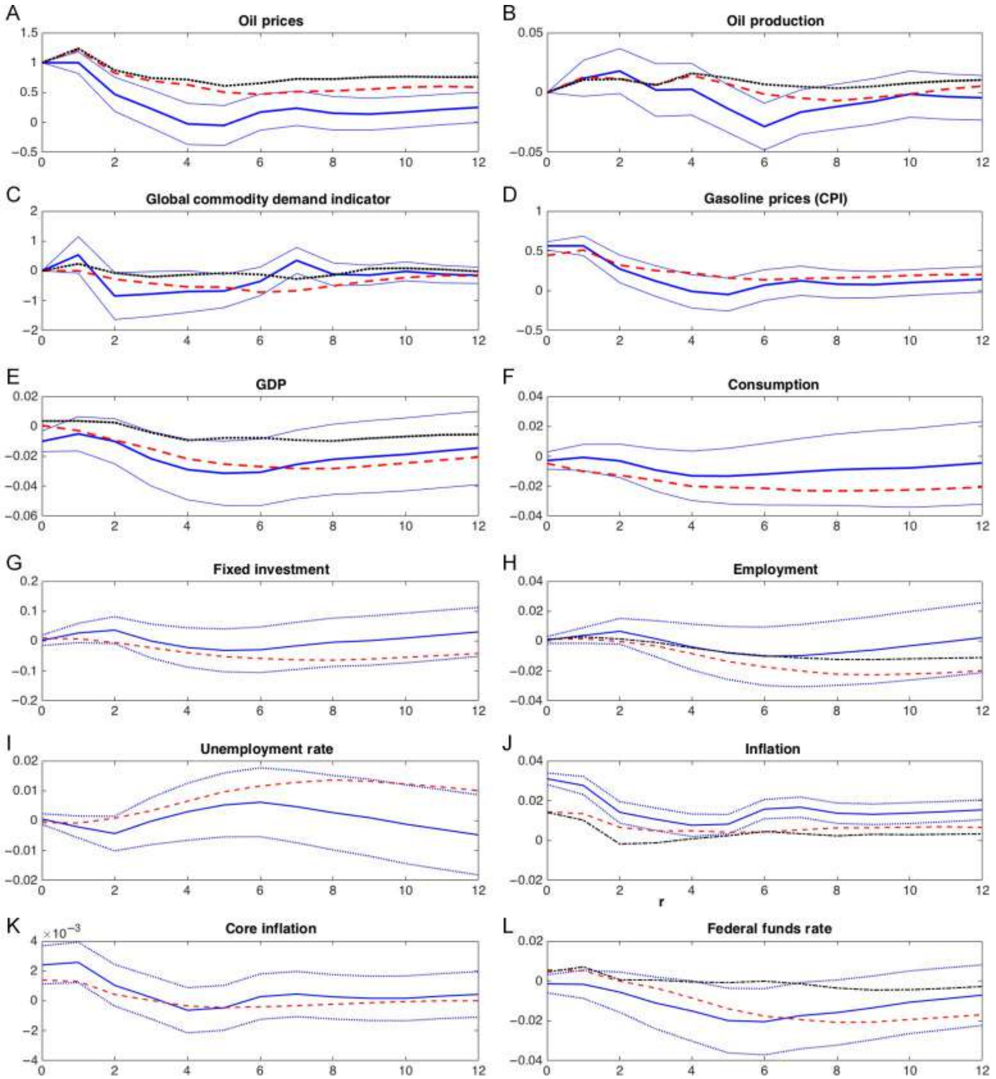


Fig. 11 Structural IRFs from the SDFM (*blue* (*dark gray* in the print version) *solid* with ±1 standard error bands), FAVAR (*red* (*gray* in the print version) *dashed*), and SVAR (*black dash–dot*) for selected variables with respect to an oil-specific demand shock: Kilian (2009) identification. Units: standard deviations for Global Commodity Demand and percentage points for all other variables.

In contrast, there are notable differences between the SDFM SIRFs for global demand shocks and the corresponding SIRFs for the FAVAR and SVAR, however the FAVAR SIRFs are quite similar to the SVAR SIRFs (Fig. 10). Broadly, the FAVAR and SVAR SIRFs are attenuated relative to the SDFM SIRFs. These features are consistent with (a) the global demand shocks—unlike the oil production shocks—being

Table 7 Forecast error variance decompositions for six periods ahead forecasts of selected variables: FAVARs and SDFMs

			B. Kilian (2009) identification					
	A. Oil price exogenous		**Oil supply**		**Global demand**		**Oil spec. demand**	
Variable	**F**	**D**	**F**	**D(O)**	**F**	**D(U)**	**F**	**D(U)**
GDP	0.07	0.07	0.04	0.01	0.02	0.04	0.09	0.04
Consumption	0.19	0.22	0.09	0.08	0.02	0.22	0.11	0.01
Fixed investment	0.04	0.04	0.05	0.04	0.03	0.04	0.03	0.01
Employment (non-ag)	0.03	0.02	0.04	0.01	0.02	0.01	0.03	0.01
Unemployment rate	0.04	0.03	0.04	0.03	0.02	0.03	0.04	0.01
PCE inflation	0.28	0.40	0.02	0.04	0.09	0.16	0.17	0.29
PCE inflation—core	0.05	0.04	0.01	0.02	0.03	0.05	0.02	0.02
Fed funds rate	0.02	0.04	0.00	0.01	0.05	0.11	0.03	0.02
Real oil price	0.81	0.53	0.14	0.10	0.22	0.44	0.42	0.09
Oil production	0.03	0.01	0.75	0.78	0.07	0.02	0.03	0.01
Global commodity shipment index	0.11	0.23	0.05	0.07	0.79	0.33	0.03	0.02
Real gasoline price	0.61	0.48	0.05	0.06	0.25	0.43	0.34	0.08

Notes: Entries are the fractions of the six periods ahead forecast error of the row variable explained by the column shock, for the "oil price exogenous" identification results (columns A) and the Kilian identification scheme (columns B). For each shock, "F" refers to the FAVAR treatment in which the factor is treated as observed and "D" refers to the SDFM treatment. In the hybrid SDFM using the Kilian (2009) identification scheme, the oil supply factor is treated as observed (the oil production variable) (D(O)) while the global demand and oil-specific demand factors are treated as unobserved (D(U)).

spanned by the space of the factor innovations, (b) the innovations in the commodity index being a noisy measure of the unobserved global factor innovations, and (c) the one step ahead forecast errors for the commodity index being close using either the factors or SVAR variables as conditioning sets. Evidence for (a) is the large fraction of the one step ahead forecast error variance of the global commodity index that is explained by the factor innovations (Table 6). But because the global commodity index is just one noisy measure of global demand, it follows from the general discussion of Section 5 that the innovations in the global commodity index in the FAVAR and SVAR models will be noisy measures of—that is, an imperfect proxy for—the innovation in global economic activity (this is point (b)). Evidence for (c) is the high correlation (0.82) between the SVAR and FAVAR estimates of the global demand shocks in Table 8.

For the oil-specific demand shock (Fig. 11), the FAVAR and SVAR SIRFs are also attenuated relative to the SDFM SIRFs. The issues associated with interpreting these differences are subtle. In addition to the oil supply and aggregate demand shocks discussed earlier, the hybrid SDFM allows for two oil price-specific shocks: one that explains some of the comovements of other macro variables, and one that is purely idiosyncratic (actually, an idiosyncratic disturbance for each oil price) which has no effect on other macro

Table 8 Correlations between identified shocks

			Oil price exogenous						Kilian (2009) identification					
			Oil price shock			Oil supply			Global demand			Oil-specific demand		
			D	F	V	D	F	V	D	F	V	D	F	V
Oil price exogenous	Oil price shock	D	1.00											
		F	0.86	1.00										
		V	0.72	0.84	1.00									
Kilian (2009) identification	Oil supply	D	−0.22	−0.24	−0.22	1.00								
		F	−0.21	−0.23	−0.23	0.95	1.00							
		V	−0.18	−0.22	−0.22	0.88	0.88	1.00						
	Global demand	D	0.70	0.63	0.56	0.00	0.06	0.07	1.00					
		F	0.45	0.35	0.31	−0.02	0.00	−0.06	0.37	1.00				
		V	0.37	0.28	0.37	0.00	−0.01	0.00	0.40	0.82	1.00			
	Oil-specific demand	D	0.63	0.50	0.43	0.00	−0.05	−0.04	0.00	0.30	0.17	1.00		
		F	0.66	0.83	0.79	0.00	0.00	0.03	0.54	0.00	0.02	0.44	1.00	
		V	0.60	0.76	0.91	−0.03	−0.04	0.00	0.48	0.00	0.00	0.39	0.88	1.00

Notes: Entries are correlations between the identified shocks. D=SDFM or hybrid SDFM, F=FAVAR, and V=SVAR.

variables. According to the FEVDs in Table 7, the oil-specific demand shock spanned by the factor innovations explains only a small amount of the forecast error in oil prices, and virtually none of the variation in major macroeconomic variables. Thus the SDFM relegates the residual variation in oil prices to the idiosyncratic disturbance, which has no effect on variables other than the oil price itself (and on PCE inflation, presumably through the oil price). In contrast, the FAVAR and SVAR have a single oil price-specific shock instead of the two in the SDFM. The single shock in the FAVAR and SVAR mix the purely idiosyncratic movements in oil prices with the oil-specific demand shock that could have broader consequences, so that this shock explains half of the six quarters ahead forecast error variance for oil prices, and one-third of that for gasoline prices, but very small amounts of the variation in other macro variables.

7.6 Discussion and Lessons

The two identification schemes provide two contrasting examples. In the "oil price exogenous" identification scheme, the oil price innovation is effectively spanned by the space of factor innovations, so it makes little difference whether oil prices are treated as an unobserved factor in a SDFM or an observed factor in a FAVAR. Moreover, because it is difficult to predict oil price changes, using the factors for that prediction or using the eight-variable VAR makes little difference. Thus, in all the models, the oil price shock is essentially the same, so the SIRFs and variance decompositions are essentially the same. For this scheme, it turns out that it matters little whether a SDFM, FAVAR, or SVAR is used.

In contrast, in the Kilian (2009) identification scheme, the results depend more sensitively on which model is used for the factors that are treated as unobserved in the SDFM. Moreover, there is the additional feature that the forecast error in oil production seems not to be spanned by the macro factor innovations, indicating both that it has little effect on the macro variables and that an attempt to treat oil production as an unobserved factor will have problems with estimation error so that it is preferable to treat oil production as an observed factor. The dependence of the results for the global activity factor and the oil-specific demand factor are consistent with the theoretical discussion in Section 5: treating those global demand and oil-specific demand as observed in a FAVAR, or as variables in a SVAR, arguably leads to measurement error in those innovations, and thus to measurement error in the IRFs. For these two shocks, it is preferable to recognize that the observed variables measure the shocks with error and thus to rely on SDFM estimates of the IRFs.

Finally, on substance, these results are consistent with the modern literature that oil supply shocks explain little of the variation of US aggregate activity since the early 1980s. Indeed, this result comes through even in the "oil price exogenous" identification scheme estimated post-1984. Instead, aggregate demand shocks are an important force in oil price movements: as estimated using the SDFM, 44% of the variance of six-quarter horizon forecast errors in oil prices is explained by global demand shocks, larger than the FAVAR estimate of 22%, consistent with the measurement error discussion earlier.

8. CRITICAL ASSESSMENT AND OUTLOOK

This section starts with some practical recommendations for empirical use of DFMs, drawn both from the literature and our own experience with these models. It then turns to a broader assessment of lessons learned from the large literature on DFMs, including touching on some remaining open methodological issues.

8.1 Some Recommendations for Empirical Practice

8.1.1 Variable Selection and Data Processing

Selection of the variables in the DFM should be guided by the purpose of the empirical application and knowledge of the data series. For the purposes of index construction, the series should have comparable scope, for example the real activity index constructed in Section 6 used the subset of real activity variables, not the full dataset. For the purposes of nowcasting, forecasting, and factor estimation, a guiding principle is that the factor innovations should span the space of the most important shocks that in turn affect the evolution of the variables of most interest.

The methods described in this chapter apply to variables that are integrated of order zero; in practice, this can require preprocessing the data to remove long-run dependence and trends. In most applications, this is done by transforming the variables to growth rates or more generally using first or second differences of the variables as appropriate. For the application in this chapter, we additionally removed remaining low-frequency swings by subtracting off a trend estimated using a lowpass filter designed to capture changes in mean growth rates at periodicities of a decade and longer. Although this step is uncommon in the literature, we believe it is important when working with US macro data because the drivers of the long-term trends in the data, such as multidecadal demographic swings, confound the short- and medium-term modeling in the DFM.

8.1.2 Parametric vs Nonparametric Methods

The parametric approach of formulating and estimating the DFM in state space has theoretical advantages: it produces the MLE and is amenable to Bayesian analysis under correct specification and it handles data irregularities such as missing observations and mixed-frequency data. But our reading of the literature and our own experience suggest that in practice the differences between parametric implementations and nonparametric implementations (principal components or other least-squares methods for estimating the factors) are slim in most applications. As discussed in Section 2.3.3, like the parametric approach, nonparametric methods can handle missing data, mixed data frequencies, and other data irregularities. The nonparametric methods have the added advantage of computational simplicity and do not require specifying a parametric dynamic model for the key step of estimating the factors. For these reasons, we therefore consider the nonparametric methods to be the appropriate default.

8.1.3 Instability

There is mounting empirical evidence that DFMs, like other time series models, can exhibit instability. This is not a surprise, for example it is well documented that changes associated with the Great Moderation go beyond reduction in variances to include changes in dynamics and reduction in predictability. Thus it is important to check for stability in DFMs, just as it is in other models with time series data. The stability tests used in this chapter are simple to implement and entail applying textbook single-equation stability tests to regressions of a single variable on the factors (other stability tests are discussed in Section 2.5.2).

One subtlety is that the PC estimator of the factors has some desirable robustness to modest amounts of time variation (see the discussion in Section 2.5.1). As a result, if there is a break in the factor loadings of some but not all of the variables, it can be appropriate to use the full sample for estimating the factors but a split sample for estimating the factor loadings, although whether this is warranted depends on the application.

8.1.4 Additional Considerations for Structural Analysis

Four sets of issues are worth stressing when a goal of the analysis is to estimate the effect of structural shocks.

The first, which is a central point of this chapter, is that identification methods developed for SVAR analysis carry over directly to SDFMs with the assistance of the unit effect normalization (32) and the named factor normalization (12).

The second concerns the potential for weak identification. This concern applies equally to SVARs, FAVARs, and SDFMs. One theme of Section 4 is that the various methods used to identify structural shocks and their IRFs in SVARs can all be interpreted as GMM, or in some cases simple instrumental variables, methods. As a result, the possibility arises that the structural parameters (the parameters of the H matrix in (20)) might be weakly identified. If so, SIRFs will in general be biased and confidence intervals will be unreliable. As of this writing, some methods for identification–robust inference in SVARs have been explored but there is not yet a comprehensive suite of tools available.

Third, inference with sign-identified SVARs, FAVARs, and SDFMs has its own challenges. As discussed in Section 4.6.2, nonlinearities in the mapping from the prior to the posterior imply that seemingly uninformative priors induce informative priors over the unidentified set. Resolving this problem is an active area of research.

The fourth issue, which arises for SDFMs but not for SVARs or FAVARs, is the possibility that the identified shock might not be spanned by the innovations of the factor loadings. This could arise either because the variables chosen for the DFM have too narrow a scope, or because the shock of interest simply has little or no macro consequence. This latter situation arose in the empirical application of Section 7, in which the factor innovations explained almost none of the forecast error in global oil production. In this case, the named factor normalization breaks down (because the latent macro factors do not include a global oil production factor so there is effectively no common component of global oil production) so the SDFM approach is not reliable. In Section 7, we addressed this problem by adopting a hybrid SDFM in which global oil production was an observed factor, which was estimated to explain very little of the post-1984 variation in US macro variables.

8.2 Assessment

We conclude by stepping back and returning to three high-level questions about whether DFMs have achieved their promise. First, has the early indication that the comovements of macro variables are well described by a small number of factors held up to scrutiny? Second, have DFMs—the first and still leading tool for "big data" analysis in macroeconomics—improved forecasts and nowcasts of macroeconomic variables? And third, do structural DFMs provide improvements over SVARs and, if so, how?

8.2.1 Do a Small Number of Factors Describe the Comovements of Macro Variables?

The repeated finding in the empirical DFM literature is that the answer is a strong yes. In the 207-variable dataset, the average R^2 of the regression of 207 variables against the eight factors is 51%. For major macroeconomic aggregates, which were not used to estimate the factors, this fraction is higher: 81% for GDP growth and 93% for the growth of nonfarm employment. This R^2 is large for other macro variables as well: 64% for the PCE deflator, 72% for the 10 year–3 month treasury spread, and 73% for the S&P 500. This high fit, for different DFMs and different variables, is evident Figs. 4, 5, and 7B in this chapter, and in many applications in this literature. This general affirmative answer does not mean that every variable is well fit by the few common factors, nor does it imply that there is no remaining common structure. But the stylized fact from Sargent and Sims (1977) of a few factors explaining a large fraction of the variation of many macro series is robust.

8.2.2 Do DFMs Improve Forecasts and Nowcasts?

Our answer is a nuanced yes. Broadly speaking, DFM forecasts are competitive with other methods, and for certain problems, such as forecasting real economic activity, DFM forecasts are in many cases the best available forecasts. For nowcasts, DFMs provide a structured and internally consistent way to handle the "ragged edge" problem with large datasets. For nowcasts, mixed-frequency methods using small datasets have proven competitive in some applications. As a practical matter, in macro forecasting and nowcasting applications DFMs are typically in the mix, sometimes provide the best forecasts, and at a minimum belong in the suite of models considered.

8.2.3 Do SDFMs Provide Improvements Over SVARs?

From the perspective of structural shock analysis, DFMs have two substantial advantages over SVARs and, in many cases, over FAVARs. First, by using many variables, they are better able to span the space of structural shocks than a low-dimensional VAR. As discussed in Section 6.4, in the US quarterly dataset the space of innovations of low-dimensional VARs does not well approximate the space of factor innovations, consistent with the individual series in the VAR having measurement error and idiosyncratic variation. This finding suggests that a method to identify shocks could fail in a SVAR because of measurement error or idiosyncratic variation, but succeed in identifying the shock in a SDFM, a general point that is consistent with the empirical results in Section 7.4.

Second, a side benefit of using many variables is that it the SDFM generates internally consistent SIRFs for a large number of variables. The SDFM separates the tasks of identifying the structural shock and estimating a SIRF for variables of interest.

ACKNOWLEDGMENTS

We thank Knut Are Aastveit, Michael Bauer, Hilde Bjørnland, Gabriel Chodorow-Reich, Domenico Giannone, Lutz Kilian, Massimiliano Marcellino, Frédéric Martinet, Serena Ng, Geert Peersman, Mikkel Plagborg-Møller, Valerie Ramey, and Leif Anders Thorsrud for helpful comments and/or discussions and Paul Ho, Eben Lazarus, and Daniel Lewis for research assistance.

REFERENCES

Aaronson, S., Cajner, T., Fallick, B., Galbis-Reig, F., Smith, C., Wascher, W., 2014. Labor force participation: recent developments and future prospects. In: Brookings Papers on Economic Activity, Fall 2014, pp. 197–275. (including discussion).

Aastveit, K.A., 2014. Oil price shocks in a data-rich environment. Energy Econ. 45, 268–279.

Aastveit, K.A., Bjørnland, H.C., Thorsrud, L.A., 2015. What drives oil prices? Emerging versus developed economies. J. Appl. Econ. 30, 1013–1028.

Aastveit, K.A., Gerdrup, K.R., Jore, A.S., Thorsrud, L.A., 2014. Nowcasting GDP in real time: a density combination approach. J. Bus. Econ. Stat. 32, 48–68.

Aguilar, O., West, M., 2000. Bayesian dynamic factor models and portfolio allocation. J. Bus. Econ. Stat. 18, 338–357.

Ahn, S.C., Horenstein, A.R., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81, 1203–1227.

Altissimo, F., Bassanetti, A., Cristadoro, R., Forni, M., Hallin, M., Lippi, M., Reichlin, L., Veronese, G., 2001. EuroCOIN: a real time coincident indicator of the euro area business cycle. In: CEPR DP3108.

Altissimo, F., Cristadoro, R., Forni, M., Lippi, M., Veronese, G., 2010. New EuroCOIN: tracking economic growth in real time. Rev. Econ. Stat. 92 (4), 1024–1034.

Amengual, D., Watson, M.W., 2007. Consistent estimation of the number of dynamic factors in a large N and T panel. J. Bus. Econ. Stat. 25, 91–96.

Andrews, D.W.K., Soares, G., 2010. Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica 78, 119–157.

Andrews, D.W.K., Stock, J.H., 2007. Inference with weak instruments. In: Blundell, R., Newey, W.K., Persson, T. (Eds.), Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, Vol. III. Cambridge University Press, Cambridge, UK.

Angelini, E., Bańbura, N., Rünstler, G., 2010. Estimating and forecasting the euro area monthly national accounts from a dynamic factor model. In: OECD Journal of Business Cycle Measurement and Analysis 7, pp. 1–22. also ECB working paper no. 953 (2008).

Arias, J.E., Rubio-Ramírez, J.F., Waggoner, D.F., 2014. Inference based on SVARs identified with sign and zero restrictions: theory and applications. Federal Reserve Bank of Atlanta. Working paper 2014-1.

Aruoba, S.B., Diebold, F.X., Scotti, C., 2009. Real-time measurement of business conditions. J. Bus. Econ. Stat. 27, 417–427.

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–172.

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.

Bai, J., Ng, S., 2006a. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. Econometrica 74, 1133–1150.

Bai, J., Ng, S., 2006b. Evaluating latent and observed factors in macroeconomics and finance. J. Econ. 131, 507–537.

Bai, J., Ng, S., 2007. Determining the number of primitive shocks in factor models. J. Bus. Econ. Stat. 25, 52–60.

Bai, J., Ng, S., 2008. Large dimensional factor analysis. Found. Trends Econ. 3 (2), 89–163.

Bai, J., Ng, S., 2013. Principal components estimation and identification of static factors. J. Econ. 176, 18–29.

Bai, J., Wang, P., 2014. Identification theory for high dimensional static and dynamic factor models. J. Econ. 178, 794–804.

Bańbura, M., Giannone, D., Modugno, M., Reichlin, L., 2013. Nowcasting and the real-time data flow. In: Elliott, G., Timmermann, A., Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting, vol. 2. Elsevier, North-Holland, pp. 195–237. Chapter 4.

Bańbura, M., Modugno, M., 2014. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. J. Appl. Econ. 29, 133–160.

Banerjee, A., Marcellino, M., 2009. Factor-augmented error correction models. In: Shephard, N., Castle, J. (Eds.), The Methodology and Practice of Econometrics: Festschrift in Honor of D.F. Hendry. Oxford University Press, Oxford, pp. 227–254. Chapter 9.

Banerjee, A., Marcellino, M., Masten, I., 2014. Forecasting with factor-augmented error correction models. Int. J. Forecast. 30, 589–612.

Banerjee, A., Marcellino, M., Masten, I., 2016. An overview of the factor-augmented error correction model. In: Koopman, S.J., Hillebrand, E. (Eds.), Dynamic Factor Models, Advances in Econometrics, 35, Emerald Group Publishing, Bingley, UK.

Bates, B., Plagborg-Møller, M., Stock, J.H., Watson, M.W., 2013. Consistent factor estimation in dynamic factor models with structural instability. J. Econ. 177, 289–304.

Baumeister, C., Hamilton, J.D., 2015a. Sign restrictions, structural vector autoregressions, and useful prior information. Econometrica 83, 1963–1999.

Baumeister, C., Hamilton, J.D., 2015b. Structural interpretation of vector autoregressions with incomplete identification: revisiting the role of oil supply and demand shocks. Manuscript. University of California, San Diego.

Baumeister, C., Kilian, L., 2016. Forty years of oil price fluctuations: why the price of oil may still surprise us. J. Econ. Perspect. 30, 139–160.

Baumeister, C., Peersman, G., 2013. Time-varying effects of oil supply shocks on the U.S. economy. Am. Econ. J. Macroecon. 5, 1–28.

Barsky, R.B., Kilian, L., 2002. Do we really know that oil caused the great stagflation? A monetary alternative. NBER Macroecon. Annu. 16, 137–183.

Baxter, M., King, R.G., 1999. Measuring business cycles: approximate band-pass filters for economic time series. Rev. Econ. Stat. 81, 575–593.

Bernanke, B.S., Boivin, J., Eliasz, P., 2005. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. Q. J. Econ. 120, 387–422.

Bernanke, B.S., Kuttner, K.N., 2005. What explains the stock market's reaction to federal reserve policy? J. Financ. 40, 1221–1257.

Bjørnland, H., Thorsrud, L.A., forthcoming. Boom or gloom? Examining the Dutch disease in two-speed economies. Econ. J.

Bjørnland, H., Thorsrud, L.A., 2015a. Commodity prices and fiscal policy design: procyclical despite a rule. CAMP working paper 5/2015.

Bjørnland, H., Thorsrud, L.A., 2015b. Applied Time Series for Macroeconomists. Gyldendal Akademisk, Oslo.

Blanchard, O.J., Galí, J., 2010. The macroeconomic effects of oil price shocks: why are the 2000s so different from the 1970s? In: Galí, J., Gertler, M.J. (Eds.), International Dimensions of Monetary Policy. University of Chicago Press for the NBER, Chicago, pp. 373–421. Chapter 7.

Blanchard, O.J., Quah, D., 1989. Dynamic effects of aggregate demand and supply disturbances. Am. Econ. Rev. 79, 655–673.

Blanchard, O.J., Watson, M.W., 1986. Are business cycles all alike? In: Gordon, R.J. (Ed.), The American Business Cycle. University of Chicago Press, Chicago.

Boivin, J., Ng, S., 2006. Are more data always better for factor analysis. J. Econ. 132, 169–194.

Breitung, J., Eickmeier, S., 2011. Testing for structural breaks in dynamic factor models. J. Econ. 163, 71–84.

Breitung, J., Tenhofen, J., 2011. GLS estimation of dynamic factor models. J. Am. Stat. Assoc. 106, 1150–1166.

Burns, A.F., Mitchell, W.C., 1946. Measuring Business Cycles. NBER, New York.

Campbell, J.R., Evans, C.L., Fisher, J.D.M., Justiniano, A., 2012. Macroeconomic effects of FOMC forward guidance. In: Brookings Papers on Economic Activity, Spring, pp. 1–80.

Carrasco, M., Rossi, B., forthcoming. In-sample inference and forecasting in misspecified factor models. J. Bus. Econ. with discussion.

Chamberlain, G., Rothschild, M., 1983. Arbitrage factor structure, and mean-variance analysis of large asset markets. Econometrica 51, 1281–1304.

Charnavoki, V., Dolado, J.J., 2014. The effects of global shocks on small commodity-exporting economies: lessons from Canada. Am. Econ. J. Macroecon. 6, 207–237.

Chen, L., Dolado, J.J., Gonzalo, J., 2014. Detecting big structural breaks in large factor models. J. Econ. 180, 30–48.

Cheng, X., Hansen, B.E., 2015. Forecasting with factor-augmented regression: a Frequentist model averaging approach. J. Econ. 186, 280–293.

Cheng, X., Liao, Z., Schorfheide, F., forthcoming. Shrinkage estimation of high-dimensional factor models with structural instabilities. Rev. Econ. Stud.

Chevillon, G., Mavroeidis, S., Zhan, Z., 2015. Robust Inference in Structural VARs with Long-Run Restrictions. manuscript. Oxford University.

Choi, I., 2012. Efficient estimation of factor models. Econ. Theory 28, 274–308.

Christiano, L.J., Eichenbaum, M.S., Evans, C.L., 1999. Monetary policy shocks: what have we learned and to what end? In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics. Elsevier Science, North-Holland, Amsterdam.

Christiano, L.J., Eichenbaum, M., Vigfusson, R., 2006. Assessing structural VARs. NBER Macroecon. Annu. 21, 1–72 (including discussion).

Clements, M.P., forthcoming. Real-time factor model forecasting and the effects of instability. Comput. Stat. Data Anal.

Cochrane, J.H., Piazzesi, M., 2002. The fed and interest rates: a high-frequency identification. Am. Econ. Rev. 92 (May), 90–95.

Cogley, Timothy, Sargent, Thomas J., 2005. Drifts and volatilities: monetary policies and outcomes in the post WWII US. Rev. Econ. Dyn. 8 (2), 262–302.

Connor, G., Korajczyk, R.A., 1986. Performance measurement with the arbitrage pricing theory. J. Financ. Econ. 15, 373–394.

Corradi, V., Swanson, N., 2014. Testing for structural stability of factor augmented forecasting models. J. Econ. 182, 100–118.

Council of Economic Advisers, 2013. Economic Report of the President 2013. U.S. Government Printing Office. Chapter 2.

Crone, T.M., Clayton-Matthews, A., 2005. Consistent economic indexes for the 50 states. Rev. Econ. Stat. 87, 593–603.

D'Agostino, A., Giannone, Domenico, 2012. Comparing alternative predictors based on large-panel factor models. Oxf. Bull. Econ. Stat. 74, 306–326.

del Negro, M., Otrok, C., 2008. Dynamic factor models with time-varying parameters: measuring changes in international business cycles: Staff Report. Federal Reserve Bank of New York. no. 326.

De Mol, C., Giannone, D., Reichlin, L., 2008. Forecasting using a large number of predictors: is Bayesian shrinkage a valid alternative to principal components? J. Econ. 146, 318–328.

Doz, C., Giannone, D., Reichlin, L., 2011. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. J. Econ. 164 (1), 188–205.

Doz, C., Giannone, D., Reichlin, L., 2012. A quasi maximum likelihood approach for large approximate dynamic factor models. Rev. Econ. Stat. 94, 1014–1024.

Durbin, J., Koopman, S.J., 2012. Time Series Analysis by State Space Methods, second ed. Oxford University Press, Oxford.

Edelstein, P., Kilian, L., 2009. How sensitive are consumer expenditures to retail energy prices? J. Mon. Econ. 56, 766–779.

Eickmeier, S., Lemke, W., Marcellino, M., 2015. A classical time varying FAVAR model: estimation, forecasting, and structural analysis. J. Royal Stat. Soc. A 178, 493–533.

Eickmeier, S., Ziegler, C., 2008. How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach. J. Forecast. 27 (3), 237–265.

Elliott, G., Müller, U., 2006. Efficient tests for general persistent time variation in regression coefficients. Rev. Econ. Stud. 73, 907–940.

Engle, R.F., Watson, M.W., 1981. A one-factor multivariate time series model of metropolitan wage rates. J. Am. Stat. Assoc. 76, 774–781.

Engle, R.F., Watson, M.W., 1983. Alternative algorithms for estimation of dynamic MIMIC, factor, and time varying coefficient regression models. J. Econ. 23, 385–400.

Evans, M.D.D., 2005. Where are we now? Real-time estimates of the macroeconomy. Int. J. Cent. Bank. 1, 127–175.

Faust, J., 1998. The robustness of identified VAR conclusions about money. Carn.-Roch. Conf. Ser. Public Policy 49, 207–244.

Faust, J., Leeper, E.M., 1997. When do long-run identifying restrictions give reliable results? J. Bus. Econ. Stat. 15, 345–353.

Faust, J., Rogers, J.H., Swanson, E., Wright, J.H., 2003. Identifying the effects of monetary policy shocks on exchange rates using high frequency data. J. Eur. Econ. Assoc. 1 (5), 1031–1057.

Faust, J., Swanson, E., Wright, J., 2004. Identifying VARs based on high-frequency futures data. J. Monet. Econ. 51 (6), 1107–1131.

Fernández-Villaverde, J., Rubio-Ramírez, J.F., Sargent, T.J., Watson, M.W., 2007. The ABCs (and Ds) of understanding VARs. Am. Econ. Rev. 97 (3), 1021–1026.

Forni, M., Gambetti, L., 2010. The dynamic effects of monetary policy: a structural factor model approach. J. Monet. Econ. 57, 203–216.

Forni, M., Giannone, D., Lippi, M., Reichlin, L., 2009. Opening the black box: structural factor models with large cross sections. Econ. Theory 25, 1319–1347.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized factor model: identification and estimation. Rev. Econ. Stat. 82, 540–554.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2005. The generalized dynamic factor model: one-sided estimation and forecasting. J. Am. Stat. Assoc. 100, 830–839.

Forni, M., Reichlin, L., 1998. Let's get real: a dynamic factor analytical approach to disaggregated business cycle. Rev. Econ. Stud. 65, 453–474.

Foroni, C., Marcellino, M., 2013. A survey of econometric methods for mixed-frequency data. Norges Bank working paper 2013-6.

Freedman, D., 1999. Wald lecture: on the Bernstein-von Mises theorem with infinite-dimensional parameters. Ann. Stat. 27 (4), 1119–1141.

Fry, R., Pagan, A., 2011. Sign restrictions in structural vector autoregressions: a critical review. J. Econ. Lit. 49 (4), 938–960.

Gafarov, B., Montiel Olea, J.L., 2015. On the Maximum and Minimum Response to an Impulse in SVARs. New York University. manuscript.

Gali, Jordi, 1999. Technology, employment, and the business cycle: do technology shocks explain aggregate fluctuations? Am. Econ. Rev. 89 (1), 249–271.

Gertler, M., Karadi, P., 2015. Monetary policy surprises, credit costs, and economic activity. Am. Econ. J. Macroecon. 7, 44–76.

Geweke, J., 1977. The dynamic factor analysis of economic time series. In: Aigner, D.J., Goldberger, A.S. (Eds.), Latent Variables in Socio-Economic Models. North-Holland, Amsterdam.

Giacomini, R., Kitagawa, T., 2014. Inference About Non-Identified SVARs. University College London. manuscript.

Giannone, D., Reichlin, L., Small, D., 2008. Nowcasting: the real-time informational content of macroeconomic data. J. Monet. Econ. 55, 665–676.

Gonçalves, S., Perron, B., 2015. Bootstrapping factor-augmented regression models. J. Econ. 182, 156–173.

Gonçalves, S., Perron, B., Djogbenou, A., forthcoming. Bootstrap prediction intervals for factor models. J. Bus. Econ. Stat.

Gordon, R.J., 2014. A new method of estimating potential real GDP growth: implications for the labor market and the debt/GDP ratio. NBER discussion paper 201423.

Gordon, R.J., 2016. The Rise and Fall of American Growth. Princeton University Press, Princeton.

Gospodinov, N., 2010. Inference in nearly nonstationary SVAR models with long-run identifying restrictions. J. Bus. Econ. Stat. 28, 1–11.

Güntner, J.H.F., 2014. How do oil producers respond to oil demand shocks. Energy Econ. 44, 1–13.

Gürkaynak, R.S., Sack, B., Swanson, E., 2005. The sensitivity of long-term interest rates to economic news: evidence and implications for macroeconomic models. Am. Econ. Rev. 95 (1), 425–436.

Hallin, M., Liška, R., 2007. The generalized dynamic factor model: determining the number of factors. J. Am. Stat. Assoc. 102, 603–617.

Hamilton, J.D., 1983. Oil and the macroeconomy since World War II. J. Polit. Econ. 91, 228–248.

Hamilton, J.D., 2003. What is an oil shock? J. Econ. 113, 363–398.

Hamilton, J.D., 2009. Causes and consequences of the oil shock of 2007–8. Brookings Papers on Economic Activity. Spring 2009, 215–261.

Hamilton, J.D., 2013. Historical oil shocks. In: Parker, R.E., Whaples, R. (Eds.), Routledge Handbook of Major Events in Economic History. Routledge Taylor and Francis Group, New York.

Hamilton, J.D., 2016. Macroeconomic regimes and regime shifts. In: Taylor, J.B. and Uhlig, H. (Eds.), Handbook of Macroeconomics, vol. 2A. Elsevier, Amsterdam, Netherlands, pp. 163–201.

Han, X., Inoue, A., 2015. Tests for parameter instability in dynamic factor models. Econ. Theory 31, 1117–1152.

Hansen, L.P., Sargent, T.J., 1991. Two difficulties in interpreting vector autoregressions. In: Hansen, L.P., Sargent, T.J. (Eds.), Rational Expectations Econometrics. Westview Press, Boulder, pp. 77–119.

Hanson, S.G., Stein, J.C., 2015. Monetary policy and long-term real rates. J. Financ. Econ. 115, 429–448.

Harvey, A.C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge, UK.

Hodrick, R.J., Prescott, E.C., 1997. Post-war U.S. business cycles: an empirical investigation. J. Money Credit Bank. 29, 1–16.

Hooker, M.A., 1996. What happened to the oil price–macroeconomy relationship? J. Mon. Econ. 38, 195–213.

Inoue, A., Kilian, L., 2013. Inference on impulse response functions in structural VARs. J. Econ. 177, 1–13.

Jungbacker, B., Koopman, S.J., van der Wel, M., 2011. Maximum likelihood estimation for dynamic factor models with missing data. J. Econ. Dyn. Control 35, 1358–1368.

Juvenal, L., Petrella, I., 2015. Speculation in the oil market. J. Appl. Econ. 30, 621–649.

Kaufmann, S., Schumacher, C., 2012. Finding relevant variables in sparse Bayesian factor models: economic applications and simulation results. Deutsche Bundesbank discussion paper 29/2012.

Kilian, L., 1998. Small-sample confidence intervals for impulse response functions. Rev. Econ. Stat. 80, 218–230.

Kilian, L., 2001. Impulse response analysis in vector autoregressions with unknown lag order. J. Forecast. 20, 161–179.

Kilian, L., 2008a. Exogenous oil supply shocks: how big are they and how much do they matter for the U.S. economy? Rev. Econ. Stat. 90, 216–240.

Kilian, L., 2008b. The economic effects of energy price shocks. J. Econ. Lit. 46, 871–909.

Kilian, L., 2009. Not all oil price shocks are alike: disentangling demand and supply shocks in the crude oil market. Am. Econ. Rev. 99, 1053–1069.

Kilian, L., 2015. Structural vector autoregressions. In: Hashimzade, N., Thornton, M.A. (Eds.), Handbook of Research Methods and Applications in Empirical Macroeconomics. Edward Elgar, Cheltenham, UK. Chapter 22.

Kilian, L., Murphy, D.P., 2012. Why agnostic sign restrictions are not enough: understanding the dynamics of oil market VAR models. J. Eur. Econ. Assoc. 10, 1166–1188.

Kilian, L., Murphy, D.P., 2014. The role of inventories and speculative trading in the global market for crude oil. J. Appl. Econ. 29, 454–478.

Kim, C.-J., Nelson, C.R., 1998. Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. Rev. Econ. Stat. 80, 188–201.

King, R.G., Plosser, C.I., Stock, J.H., Watson, M.W., 1991. Stochastic trends and economic fluctuations. Am. Econ. Rev. 81, 819–840.

Koopman, S.J., Mesters, G., forthcoming. Empirical Bayes methods for dynamic factor models. Rev. Econ. Stat.

Korobilis, D., 2014. Data-Based Priors for Vector Autoregressions with Drifting Coefficients. University of Glasgow. manuscript.

Kose, A.M., Otrok, C., Whiteman, C.H., 2003. International business cycles: world, region, and country-specific factors. Am. Econ. Rev. 93, 1216–1239.

Kuttner, K.N., 2001. Monetary policy surprises and interest rates: evidence fropm the Fed fudns futures market. J. Monet. Econ. 47, 523–544.

Lanne, M., Lütkepohl, H., Maciejowska, K., 2010. Structural vector autoregressions with Markov switching. J. Econ. Dyn. Control 34, 121–131.

Leeper, E.M., Walker, T.B., Yang, S.-C.S., 2013. Fiscal foresight and information flows. Econometrica 81, 1115–1145.

Lippi, F., Nobili, A., 2012. Oil and the macroeconomy: a quantitative structural analysis. J. Eur. Econ. Assoc. 10, 1059–1083.

Lütkepohl, H., 2015. New Introduction to Multiple Time Series Analysis. Springer-Verlag, Berlin.

Lütkepohl, H., Netšunajev, A., 2014. Disentangling demand and supply shocks in the crude oil market: how to check sign restrictions in structural VARs. J. Appl. Econ. 29, 479–496.

Lütkepohl, H., Netšunajev, A., 2015. Structural vector autoregressions with heteroskedasticity: a comparison of different volatility models. Humboldt University. SFB 649 discussion paper 2015-015.

Magnusson, L.M., Mavroeidis, S., 2014. Identification using stability restrictions. Econometrica 82, 1799–1851.

Marcellino, M., Sivec, V., 2014. Monetary, Fiscal, and Oil Shocks: Evidence Based on Mixed Frequency Structural FAVARs. forthcoming, J. Econometrics.

Mariano, R.S., Murasawa, Y., 2003. A new coincident index of business cycles based on monthly and quarterly series. J. Appl. Econ. 18, 427–443.

Mariano, R.S., Murasawa, Y., 2010. A coincident index, common factors, and monthly real GDP. Oxf. Bull. Econ. Stat. 72 (1), 27–46.

Mavroeidis, S., Plagborg-Møller, M., Stock, J.H., 2014. Empirical evidence on inflation expectations in the New Keynesian Phillips curve. J. Econ. Lit. 52, 124–188.

McCracken, M., Ng, S., 2015. FRED-MD: a monthly database for macroeconomic research. Federal Reserve Bank of St. Louis. Working paper 2015-012B.

Mertens, K., Ravn, M.O., 2013. The dynamic effects of personal and corporate income tax changes in the United States. Am. Econ. Rev. 103, 1212–1247.

Montiel Olea, J., Stock, J.H., Watson, M.W., 2016. Inference in structural VARs with external instruments. Manuscript. Harvard University.

Moon, H.R., Schorfheide, F., 2012. Bayesian and Frequentist inference in partially identified models. Econometrica 80, 755–782.

Moon, H.R., Schorfheide, F., Granziera, E., 2013. Inference for VARs Identified with Sign Restrictions. Manuscript. University of Pennsylvania.

Mumtaz, H., Surico, P., 2012. Evolving international inflation dynamics: world and country-specific factors. J. Eur. Econ. Assoc. 10 (4), 716–734.

Nakamura, E., Steinsson, J., 2015. High Frequency Identification of Monetary Non-Neutrality. Manuscript. Columbia University.

Nelson, C.R., Startz, R., 1990a. The distribution of the instrumental variable estimator and its *t* ratio when the instrument is a poor one. J. Bus. 63, S125–S140.

Nelson, C.R., Startz, R., 1990b. Some further results on the exact small sample properties of the instrumental variables estimator. Econometrica 58, 967–976.

Normandin, M., Phaneuf, L., 2004. Monetary policy shocks: testing identification conditions under time-varying conditional volatility. J. Monet. Econ. 51, 1217–1243.

Onatski, A., 2009. Testing hypotheses about the number of factors in large factor models. Econometrica 77, 1447–1479.

Onatski, A., 2010. Determining the number of factors from empirical distribution of eigenvalues. Rev. Econ. Stat. 92, 1004–1016.

Otrok, C., Whiteman, C.H., 1998. Bayesian leading indicators: measuring and predicting economic conditions in Iowa. Int. Econ. Rev. 39, 997–1014.

Pagan, A.R., Robertson, J.C., 1998. Structural models of the liquidity effect. Rev. Econ. Stat. 80, 202–217.

Peersman, G., Van Robays, I., 2009. Oil and the Euro area. Economic Policy October 2009, 605–651.

Plagborg-Møller, M., 2015. Bayesian Inference on Structural Impulse Response Functions. Harvard University. manuscript.

Quah, D., Sargent, T.J., 1993. A dynamic index model for large cross sections (with discussion). In: Stock, J.H., Watson, M.W. (Eds.), Business Cycles, Indicators, and Forecasting. University of Chicago Press for the NBER, Chicago, pp. 285–310.

Ramey, V.A., 2011. Identifying government spending shocks: it's all in the timing. Q. J. Econ. 126, 1–50.

Ramey, V.A., 2016. Macroeconomic shocks and their propagation. In: Taylor, J.B. and Uhlig, H. (Eds.), Handbook of Macroeconomics, vol. 2A. Elsevier, Amsterdam, Netherlands, pp. 71–162.

Ramey, V.A., Shapiro, M., 1998. "Costly capital reallocation and the effects of government spending" with discussion. Carn.-Roch. Conf. Ser. Public Policy 48, 145–209.

Rigobon, R., 2003. Identification through heteroskedasticity. Rev. Econ. Stat. 85, 777–792.

Rigobon, R., Sack, B., 2003. Measuring the reaction of monetary policy to the stock market. Q. J. Econ. 118, 639–669.

Rigobon, R., Sack, B., 2004. The impact of monetary policy on asset prices. J. Monet. Econ. 51, 1553–1575.

Romer, C.D., Romer, D.H., 1989. Does monetary policy matter? A new test in the spirit of Friedman and Schwartz. In: Blanchard, O.J., Fischer, S. (Eds.), NBER Macroeconomics Annual 1989. MIT Press, Cambridge, MA, pp. 121–170.

Romer, C.D., Romer, D.H., 2004. A new measure of monetary shocks: derivation and implications. Am. Econ. Rev. 94, 1055–1084.

Romer, C.D., Romer, D.H., 2010. The macroeconomic effects of tax changes: estimates based on a new measure of fiscal shocks. Am. Econ. Rev. 100 (3), 763–801.

Romer, C.D., Romer, D.H., 2015. New Evidence on the Impact of the Financial Crisis. Manuscript, University of California, Berkeley.

Rubio-Ramírez, J.F., Waggoner, D.F., Zha, T., 2010. Structural vector autoregressions: theory of identification and algorithms for inference. Rev. Econ. Stud. 77, 665–696.

Rudebusch, G.D., 1998. Do measures of monetary policy in a VAR make sense? Int. Econ. Rev. 39, 907–931.

Sargent, T.J., 1987. Macroeconomic Theory, second ed. Harcourt Brace Jovanovich Academic Press, Orlando.

Sargent, T.J., 1989. Two models of measurements and the investment accelerator. J. Polit. Econ. 97, 251–287.

Sargent, T.J., Sims, C.A., 1977. Business cycle modeling without pretending to have too much a-priori economic theory. In: Sims, C. et al., (Ed.), New Methods in Business Cycle Research. Federal Reserve Bank of Minneapolis, Minneapolis.

Sarte, P.-D.G., 1997. On the identification of structural vector autoregressions. Richmond Fed Econ. Q. 83 (3), 45–67.

Sentana, E., Fiorentini, G., 2001. Identification, estimation, and testing of conditionally heteroskedastic factor models. J. Econ. 102, 143–164.

Shapiro, M.D., Watson, M.W., 1988. Sources of business cycle fluctuations. NBER Macroecon. Annu. 3, 111–156.

Sims, C.A., 1980. Macroeconomics and reality. Econometrica 48, 1–48.

Sims, C.A., Zha, T., 1998. Bayesian methods for dynamic multivariate models. Int. Econ. Rev. 39, 949–968.

Sims, C.A., Zha, T., 1999. Error bands for impulse responses. Econometrica 67, 1113–1155.

Sims, C.A., Zha, T., 2006. Were there regime shifts in U.S. monetary policy. Am. Econ. Rev. 96 (1), 54–81.

Staiger, D., Stock, J.H., 1997. Instrumental variable regression with weak instruments. Econometrica 65 (3), 557–586.

Stock, J.H., 2008. What's new in econometrics: time series, lecture 7. Short course lectures, NBER Summer Institute, at http://www.nber.org/minicourse_2008.html.

Stock, J.H., Watson, M.W., 1989. New indexes of coincident and leading economic indicators. NBER Macroecon. Annu. 4, 351–393.

Stock, J.H., Watson, M.W., 1991. A probability model of the coincident economic indicators. In: Moore, G., Lahiri, K. (Eds.), The Leading Economic Indicators: New Approaches and Forecasting Records. Cambridge University Press, Cambridge, pp. 63–90.

Stock, J.H., Watson, M.W., 1993. A procedure for predicting recessions with leading indicators: econometric issues and recent experience. In: Stock, J.H., Watson, M.W. (Eds.), Business Cycles, Indicators and Forecasting. NBER Studies in Business Cycles, vol. 28. University of Chicago Press for the NBER, Chicago.

Stock, J.H., Watson, M.W., 1996. Evidence on structural instability in macroeconomic time series relations. J. Bus. Econ. Stat. 14, 11–30.

Stock, J.H., Watson, M.W., 1998. Median unbiased estimation of coefficient variance in a time varying parameter model. J. Am. Stat. Assoc. 93, 349–358.

Stock, J.H., Watson, M.W., 1999. Forecasting inflation. J. Monet. Econ. 44 (2), 293–335.

Stock, J.H., Watson, M.W., 2002a. Forecasting using principal components from a large number of predictors. J. Am. Stat. Assoc. 97, 1167–1179.

Stock, J.H., Watson, M.W., 2002b. Macroeconomic forecasting using diffusion indexes. J. Bus. Econ. Stat. 20, 147–162.

Stock, J.H., Watson, M.W., 2005. Implications of Dynamic Factor Models for VAR Analysis. Harvard University. manuscript.

Stock, J.H., Watson, M.W., 2009. Forecasting in dynamic factor models subject to structural instability. In: Shephard, Neil, Castle, Jennifer (Eds.), The Methodology and Practice of Econometrics: Festschrift in Honor of D.F. Hendry. Oxford University Press, Oxford. Chapter 7.

Stock, J.H., Watson, M.W., 2011. Dynamic factor models. In: Clements, M.J., Hendry, D.F. (Eds.), Oxford Handbook on Economic Forecasting. Oxford University Press, Oxford, pp. 35–59. Chapter 2.

Stock, J.H., Watson, M.W., 2012a. Disentangling the channels of the 2007–09 recession. In: Brookings Papers on Economic Activity, No. 1, pp. 81–135.

Stock, J.H., Watson, M.W., 2012b. Generalized shrinkage methods for forecasting using many predictors. J. Bus. Econ. Stat. 30, 481–493.

Stock, J.H., Watson, M.W., 2015. Core inflation and trend inflation. NBER working paper 21282.

Stock, J.H., Wright, J.H., 2000. GMM with weak identification. Econometrica 68, 1055–1096.

Uhlig, H., 2005. What are the effects of monetary policy on output? Results from an agnostic identification procedure. J. Monet. Econ. 52, 381–419.

Watson, M.W., 2006. Comment on Christiano, Eichenbaum, and Vigfusson's 'assessing structural VARs'. NBER Macroecon. Annu. 2006, 21, 97–102.

Wold, H., 1954. Causality and econometrics. Econometrica 22 (2), 162–177.

Wright, J., 2012. What does monetary policy to do long-term interest rates at the zero lower bound? Econ. J. 122, F447–F466.

Yamamoto, Y., 2012. Bootstrap inference for impulse response functions in factor-augmented vector autoregressions. Hitotsubashi University, Global COE Hi-Stat Discussion Paper Series 249.

CHAPTER 9

Solution and Estimation Methods for DSGE Models

J. Fernández-Villaverde[*], J.F. Rubio-Ramírez[†,‡,§,¶], F. Schorfheide[*]
[*]University of Pennsylvania, Philadelphia, PA, United States
[†]Emory University, Atlanta, GA, United States
[‡]Federal Reserve Bank of Atlanta, Atlanta, GA, United States
[§]BBVA Research, Madrid, Madrid, Spain
[¶]Fulcrum Asset Management, London, England, United Kingdom

Contents

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.006

Abstract

This chapter provides an overview of solution and estimation techniques for dynamic stochastic general equilibrium models. We cover the foundations of numerical approximation techniques as well as statistical inference and survey the latest developments in the field.

Keywords

Approximation error analysis, Bayesian inference, DSGE model, Frequentist inference, GMM estimation, Impulse response function matching, Likelihood-based inference, Metropolis-Hastings algorithm, Minimum distance estimation, Particle filter, Perturbation methods, Projection methods, Sequential Monte Carlo.

JEL Classification Codes

C11, C13, C32, C52, C61, C63, E32, E52

1. INTRODUCTION

The goal of this chapter is to provide an illustrative overview of the state-of-the-art solution and estimation methods for dynamic stochastic general equilibrium (DSGE) models. DSGE models use modern macroeconomic theory to explain and predict comovements of aggregate time series over the business cycle. The term *DSGE model* encompasses a broad class of macroeconomic models that spans the standard neoclassical growth model discussed in King et al. (1988) as well as New Keynesian monetary models with numerous real and nominal frictions along the lines of Christiano et al. (2005) and Smets and Wouters (2003). A common feature of these models is that decision rules of economic agents are derived from assumptions about preferences, technologies, information, and the prevailing fiscal and monetary policy regime by solving intertemporal optimization problems. As a consequence, the DSGE model paradigm delivers empirical models with a strong degree of theoretical coherence that are attractive as a laboratory for policy experiments. Modern DSGE models are flexible enough to accurately track and forecast macroeconomic time series fairly well. They have become one of the workhorses of monetary policy analysis in central banks.

The combination of solution and estimation methods in a single chapter reflects our view of the central role of the tight integration of theory and data in macroeconomics. Numerical solution methods allow us to handle the rich DSGE models that are needed for business cycle analysis, policy analysis, and forecasting. Estimation methods enable us to take these models to the data in a rigorous manner. DSGE model solution and estimation techniques are the two pillars that form the basis for understanding the behavior of aggregate variables such as GDP, employment, inflation, and interest rates, using the tools of modern macroeconomics.

Unfortunately for PhD students and fortunately for those who have worked with DSGE models for a long time, the barriers to entry into the DSGE literature are quite high. The solution of DSGE models demands familiarity with numerical approximation techniques and the estimation of the models is nonstandard for a variety of reasons, including a state-space representation that requires the use of sophisticated filtering techniques to evaluate the likelihood function, a likelihood function that depends in a complicated way on the underlying model parameters, and potential model misspecification that renders traditional econometric techniques based on the "axiom of correct specification" inappropriate. The goal of this chapter is to lower the barriers to entry into this field by providing an overview of what have become the "standard" methods of solving and estimating DSGE models in the past decade and by surveying the most recent technical developments. The chapter focuses on methods more than substantive applications, though we provide detailed numerical illustrations as well as references to applied research. The material is grouped into two parts. Part I: Solving DSGE Models (Sections 2–7) is devoted to solution techniques, which are divided into perturbation and projection techniques. Part II: Estimating DSGE Models (Sections 8–12) focuses on estimation. We cover both Bayesian and frequentist estimation and inference techniques.

PART I. SOLVING DSGE MODELS

2. SOLUTION METHODS FOR DSGE MODELS

DSGE models do not admit, except in a few cases, a closed-form solution to their equilibrium dynamics that we can derive with "paper and pencil." Instead, we have to resort to numerical methods and a computer to find an approximated solution.

However, numerical analysis and computer programming are not a part of the standard curriculum for economists at either the undergraduate or the graduate level. This educational gap has created three problems. The first problem is that many macroeconomists have been reluctant to accept the limits imposed by analytic results. The cavalier assumptions that are sometimes taken to allow for closed-form solutions may confuse more than clarify. While there is an important role for analytic results for building intuition, for understanding economic mechanisms, and for testing numerical approximations, many of the questions that DSGE models are designed to address require a quantitative answer that only numerical methods can provide. Think, for example, about the optimal response of monetary policy to a negative supply shock. Suggesting that the monetary authority should lower the nominal interest rate to smooth output is not enough for real-world advice. We need to gauge the magnitude and the duration of such an interest rate reduction. Similarly, showing that an increase in government spending raises output does not provide enough information to design an effective countercyclical fiscal package.

The second problem is that the lack of familiarity with numerical analysis has led to the slow diffusion of best practices in solution methods and little interest in issues such as the assessment of numerical errors. Unfortunately, the consequences of poor approximations can be severe. Kim and Kim (2003) document how inaccurate solutions may cause spurious welfare reversals. Similarly, the identification of parameter values may depend on the approximated solution. For instance, van Binsbergen et al. (2012) show that a

DSGE model with recursive preferences needs to be solved with higher-order approximations for all parameters of interest to be identified. Although much progress in the quality of computational work has been made in the last few years, there is still room for improvement. This is particularly important as essential nonlinearities—such as those triggered by nonstandard utility functions, time-varying volatility, or occasionally binding constraints—are becoming central to much research on the frontier of macroeconomics. Nonstandard utility functions such as the very popular Epstein–Zin preferences (Epstein and Zin, 1989) are employed in DSGE models by Tallarini (2000), Piazzesi and Schneider (2006), Rudebusch and Swanson (2011, 2012), van Binsbergen et al. (2012), and Fernández-Villaverde et al. (2014), among many others. DSGE models with time-varying volatility include Fernández-Villaverde and Rubio-Ramírez (2007), Justiniano and Primiceri (2008), Bloom (2009), Fernández-Villaverde et al. (2011, 2015b), also among many others. Occasionally binding constraints can be caused by many different mechanisms. Two popular ones are the zero lower bound (ZLB) of nominal interest rates (Eggertsson and Woodford, 2003; Christiano et al., 2011; Fernández-Villaverde et al., 2015a; Aruoba and Schorfheide, 2015; and Gust et al., 2016) and financial frictions (such as in Bernanke and Gertler, 1989; Carlstrom and Fuerst, 1997; Bernanke et al., 1999; Fernández-Villaverde, 2010; Christiano et al., 2014; and dozens of others). Inherent nonlinearities force macroeconomists to move beyond traditional linearization methods.

The third problem is that, even within the set of state-of-the-art solution methods, researchers have sometimes been unsure about the trade-offs (for example, regarding speed vs accuracy) involved in choosing among different algorithms.

Part I of the chapter covers some basic ideas about solution methods for DSGE models, discusses the trade-offs created by alternative algorithms, and introduces basic concepts related to the assessment of the accuracy of the solution. Throughout the chapter, we will include remarks with additional material for those readers willing to dig deeper into technical details.

Because of space considerations, there are important topics we cannot cover in what is already a lengthy chapter. First, we will not deal with value and policy function iteration. Rust (1996) and Cai and Judd (2014) review numerical dynamic programming in detail. Second, we will not discuss models with heterogeneous agents, a task already well accomplished by Algan et al. (2014) and Nishiyama and Smetters (2014) (the former centering on models in the Krusell and Smith (1998) tradition and the latter focusing on overlapping generations models). Although heterogeneous agent models are, indeed, DSGE models, they are often treated separately for simplicity. For the purpose of this chapter, a careful presentation of issues raised by heterogeneity will consume many pages. Suffice it to say, nevertheless, that most of the ideas in our chapter can also be applied, with suitable modifications, to models with heterogeneous agents. Third, we will not spend much time explaining the peculiarities of Markov-switching regime models and models with stochastic volatility. Finally, we will not explore how the massively parallel programming allowed by graphic processor units (GPUs) is a game-changer that opens the door to the solution of a much richer class of models. See, for example, Aldrich et al. (2011) and Aldrich (2014). Finally, for general background, the reader may want to consult a good numerical analysis book for economists. Judd (1998) is still the classic reference.

Two additional topics—a survey of the evolution of solution methods over time and the contrast between the solution of models written in discrete and continuous time—are briefly addressed in the next two remarks.

Remark 1 (The evolution of solution methods) We will skip a detailed historical survey of methods employed for the solution of DSGE models (or more precisely, for their ancestors during the first two decades of the rational expectations revolution). Instead, we will just mention four of the most influential approaches.

Fair and Taylor (1983) presented an extended path algorithm. The idea was to solve, for a terminal date sufficiently far into the future, the path of endogenous variables using a shooting algorithm. Recently, Maliar et al. (2015) have proposed a promising derivation of this idea, the extended function path (EFP), to analyze applications that do not admit stationary Markov equilibria.

Kydland and Prescott (1982) exploited the fact that the economy they were analyzing was Pareto optimal to solve the social planner's problem instead of the recursive equilibrium of their model. To do so, they substituted a linear quadratic approximation to the original social planner's problem and exploited the fast solution algorithms existing for that class of optimization problems. We will discuss this approach and its relation with perturbation in Remark 13.

King, Plosser, and Rebelo (in the widely disseminated technical appendix, not published until 2002), building on Blanchard and Kahn (1980)'s approach, linearized the equilibrium conditions of the model (optimality conditions, market clearing conditions, etc.), and solved the resulting system of stochastic linear difference equations. We will revisit linearization below by interpreting it as a first-order perturbation.

Christiano (1990) applied value function iteration to the social planner's problem of a stochastic neoclassical growth model.

Remark 2 (Discrete vs continuous time) In this chapter, we will deal with DSGE models expressed in discrete time. We will only make passing references to models in continuous time. We do so because most of the DSGE literature is in discrete time. This, however, should not be a reason to forget about the recent advances in the computation of DSGE models in continuous time (see Parra-Alvarez, 2015) or to underestimate the analytic power of continuous time. Researchers should be open to both specifications and opt, in each particular application, for the time structure that maximizes their ability to analyze the model and take it to the data successfully.

3. A GENERAL FRAMEWORK

A large number of solution methods have been proposed to solve DSGE models. It is, therefore, useful to have a general notation to express the model and its solution. This general notation will make the similarities and differences among the solution methods clear and will help us to link the different approaches with mathematics, in particular with the well-developed study of functional equations.

Indeed, we can cast numerous problems in economics in the form of a functional equation.[a] Let us define a functional equation more precisely. Let J^1 and J^2 be two functional spaces, $\Omega \subseteq \mathbb{R}^n$ (where Ω is the state space), and $\mathcal{H} : J^1 \rightarrow J^2$ be an operator between these two spaces. A *functional equation problem* is to find a function $d \subseteq J^1 : \Omega \rightarrow \mathbb{R}^m$ such that:

$$\mathcal{H}(d) = \mathbf{0}. \tag{1}$$

From Eq. (1), we can see that regular equations are nothing but particular examples of functional equations. Also, note that $\mathbf{0}$ is the space zero, different in general than the zero in the real numbers.

Examples of problems in macroeconomics that can be framed as a functional equation include value functions, Euler equations, and conditional expectations. To make this connection explicit, we introduce first the stochastic neoclassical growth model, the ancestor of all modern DSGE models. Second, we show how we can derive a functional equation problem that solves for the equilibrium dynamics of the model in terms of either a value function, an Euler equation, or a conditional expectation. After this example, the reader will be able to extend the steps in our derivations to her application.

3.1 The Stochastic Neoclassical Growth Model

We have an economy with a representative household that picks a sequence of consumption c_t and capital k_t to solve

$$\max_{\{c_t, k_{t+1}\}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t) \tag{2}$$

where $\mathbb{E}_t$ is the conditional expectation operator evaluated at period t, β is the discount factor, and u is the period utility function. For simplicity, we have eliminated the labor supply decision.

[a] Much of we have to say in this chapter is not, by any means, limited to macroeconomics. Similar problems appear in fields such as finance, industrial organization, international finance, etc.

The resource constraint of the economy is given by

$$c_t + k_{t+1} = e^{z_t} k_t^{\alpha} + (1-\delta)k_t \tag{3}$$

where δ is the depreciation rate and z_t is an AR(1) productivity process:

$$z_t = \rho z_{t-1} + \sigma \varepsilon_t, \varepsilon_t \sim N(0,1) \text{ and } |\rho| < 1. \tag{4}$$

Since both fundamental welfare theorems hold in this economy, we can jump between the social planner's problem and the competitive equilibrium according to which approach is more convenient in each moment. In general, this would not be possible, and some care is required to stay on either the equilibrium problem or the social planner's problem according to the goals of the exercise.

3.2 A Value Function

Under standard technical conditions (Stokey et al., 1989), we can transform the sequential problem defined by Eqs. (2)–(4) into a recursive problem in terms of a value function $V(k_t, z_t)$ for the social planner that depends on the two state variables of the economy, capital, k_t, and productivity, z_t. More concretely, $V(k_t, z_t)$ is defined by the Bellman operator:

$$V(k_t, z_t) = \max_{k_{t+1}} \left[u\left(e^{z_t} k_t^{\alpha} + (1-\delta)k_t - k_{t+1}\right) + \beta \mathbb{E}_t V(k_{t+1}, z_{t+1}) \right] \tag{5}$$

where we have used the resource constraint (3) to substitute for c_t in the utility function and the expectation in (5) is taken with respect to (4). This value function has an associated decision rule $g: \mathbb{R}_+ \times \mathbb{R} \rightarrow \mathbb{R}_+$:

$$k_{t+1} = g(k_t, z_t)$$

that maps the states k_t and z_t into optimal choices of k_{t+1} (and, therefore, optimal choices of $c_t = e^{z_t} k_t^{\alpha} + (1-\delta)k_t - g(k_t, z_t)$).

Expressing the model as a value function problem is convenient for several reasons. First, we have many results about the properties of value functions and the decision rules associated with them (for example, regarding their differentiability). These results can be put to good use both in the economic analysis of the problem and in the design of numerical methods. The second reason is that, as a default, we can use value function iteration (as explained in Rust, 1996 and Cai and Judd, 2014), a solution method that is particularly reliable, although often slow.

We can rewrite the Bellman operator as:

$$V(k_t, z_t) - \max_{k_{t+1}} \left[u\left(e^{z_t} k_t^{\alpha} + (1-\delta)k_t - k_{t+1}\right) + \beta \mathbb{E}_t V\ (k_{t+1}, z_{t+1}) \right] = 0,$$

for all k_t and z_t. If we define:

$$\mathcal{H}(d) = V(k_t, z_t) - \max_{k_{t+1}} \left[u\left(e^{z_t}k_t^{\alpha} + (1-\delta)k_t - k_{t+1}\right) + \beta\mathbb{E}_t V(k_{t+1}, z_{t+1})\right] = 0, \tag{6}$$

for all k_t and z_t, where $d(\cdot,\cdot) = V(\cdot,\cdot)$, we see how the operator $\mathcal{H}$, a rewrite of the Bellman operator, takes the value function $V(\cdot,\cdot)$ and obtains a zero. More precisely, Eq. (6) is an integral equation given the presence of the expectation operator. This can lead to some nontrivial measure theory considerations that we leave aside.

3.3 Euler Equation

We have outlined several reasons why casting the problem in terms of a value function is attractive. Unfortunately, this formulation can be difficult. If the model does not satisfy the two fundamental welfare theorems, we cannot easily move between the social planner's problem and the competitive equilibrium. In that case, also, the value function of the household and firms will require laws of motion for individual and aggregate state variables that can be challenging to characterize.[b]

An alternative is to work directly with the set of equilibrium conditions of the model. These include the first-order conditions for households, firms, and, if specified, government, budget and resource constraints, market clearing conditions, and laws of motion for exogenous processes. Since, at the core of these equilibrium conditions, we will have the Euler equations for the agents in the model that encode optimal behavior (with the other conditions being somewhat mechanical), this approach is commonly known as the Euler equation method (sometimes also referred to as solving the equilibrium conditions of the models). This solution strategy is extremely general and it allows us to handle non-Pareto efficient economies without further complications.

In the case of the stochastic neoclassical growth model, the Euler equation for the sequential problem defined by Eqs. (2)–(4) is:

$$u'(c_t) = \beta\mathbb{E}_t\left[u'(c_{t+1})\left(\alpha e^{z_{t+1}}k_{t+1}^{\alpha-1} + 1 - \delta\right)\right]. \tag{7}$$

Again, under standard technical conditions, there is a decision rule $g: \mathbb{R}_+ \times \mathbb{R} \rightarrow \mathbb{R}_+^2$ for the social planner that gives the optimal choice of consumption ($g^1(k_t, z_t)$) and capital tomorrow ($g^2(k_t, z_t)$) given capital, k_t, and productivity, z_t, today. Then, we can rewrite the first-order condition as:

$$u'\left(g^1(k_t, z_t)\right) = \beta\mathbb{E}_t\left[u'\left(g^1\left(g^2(k_t, z_t), z_{t+1}\right)\right)\left(\alpha e^{\rho z_t + \sigma\varepsilon_{t+1}}\left(g^2(k_t, z_t)\right)^{\alpha-1} + 1 - \delta\right)\right],$$

[b] See Hansen and Prescott (1995), for examples of how to recast a non-Pareto optimal economy into the mold of an associated Pareto-optimal problem.

for all k_t and z_t, where we have used the law of motion for productivity (4) to substitute forz_{t+1} or, alternatively:

$$\left(\begin{array}{c} u'\left(g^1(k_t,z_t)\right) \\ -\beta\mathbb{E}_t\left[u'\left(g^1\left(g^2(k_t,z_t),z_{t+1}\right)\right)\left(\alpha e^{\rho z_t+\sigma\varepsilon_{t+1}}\left(g^2(k_t,z_t)\right)^{\alpha-1}+1-\delta\right)\right]\end{array}\right)=0, \tag{8}$$

for all k_t and z_t (note the composition of functions $g^1\left(g^2(k_t,z_t),z_{t+1}\right)$ when evaluating consumption at $t+1$). We also have the resource constraint:

$$g^1(k_t,z_t)+g^2(k_t,z_t)=e^{z_t}k_t^\alpha+(1-\delta)k_t \tag{9}$$

Then, we have a functional equation where the unknown object is the decision rule g. Mapping Eqs. (8) and (9) into our operator $\mathcal{H}$ is straightforward:

$$\mathcal{H}(d)=\left\{\begin{array}{c} u'\left(g^1(k_t,z_t)\right) \\ -\beta\mathbb{E}_t\left[u'\left(g^1\left(g^2(k_t,z_t),z_{t+1}\right)\right)\left(\alpha e^{\rho z_t+\sigma\varepsilon_{t+1}}\left(g^2(k_t,z_t)\right)^{\alpha-1}+1-\delta\right)\right] \\ g^1(k_t,z_t)+g^2(k_t,z_t)-e^{z_t}k_t^\alpha-(1-\delta)k_t\end{array}\right.=\mathbf{0},$$

for all k_t and z_t, where $d=g$.

In this simple model, we could also have substituted the resource constraint in Eq. (8) and solved for a one-dimensional decision rule, but by leaving Eqs. (8) and (9), we illustrate how to handle cases where this substitution is either infeasible or inadvisable.

An additional consideration that we need to take care of is that the Euler equation (7) is only a necessary condition. Thus, after finding $g(\cdot,\cdot)$, we would also need to ensure that a transversality condition of the form:

$$\lim_{t\to\infty}\beta^t\frac{u'(c_t)}{u'(c_0)}k_t=0$$

(or a related one) is satisfied. We will describe below how we build our solution methods to ensure that this is, indeed, the case.

3.4 Conditional Expectations

We have a considerable degree of flexibility in how we specify $\mathcal{H}$ and d. For instance, if we go back to the Euler equation (7):

$$u'(c_t)=\beta\mathbb{E}_t\left[u'(c_{t+1})\left(\alpha e^{z_{t+1}}k_{t+1}^{\alpha-1}+1-\delta\right)\right]$$

we may want to find the unknown conditional expectation:

$$\mathbb{E}_t\left[u'(c_{t+1})\left(\alpha e^{z_{t+1}}k_{t+1}^{\alpha-1}+1-\delta\right)\right].$$

This may be the case either because the conditional expectation is the object of interest in the analysis or because solving for the conditional expectation avoids problems associated with the decision rule. For example, we could enrich the stochastic neoclassical growth model with additional constraints (such as a nonnegative investment: $k_{t+1} \geq (1-\delta)k_t$) that induce kinks or other undesirable properties in the decision rules. Even when those features appear, the conditional expectation (since it smooths over different realizations of the productivity shock) may still have properties such as differentiability that the researcher can successfully exploit either in her numerical solution or later in the economic analysis.[c]

To see how this would work, we can define $g: \mathbb{R}_+ \times \mathbb{R} \rightarrow \mathbb{R}_+$:

$$g(k_t, z_t) = \mathbb{E}_t\left[u'(c_{t+1})\left(\alpha e^{z_{t+1}} k_{t+1}^{\alpha-1} + 1 - \delta\right)\right] \tag{10}$$

where we take advantage of $\mathbb{E}_t$ being a function of the states of the economy. Going back to our the Euler equation (7) and the resource constraint (3), if we have access to g, we can find:

$$c_t = u'(\beta g(k_t, z_t))^{-1} \tag{11}$$

and

$$k_{t+1} = e^{z_t} k_t^{\alpha} + (1-\delta)k_t - u'(\beta g(k_t, z_t))^{-1}.$$

Thus, knowledge of the conditional expectation allows us to recover all the other endogenous variables of interest in the model. To save on notation, we write $c_t = c_{g,t}$ and $k_{t+1} = k_{g,t}$ to denote the values of c_t and k_{t+1} implied by g. Similarly:

$$c_{t+1} = c_{g,t+1} = u'(\beta g(k_{t+1}, z_{t+1}))^{-1} = u'\left(\beta g\left(k_{g,t}, z_{t+1}\right)\right)^{-1}$$

is the value of c_{t+1} implied by the recursive application of g.

To solve for g, we use its definition in Eq. (10):

$$g(k_t, z_t) = \beta \mathbb{E}_t\left[u'\left(c_{g,t+1}\right)\left(\alpha e^{\rho z_t + \sigma \varepsilon_{t+1}} k_{g,t}^{\alpha-1} + 1 - \delta\right)\right]$$

and write:

$$\mathcal{H}(d) = g(k_t, z_t) - \beta \mathbb{E}_t\left[u'\left(c_{g,t+1}\right)\left(\alpha e^{\rho z_t + \sigma \varepsilon_{t+1}} k_{g,t}^{\alpha-1} + 1 - \delta\right)\right] = \mathbf{0}$$

where $d = g$.

[c] See Fernández-Villaverde et al. (2015a) for an example. The paper is interested in solving a New Keynesian business cycle model with a zero lower bound (ZLB) on the nominal interest rate. This ZLB creates a kink on the function that maps states of the model into nominal interest rates. The paper gets around this problem by solving for consumption, inflation, and an auxiliary variable that encodes information similar to that of a conditional expectation. Once these functions have been found, the rest of the endogenous variables of the model, including the nominal interest rate, can be derived without additional approximations. In particular, the ZLB is always satisfied.

3.5 The Way Forward

We have argued that a large number of problems in macroeconomics can be expressed in terms of a functional equation problem

$$\mathcal{H}(d) = \mathbf{0}$$

and we have illustrated our assertion by building the operator $\mathcal{H}$ for a value function, for an Euler equation problem, and for a conditional expectation problem. Our examples, though, do not constitute an exhaustive list. Dozens of other cases can be constructed following the same ideas.

We will move now to study the two main families of solution methods for functional equation problems: perturbation and projection methods. Both families replace the unknown function d for an approximation $d^j(\mathbf{x}, \theta)$ in terms of the state variables of the model $\mathbf{x}$ and a vector of coefficients θ and a degree of approximation j (we are deliberately being ambiguous about the interpretation of that degree). We will use the terminology "parameters" to refer to objects describing the preferences, technology, and information sets of the model. The discount factor, risk aversion, the depreciation rate, or the persistence of the productivity shock are examples of parameters. We will call the numerical terms "coefficients" in the numerical solution. While the "parameters" usually have a clear economic interpretation associated with them, the "coefficients" will, most of the time, lack such interpretation.

Remark 3 (Structural parameters?) We are carefully avoiding the adjective "structural" when we discuss the parameters of the model. Here we follow Hurwicz (1962), who defined a "structural parameter" as a parameter that was invariant to a class of policy interventions the researcher is interested in analyzing. Many parameters of interest may not be "structural" in Hurwicz's sense. For example, the persistence of a technology shock may depend on the barriers to entry/exit in the goods and services industries and how quickly technological innovations can diffuse. These barriers may change with variations in competition policy. See a more detailed discussion on the "structural" character of parameters in DSGE models as well as empirical evidence in Fernández-Villaverde and Rubio-Ramírez (2008).

The states of the model will be determined by the structure of the model. Even if, in the words of Thomas Sargent, "finding the states is an art" (meaning both that there is no constructive algorithm to do so and that the researcher may be able to find different sets of states that accomplish the goal of fully describing the situation of the model, some of which may be more useful than the others in one context but less so in another one), determining the states is a step previous to the numerical solution of the model and, therefore, outside the purview of this chapter.

4. PERTURBATION

Perturbation methods build approximate solutions to a DSGE economy by starting from the exact solution of a particular case of the model or from the solution of a nearby model whose solution we have access to. Perturbation methods are also known as asymptotic methods, although we will avoid such a name because it risks confusion with related techniques regarding the large sample properties of estimators as the ones we will introduce in Part II of the chapter. In their more common incarnation in macroeconomics, perturbation algorithms build Taylor series approximations to the solution of a DSGE model around its deterministic steady state using implicit-function theorems. However, other perturbation approaches are possible, and we should always talk about *a* perturbation of the model instead of *the* perturbation. With a long tradition in physics and other natural sciences, perturbation theory was popularized in economics by Judd and Guu (1993) and it has been authoritatively presented by Judd (1998), Judd and Guu (2001), and Jin and Judd (2002).[d] Since there is much relevant material about perturbation problems in economics (including a formal mathematical background regarding solvability conditions, and more advanced perturbation techniques such as gauges and Padé approximants) that we cannot cover in this chapter, we refer the interested reader to these sources.

Over the last two decades, perturbation methods have gained much popularity among researchers for four reasons. First, perturbation solutions are accurate around an approximation point. Perturbation methods find an approximate solution that is inherently local. In other words, the approximated solution is extremely close to the exact, yet unknown, solution around the point where we take the Taylor series expansion. However, researchers have documented that perturbation often displays good global properties along a wide range of state variable values. See the evidence in Judd (1998); Aruoba et al. (2006) and Caldara et al. (2012). Also, as we will discuss below, the perturbed solution can be employed as an input for other solution methods, such as value function iteration. Second, the structure of the approximate solution is intuitive and easily interpretable. For example, a second-order expansion of a DSGE model includes a term that corrects for the standard deviation of the shocks that drive the stochastic dynamics of the economy. This term, which captures precautionary behavior, breaks the certainty equivalence of linear approximations that makes the discussion of welfare and risk in a linearized world challenging. Third, as we will explain below, a traditional linearization is nothing but a first-order perturbation. Hence, economists can import into perturbation theory much of their knowledge and practical experience while, simultaneously, being able to incorporate the formal results developed in applied mathematics. Fourth, thanks

[d] Perturbation approaches were already widely used in physics in the 19th century. They became a central tool in the natural sciences with the development of quantum mechanics in the first half of the 20th century. Good general references on perturbation methods are Simmonds and Mann (1997) and Bender and Orszag (1999).

to open-source software such as Dynare and Dynare++ (developed by Stéphane Adjemian, Michel Juillard, and their team of collaborators), or Perturbation AIM (developed by Eric Swanson, Gary Anderson, and Andrew Levin) higher-order perturbations are easy to compute even for practitioners less familiar with numerical methods.[e]

4.1 The Framework

Perturbation methods solve the functional equation problem:

$$\mathcal{H}(d) = \mathbf{0}$$

by specifying a Taylor series expansion to the unknown function $d : \boldsymbol{\Omega} \rightarrow \mathbb{R}^m$ in terms of the n state variables of the model $\mathbf{x}$ and some coefficients θ. For example, a second-order Taylor expansion has the form:

$$d_i^2(\mathbf{x},\theta) = \theta_{i,0} + \theta_{i,1}(\mathbf{x}-\mathbf{x}_0)' + (\mathbf{x}-\mathbf{x}_0)\theta_{i,2}(\mathbf{x}-\mathbf{x}_0)', \text{ for } i = 1,\ldots,m \tag{12}$$

where $\mathbf{x}'$ is the transpose of $\mathbf{x}$, $\mathbf{x}_0$ is the point around which we build our perturbation solution, $\theta_{i,0}$ is a scalar, $\theta_{i,1}$ is an n-dimensional vector, $\theta_{i,2}$ is a $n \times n$ matrix, and where $\theta_{i,0}$, $\theta_{i,1}$, and $\theta_{i,2}$ depend on the derivatives of d that we will find using implicit-function theorems.[f]

In comparison, the traditional linearization approach popularized by King et al. (2002) delivers a solution of the form:

$$d_i^1(\mathbf{x},\theta) = \widetilde{\theta}_{i,0} + \theta_{i,1}(\mathbf{x}-\mathbf{x}_0)'$$

where the vector $\theta_{i,1}$ is the same as in Eq. (12) and $\widetilde{\theta}_{i,0} = \theta_{i,0}$ if $j = 1$. In other words, linearization is nothing more than a first-order perturbation. Higher-order approximations generalize the structure of the linearized solution by including additional terms. Instead of being an *ad hoc* procedure (as it was sometimes understood in the 1980s and 1990s), linearization can borrow from a large set of well-established results in perturbation theory. But the direction of influence also goes in the opposite direction: we can use much of our accumulated understanding on linearized DSGE models (such as how to efficiently solve for the coefficients $\theta_{i,0}$ and $\theta_{i,1}$ and how to interpret their economic meaning) in perturbation.

[e] Dynare (a toolbox for Matlab) and Dynare++ (a stand-alone application) allow the researcher to write, in a concise and transparent language, the equilibrium conditions of a DSGE model and find a perturbation solution to it, up to the third order in Dynare and an arbitrary order in Dynare++. See http://www.dynare.org/. Perturbation AIM follows a similar philosophy, but with the additional advantage of being able to rely on Mathematica and its efficient use of arbitrary-precision arithmetic. This is important, for example, in models with extreme curvature such as those with Epstein–Zin preferences or habit persistence. See http://www.ericswanson.us/perturbation.html.

[f] Strictly speaking, the order of the approximation is given by the first nonzero or dominant term, but since in DSGE models the $\theta_{i,1}$ are typically different from zero, we can proceed without further qualifications.

Remark 4 (Linearization vs loglinearization) Linearization and, more generally, perturbation, can be performed in the level of the state variables or after applying some change of variables to any (or all) the variables of the model. Loglinearization, for example, approximates the solution of the model in terms of the log-deviations of the variables with respect to their steady state. That is, for a variable $x \in \mathbf{x}$, we define:

$$\hat{x} = \log \frac{x}{\bar{x}}$$

where $\bar{x}$ is its steady-state value, and then we find a second-order approximation:

$$d_i^2(\hat{\mathbf{x}}, \theta) = \theta_{i,0} + \theta_{i,1}(\hat{\mathbf{x}} - \hat{\mathbf{x}}_0)' + (\hat{\mathbf{x}} - \hat{\mathbf{x}}_0)\theta_{i,2}(\hat{\mathbf{x}} - \hat{\mathbf{x}}_0)', \text{ for } i = 1, \ldots, m.$$

If $\mathbf{x}_0$ is the deterministic steady state (this is more often than not the case), $\hat{\mathbf{x}}_0 = \mathbf{0}$, since for all variables $x \in \mathbf{x}$

$$\hat{x}_0 = \log \frac{x}{\bar{x}} = 0.$$

This result provides a compact representation:

$$d_i^2(\hat{\mathbf{x}}, \theta) = \theta_{i,0} + \theta_{i,1}\hat{\mathbf{x}}' + \hat{\mathbf{x}}\theta_{i,2}\hat{\mathbf{x}}', \text{ for } i = 1, \ldots, m.$$

Loglinear solutions are easy to read (the loglinear deviation is an approximation of the percentage deviation with respect to the steady state) and, in some circumstances, they can improve the accuracy of the solution. We will revisit the change of variables later in the chapter.

Before getting into technical details of how to implement perturbation methods, we will briefly distinguish between regular and singular perturbations. A regular perturbation is a situation where a *small* change in the problem induces a *small* change in the solution. An example is a standard New Keynesian model (Woodford, 2003). A small change in the standard deviation of the monetary policy shock will lead to a small change in the properties of the equilibrium dynamics (ie, the standard deviation and autocorrelation of variables such as output or inflation). A singular perturbation is a situation where a *small* change in the problem induces a *large* change in the solution. An example can be an excess demand function. A small change in the excess demand function may lead to an arbitrarily large change in the price that clears the market.

Many problems involving DSGE models will result in regular perturbations. Thus, we will concentrate on them. But this is not necessarily the case. For instance, introducing a new asset in an incomplete market model can lead to large changes in the solution. As researchers pay more attention to models with financial frictions and/or market incompleteness, this class of problems may become common. Researchers will need to learn more about how to apply singular perturbations. See, for pioneering work, Judd and Guu (1993), and a presentation of bifurcation methods for singular problems in Judd (1998).

4.2 The General Case

We are now ready to deal with the details of how to implement a perturbation. We present first the general case of how to find a perturbation solution of a DSGE model by (1) using the equilibrium conditions of the model and (2) by finding a higher-order Taylor series approximation. Once we have mastered this task, it would be straightforward to extend the results to other problems, such as the solution of a value function, and to conceive other possible perturbation schemes. This section follows much of the structure and notation of section 3 in Schmitt-Grohé and Uribe (2004).

We start by writing the equilibrium conditions of the model as

$$\mathbb{E}_t \mathcal{H}(\mathbf{y}, \mathbf{y}', \mathbf{x}, \mathbf{x}') = 0, \tag{13}$$

where $\mathbf{y}$ is an $n_y \times 1$ vector of controls, $\mathbf{x}$ is an $n_x \times 1$ vector of states, and $n = n_x + n_y$. The operator $\mathcal{H} : \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^n$ stacks all the equilibrium conditions, some of which will have expectational terms, some of which will not. Without loss of generality, and with a slight change of notation with respect to Section 3, we place the conditional expectation operator outside $\mathcal{H}$: for those equilibrium conditions without expectations, the conditional expectation operator will not have any impact. Moving $\mathbb{E}_t$ outside $\mathcal{H}$ will make some of the derivations below easier to follow. Also, to save on space, when there is no ambiguity, we will employ the recursive notation where x represents a variable at period t and x' a variable at period $t + 1$.

It will also be convenient to separate the endogenous state variables (capital, asset positions, etc.) from the exogenous state variables (productivity shocks, preference shocks, etc.). In that way, it will be easier to see the variables on which the perturbation parameter that we will introduce below will have a direct effect. Thus, we partition the state vector $\mathbf{x}$ (and taking transposes) as

$$\mathbf{x} = [\mathbf{x}_1'; \mathbf{x}_2']'.$$

where $\mathbf{x}_1$ is an $(n_x - n_\epsilon) \times 1$ vector of endogenous state variables and $\mathbf{x}_2$ is an $n_\epsilon \times 1$ vector of exogenous state variables. Let $\tilde{n} = n_x - n_\epsilon$.

4.2.1 Steady State

If we suppress the stochastic component of the model (more details below), we can define the deterministic steady-state of the model as vectors $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ such that:

$$\mathcal{H}(\bar{\mathbf{y}}, \bar{\mathbf{y}}, \bar{\mathbf{x}}, \bar{\mathbf{x}}) = 0. \tag{14}$$

The solution $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ of this problem can often be found analytically. When this cannot be done, it is possible to resort to a standard nonlinear equation solver.

The previous paragraph glossed over the possibility that the model we are dealing with either does not have a steady state or that it has several of them (in fact, we can even have a continuum of steady states). Given our level of abstraction with the definition of Eq. (13), we cannot rule out any of these possibilities. Galor (2007) discusses in detail the existence and stability (local and global) of steady states in discrete time dynamic models.

A case of interest is when the model, instead of having a steady state, has a balanced growth path (BGP): that is, when the variables of the model (with possibly some exceptions such as labor) grow at the same rate (either deterministic or stochastic). Given that perturbation is an inherently local solution method, we cannot deal directly with solving such a model. However, on many occasions, we can rescale the variables x_t in the model by the trend μ_t:

$$\hat{x}_t = \frac{x_t}{\mu_t}$$

to render them stationary (the trend itself may be a complicated function of some technological processes in the economy, as when we have both neutral and investment-specific technological change; see Fernández-Villaverde and Rubio-Ramírez, 2007). Then, we can undertake the perturbation in the rescaled variable $\hat{x}_t$ and undo the rescaling when using the approximated solution for analysis and simulation.[g]

Remark 5 (Simplifying the solution of $(\bar{x}, \bar{y})$***)*** Finding the solution $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ can often be made much easier by using two "tricks." One is to substitute some of the variables away from the operator $\mathcal{H}(\cdot)$ and reduce the system from being one of n equations in n unknowns into a system of $n' < n$ equations in n' unknowns. For example, if we have a law of motion for capital involving capital next period, capital next period, and investment:

$$k_{t+1} = (1-\delta)k_t + i_t$$

we can substitute out investment throughout the whole system just by writing:

$$i_t = k_{t+1} - (1-\delta)k_t.$$

Since the complexity of solving a nonlinear system of equations grows exponentially in the dimension of the problem (see Sikorski, 1985, for classic results on computational complexity), even a few substitutions can produce considerable improvements.

A second possibility is to select parameter values to pin down one or more variables of the model and then to solve all the other variables as a function of the fixed variables. To illustrate this point, let us consider a simple stochastic neoclassical growth model with a representative household with utility function:

[g] This rescaling is also useful with projection methods since they need a bounded domain of the state variables.

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left(\log c_t - \psi \frac{l_t^{1+\eta}}{1+\eta} \right)$$

where the notation is the same as in Section 3 and a production function:

$$output_t = A_t k_t^{\alpha} l_t^{1-\alpha}$$

where A_t is the productivity level and a law of motion for capital:

$$k_{t+1} = output_t + (1-\delta) k_t - c_t.$$

This model has a static optimality condition for labor supply of the form:

$$\psi c_t l_t^{\eta} = w_t$$

where w_t is the wage. Since with the log-CRRA utility function that we selected l_t does not have a natural unit, we can fix its deterministic steady-state value, for example, $\bar{l} = 1$. This normalization is as good as any other and the researcher can pick the normalization that best suits her needs.

Then, we can analytically solve the rest of the equilibrium conditions of the model for all other endogenous variables as a function of $\bar{l} = 1$. After doing so, we return to the static optimality condition to obtain the value of the parameter ψ as:

$$\psi = \frac{\bar{w}}{\bar{c}\bar{l}^{\eta}} = \frac{\bar{w}}{\bar{c}}$$

where $\bar{c}$ and $\bar{w}$ are the deterministic steady-state values of consumption and wage, respectively. An alternative way to think about this procedure is to realize that it is often easier to find parameter values that imply a particular endogenous variable value than to solve for those endogenous variable values as a function of an arbitrary parameter value.

Although not strictly needed to find $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, other good practices include picking units that make algebraic and numerical computations convenient to handle. For example, we can pick units to make $\overline{output} = 1$. Again, in the context of the stochastic neoclassical growth model, we will have:

$$\overline{output} = 1 = \overline{Ak}^{\alpha} \bar{l}^{1-\alpha} = \overline{Ak}^{\alpha}.$$

Then, we can find:

$$\bar{A} = \frac{1}{\bar{k}^{\alpha}}$$

and wages:

$$\bar{w} = (1-\alpha) \frac{\overline{output}}{\bar{l}} = 1 - \alpha.$$

Going back to the intertemporal Euler equation:

$$\frac{1}{\bar{c}}=\frac{1}{\bar{c}}\beta(1+\bar{r}-\delta)$$

where r is the rental rate of capital and δ is depreciation, we find:

$$\bar{r}=\frac{1}{\beta}-1+\delta.$$

Since:

$$\bar{r}=\alpha\frac{\overline{output}}{\bar{k}}=\frac{\alpha}{\bar{k}}$$

we get:

$$\bar{k}=\frac{\alpha}{\frac{1}{\beta}-1+\delta}$$

and:

$$\bar{c}=\overline{output}-\delta\bar{k}=1-\delta\frac{\alpha}{\frac{1}{\beta}-1+\delta},$$

from which:

$$\psi=\frac{\overline{w}}{\bar{c}}=\frac{1-\alpha}{1-\delta\frac{\alpha}{\frac{1}{\beta}-1+\delta}}$$

In this example, two judicious choices of units ($\bar{l}=\overline{output}=1$) render the solution of the deterministic steady state a straightforward exercise. While the deterministic steady state of more complicated models would be harder to solve, experience suggests that following the advice in this remark dramatically simplifies the task in many situations.

The deterministic steady state $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ is different from a fixed point $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ of (13):

$$\mathbb{E}_t\mathcal{H}(\hat{\mathbf{y}},\hat{\mathbf{y}},\hat{\mathbf{x}},\hat{\mathbf{x}})=0,$$

because in the former case we eliminate the conditional expectation operator while in the latter we do not. The vector $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ is sometimes known as the stochastic steady state (although, since we find the idea of mixing the words "stochastic" and "steady state" in the same term confusing, we will avoid that terminology).

4.2.2 Exogenous Stochastic Process

For the exogenous stochastic variables, we specify a stochastic process of the form:

$$\mathbf{x}_2'=\mathbf{C}(\mathbf{x}_2)+\sigma\eta_\epsilon\epsilon' \tag{15}$$

where $\mathbf{C}$ is a potentially nonlinear function. At our current level of abstraction, we are not imposing much structure on $\mathbf{C}$, but in concrete applications, we will need to add more constraints. For example, researchers often assume that all the eigenvalues of the Hessian matrix of $\mathbf{C}$ evaluated at the steady state $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ lie within the unit circle. The vector ϵ' contains the n_ϵ exogenous zero-mean innovations. Initially, we only assume that ϵ' is independent and identically distributed with finite second moments, meaning that we do not rely on any distributional assumption. Thus, the innovations may be non-Gaussian. This is denoted by $\epsilon' \sim iid(\mathbf{0}, \mathbf{I})$. Additional moment restrictions will be introduced as needed in each concrete application. Finally, η_ϵ is an $n_\epsilon \times n_\epsilon$ matrix that determines the variances-covariances of the innovations, and $\sigma \geq 0$ is a perturbation parameter that scales η_ϵ.

Often, it will be the case that $\mathbf{C}$ is linear:

$$x_2' = Cx_2 + \sigma\eta_\epsilon\epsilon'$$

where C is an $n_\epsilon \times n_\epsilon$ matrix, with all its eigenvalues with modulus less than one.

Remark 6 (Linearity of innovations) The assumption that innovations enter linearly in Eq. (15) may appear restrictive, but it is without loss of generality. Imagine that instead of Eq. (15), we have:

$$\mathbf{x}_{2,t} = \mathbf{D}(\mathbf{x}_{2,t-1}, \sigma\eta_\epsilon\epsilon_t).$$

This richer structure can be handled by extending the state vector by incorporating the innovations ϵ in the state vector. In particular, let

$$\tilde{\mathbf{x}}_{2,t} = \begin{bmatrix} \mathbf{x}_{2,t-1} \\ \epsilon_t \end{bmatrix}$$

and

$$\tilde{\epsilon}_{t+1} = \begin{bmatrix} \mathbf{0}_{n_\epsilon \times 1} \\ \epsilon_{t+1} \end{bmatrix}$$

Then, we can write

$$\mathbf{x}_{2,t} = \tilde{\mathbf{D}}\ (\tilde{\mathbf{x}}_{2,t}, \sigma\eta_\epsilon).$$

The new stochastic process is given by:

$$\begin{bmatrix} \mathbf{x}_{2,t} \\ \epsilon_{t+1} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{D}}\ (\tilde{\mathbf{x}}_{2,t}, \sigma\eta_\epsilon) \\ 0 \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{n_\epsilon \times 1} \\ \epsilon_{t+1} \end{bmatrix}$$

where $\mathbf{u}_{t+1} \sim iid(\mathbf{0}, \mathbf{I})$ or, switching back to the recursive notation:

$$\tilde{\mathbf{x}}_2' = \mathbf{C}(\tilde{\mathbf{x}}_2) + \tilde{\epsilon}'$$

To illustrate this point, we use the popular case of time-varying volatility, which, it has been argued, is of considerable importance to understand the dynamics of aggregate variables (see Bloom, 2009 and Fernández-Villaverde et al., 2011). Imagine that we have a stochastic volatility process for productivity a_t:

$$\log a_t = \rho_a \log a_{t-1} + \lambda_t \upsilon_t, \upsilon_t \sim N(0,1)$$

where λ_t is the standard deviation of the innovation υ_t. The standard deviation follows another autoregressive process:

$$\log \lambda_t = \overline{\lambda} + \rho_\lambda \log \lambda_{t-1} + \psi \eta_t, \eta_t \sim N(0,1).$$

To fit this system into our notation, we only need to define:

$$\tilde{\mathbf{x}}_{2,t} = \begin{bmatrix} \log a_{t-1} \\ \log \lambda_{t-1} \\ \upsilon_t \\ \eta_t \end{bmatrix}$$

and

$$\tilde{\epsilon}_{t+1} = \begin{bmatrix} \mathbf{0}_{2\times 1} \\ \epsilon_{t+1} \end{bmatrix}.$$

Note, also, how the perturbation parameter controls both the innovation υ_t and its standard deviation λ_t.

Perturbation methods are well suited to the solution of models with time-varying volatility because these models have a richness of state variables: for each stochastic process, we need to keep track of the level of the process and its variance. The projection methods that we will describe in the next section will have problems dealing with this large number of state variables.

Only one perturbation parameter appears in Eq. (15), even if we have a model with many innovations. The matrix η_ϵ takes account of relative sizes (and comovements) of the different innovations. If we set $\sigma = 0$, we have a deterministic model.

Remark 7 (Perturbation parameter) In the main text, we introduced the perturbation parameter as controlling the standard deviation of the stochastic process:

$$\mathbf{x}_2' = \mathbf{C}(\mathbf{x}_2) + \sigma \eta_\epsilon \epsilon'.$$

However, we should not hew too closely to this choice. First, there may be occasions where placing the perturbation in another parameter could offer better accuracy and/or deeper insights into the behavior of the model. For example, in models with Epstein–Zin preferences, Hansen et al. (2008) perform a perturbation around an elasticity of intertemporal substitution equal to 1. Also, the choice of perturbation would be different in a continuous time model, where it is usually more convenient to control the variance.

We depart from Samuelson (1970) and Jin and Judd (2002), who impose a bounded support for the innovations of the model. By doing so, these authors avoid problems with the stability of the simulations coming from the perturbation solution that we will discuss below. Instead, we will introduce pruning as an alternative strategy to fix these problems.

4.2.3 Solution of the Model

The solution of the model will be given by a set of decision rules for the control variables

$$\mathbf{y} = \mathbf{g}(\mathbf{x}; \sigma), \tag{16}$$

and for the state variables

$$\mathbf{x}' = \mathbf{h}(\mathbf{x}; \sigma) + \sigma \eta \epsilon', \tag{17}$$

where $\mathbf{g}$ maps $\mathbb{R}^{n_x} \times \mathbb{R}^+$ into R^{n_y} and $\mathbf{h}$ maps $\mathbb{R}^{n_x} \times \mathbb{R}^+$ into $\mathbb{R}^{n_x}$. Note our timing convention: controls depend on current states, while states next period depend on states today and the innovations tomorrow. By defining additional state variables that store the information of states with leads and lags, this structure is sufficiently flexible to capture rich dynamics. Also, we separate states $\mathbf{x}$ and the perturbation parameter σ by a semicolon to emphasize the difference between both elements.

The $n_x \times n_\epsilon$ matrix η is:

$$\eta = \begin{bmatrix} \emptyset \\ \eta_\epsilon \end{bmatrix}$$

where the first n_x rows come from the states today determining the endogenous states tomorrow and the last n_ϵ rows come from the exogenous states tomorrow depending on the states today and the innovations tomorrow.

The goal of perturbation is to find a Taylor series expansion of the functions $\mathbf{g}$ and $\mathbf{h}$ around an appropriate point. A natural candidate for this point is the deterministic steady state, $\mathbf{x}_t = \bar{\mathbf{x}}$ and $\sigma = 0$. As we argued above, we know how to compute this steady state and, consequently, how to evaluate the derivatives of the operator $\mathcal{H}(\cdot)$ that we will require.

First, note by the definition of the deterministic steady state (14) we have that

$$\bar{\mathbf{y}} = \mathbf{g}(\bar{\mathbf{x}}; 0) \tag{18}$$

and

$$\bar{\mathbf{x}} = \mathbf{h}(\bar{\mathbf{x}}; 0). \tag{19}$$

Second, we plug-in the unknown solution on the operator $\mathcal{H}$ and define the new operator $F : \mathbb{R}^{n_x+1} \rightarrow \mathbb{R}^n$:

$$F(\mathbf{x};\sigma) \equiv \mathbb{E}_t \mathcal{H}(\mathbf{g}(\mathbf{x};\sigma), \mathbf{g}(\mathbf{h}(\mathbf{x};\sigma) + \sigma\eta\epsilon', \sigma), \mathbf{x}, \mathbf{h}(\mathbf{x};\sigma) + \sigma\eta\epsilon') = \mathbf{0}.$$

Since $F(\mathbf{x};\sigma) = 0$ for any values of $\mathbf{x}$ and σ , any derivatives of F must also be zero:

$$F_{x_i^k \sigma^j}(\mathbf{x};\sigma) = 0, \forall \mathbf{x}, \sigma, i, k, j,$$

where $F_{x_i^k \sigma^j}(\mathbf{x};\sigma)$ is the derivative of F with respect to the i-th component x_i of $\mathbf{x}$ taken k times and with respect to σ taken j times evaluated at $(\mathbf{x};\sigma)$. Intuitively, the solution of the model must satisfy the equilibrium conditions for all possible values of the states and σ. Thus, any change in the values of the states or of σ must still keep the operator F exactly at $\mathbf{0}$. We will exploit this important fact repeatedly.

Remark 8 (Existence of derivatives) We will assume, without further discussion, that all the relevant derivatives of the operator F exist in a neighborhood of $\bar{\mathbf{x}}$. These differentiability assumptions may be hard to check in concrete applications and more research in the area would be welcomed (see the classic work of Santos, 1992). However, the components that enter into F (utility functions, production functions, etc.) are usually smooth when we deal with DSGE models, which suggest that the existence of these derivatives is a heroic assumption (although the examples in Santos, 1993 are a cautionary sign). Judd (1998, p. 463) indicates, also, that if the derivative conditions were violated, our computations would display telltale signs that would alert the researcher to the underlying problems.

The derivative assumption, however, traces the frontiers of problems suitable for perturbation: if, for example, some variables are discrete or the relevant equilibrium conditions are nondifferentiable, perturbation cannot be applied. Two caveats about the previous statement are, nevertheless, worthwhile to highlight. First, the presence of expectations often transforms problems that appear discrete into continuous ones. For example, deciding whether or not to go to college can be "smoothed out" by a stochastic shock to college costs or by an effort variable that controls how hard the prospective student is applying to college or searching for funding. Second, even if the derivative assumption breaks down and the perturbation solution is not valid, it may still be an excellent guess for another solution method.

Remark 9 (Taking derivatives) The previous exposition demonstrates the central role of derivatives in perturbation methods. Except for simple examples, manually calculating these derivatives is too onerous. Thus, researchers need to rely on computers. A first possibility, numerical derivatives, is inadvisable Judd (1998, chapter 7). The errors created by numerical derivatives quickly accumulate and, after the second or third derivative, the perturbation solution is too contaminated by them to be of any real use. A second possibility is to exploit software that takes analytic derivatives, such as Mathematica or the symbolic toolbox of Matlab. This route is usually straightforward, but it may slow down the computation and require an inordinate amount of memory. A third final alternative is to employ automatic differentiation, a technique that takes advantage of the application of the chain rule to a series of elementary arithmetic operations and functions (for how automatic differentiation can be applied to DSGE models, see Bastani and Guerrieri, 2008).

4.2.4 First-Order Perturbation

A first-order perturbation approximates $\mathbf{g}$ and $\mathbf{h}$ around $(\mathbf{x};\sigma)=(\bar{\mathbf{x}};0)$ as:

$$\begin{aligned}\mathbf{g}(\mathbf{x};\sigma) &= \mathbf{g}(\bar{\mathbf{x}};0)+\mathbf{g}_{\mathbf{x}}(\bar{\mathbf{x}};0)(\mathbf{x}-\bar{\mathbf{x}})'+\mathbf{g}_{\sigma}(\bar{\mathbf{x}};0)\sigma\\ \mathbf{h}(\mathbf{x};\sigma) &= \mathbf{h}(\bar{\mathbf{x}};0)+\mathbf{h}_{\mathbf{x}}(\bar{\mathbf{x}};0)(\mathbf{x}-\bar{\mathbf{x}})'+\mathbf{h}_{\sigma}(\bar{\mathbf{x}};0)\sigma\end{aligned}$$

where $\mathbf{g}_{\mathbf{x}}$ and $\mathbf{h}_{\mathbf{x}}$ are the gradients of $\mathbf{g}$ and $\mathbf{h}$, respectively (including only the partial derivatives with respect to components of $\mathbf{x}$) and $\mathbf{g}_{\sigma}$ and $\mathbf{h}_{\sigma}$ the derivatives of $\mathbf{g}$ and $\mathbf{h}$ with respect to the perturbation parameter σ.

Using Eqs. (18) and (19), we can write

$$\begin{aligned}\mathbf{g}(\mathbf{x};\sigma)-\bar{\mathbf{y}} &= \mathbf{g}_{\mathbf{x}}(\bar{\mathbf{x}};0)(\mathbf{x}-\bar{\mathbf{x}})'+\mathbf{g}_{\sigma}(\bar{\mathbf{x}};0)\sigma\\ \mathbf{h}(\mathbf{x};\sigma)-\bar{\mathbf{x}} &= \mathbf{h}_{\mathbf{x}}(\bar{\mathbf{x}};0)(\mathbf{x}-\bar{\mathbf{x}})'+\mathbf{h}_{\sigma}(\bar{\mathbf{x}};0)\sigma.\end{aligned}$$

Since we know $(\bar{\mathbf{x}},\bar{\mathbf{y}})$, we only need to find $\mathbf{g}_{\mathbf{x}}(\bar{\mathbf{x}};0)$, $\mathbf{g}_{\sigma}(\bar{\mathbf{x}};0)$, $\mathbf{h}_{\mathbf{x}}(\bar{\mathbf{x}};0)$, and $\mathbf{h}_{\sigma}(\bar{\mathbf{x}};0)$ to evaluate the approximation at any arbitrary point $(\mathbf{x},\sigma)$. We are searching for $n\times(n_x+1)$ coefficients (the $n_x \times n_y$ terms in $\mathbf{g}_{\mathbf{x}}(\bar{\mathbf{x}};0)$, the $n_x \times n_x$ terms in $\mathbf{h}_{\mathbf{x}}(\bar{\mathbf{x}};0)$, the n_y terms in $\mathbf{g}_{\sigma}(\bar{\mathbf{x}};0)$, and the n_x terms in $\mathbf{h}_{\sigma}(\bar{\mathbf{x}};0)$).

These coefficients can be found by using:

$$F_{x_i}(\bar{\mathbf{x}};0)=0, \forall i,$$

which gives us $n \times n_x$ equations and

$$F_{\sigma}(\bar{\mathbf{x}};0)=0,$$

which gives us n equations.

But before doing so, and to avoid runaway notation, we need to introduce the use of tensors.

Remark 10 (Tensor notation) Tensor notation (or Einstein summation notation), commonly used in physics, keeps the algebra required to perform a perturbation at a manageable level by eliminating $\sum$ and ∂ signs. To further reduce clutter, the points of evaluation of a derivative are skipped when they are unambiguous from context. An n^{th}-rank tensor in an m-dimensional space is an operator that has n indices and m^n components and obeys certain transformation rules. In our environment, $[\mathcal{H}_y]^i_\alpha$ is the (i, α) element of the derivative of $\mathcal{H}$ with respect to y:

1. The derivative of $\mathcal{H}$ with respect to y is an $n \times n_y$ matrix.
2. Thus, $[\mathcal{H}_y]^i_\alpha$ is the i-th row and α-th column element of this matrix.

3. When a subindex appears as a superindex in the next term, we are omitting a sum operator. For example,

$$[\mathcal{H}_y]^i_\alpha [\mathbf{g}_x]^\alpha_\beta [\mathbf{h}_x]^\beta_j = \sum_{\alpha=1}^{n_y} \sum_{\beta=1}^{n_x} \frac{\partial \mathcal{H}^i}{\partial y^\alpha} \frac{\partial \mathbf{g}^\alpha}{\partial x^\beta} \frac{\partial \mathbf{h}^\beta}{\partial x^j}.$$

4. The generalization to higher derivatives is direct. If we have $[\mathcal{H}_{y'y'}]^i_{\alpha\gamma}$:
   (a) $\mathcal{H}_{y'y'}$ is a three-dimensional array with n rows, n_y columns, and n_y pages.
   (b) Thus, $[\mathcal{H}_{y'y'}]^i_{\alpha\gamma}$ denotes the i-th row, α-th column element, and γ-th page of this matrix.

With the tensor notation, we can get into solving the system. First, $\mathbf{g_x}(\bar{\mathbf{x}};0)$ and $\mathbf{h_x}(\bar{\mathbf{x}};0)$ are the solution to:

$$[F_x(\bar{\mathbf{x}};0)]^i_j = [\mathcal{H}_{y'}]^i_\alpha [\mathbf{g}_x]^\alpha_\beta [\mathbf{h}_x]^\beta_j + [\mathcal{H}_y]^i_\alpha [\mathbf{g}_x]^\alpha_j + [\mathcal{H}_{x'}]^i_\beta [\mathbf{h}_x]^\beta_j + [\mathcal{H}_x]^i_j = \mathbf{0}; \\ i=1,\ldots,n;\ \ j,\beta=1,\ldots,n_x;\ \ \alpha=1,\ldots,n_y. \tag{20}$$

The derivatives of $\mathcal{H}$ evaluated at $(\mathbf{y},\mathbf{y}',\mathbf{x},\mathbf{x}') = (\bar{\mathbf{y}},\bar{\mathbf{y}},\bar{\mathbf{x}},\bar{\mathbf{x}})$ are known. Therefore, we have a system of $n \times n_x$ quadratic equations in the $n \times n_x$ unknowns given by the elements of $\mathbf{g_x}(\bar{\mathbf{x}};0)$ and $\mathbf{h_x}(\bar{\mathbf{x}};0)$. After some algebra, the system (20) can be written as:

$$AP^2 - BP - C = \mathbf{0}$$

where the $\tilde{n} \times \tilde{n}$ matrix A, the $\tilde{n} \times \tilde{n}$ matrix B and the $\tilde{n} \times \tilde{n}$ matrix C involve terms from $[\mathcal{H}_{y'}]^i_\alpha$, $[\mathcal{H}_y]^i_\alpha$, $[\mathcal{H}_{x'}]^i_\beta$, and $[\mathcal{H}_x]^i_j$ and the $\tilde{n} \times \tilde{n}$ matrix P the terms $[\mathbf{h}_x]^\beta_j$ related to the law of motion of $\mathbf{x}_1$ (in our worked-out example of the next subsection, we will make this algebra explicit). We can solve this system with a standard quadratic matrix equation solver.

Remark 11** **(Quadratic equation solvers) The literature has proposed several procedures to solve quadratic systems. Without being exhaustive, we can list Blanchard and Kahn (1980), King and Watson (1998), Uhlig (1999), Klein (2000), and Sims (2002). These different approaches vary in the details of how the solution to the system is found and how general they are (regarding the regularity conditions they require). But, conditional on applicability, all methods find the same policy functions since the linear space approximating a nonlinear space is unique.

For concision, we will only present one of the simplest of these procedures, as discussed by Uhlig (1999, pp. 43–45). Given

$$AP^2 - BP - C = \mathbf{0},$$

define the $2\tilde{n} \times 2\tilde{n}$ matrix:

$$D = \begin{bmatrix} A & \mathbf{0}_{\tilde{n}} \\ \mathbf{0}_{\tilde{n}} & I_{\tilde{n}} \end{bmatrix}$$

where $I_{\tilde{n}}$ is the $\tilde{n} \times \tilde{n}$ identity matrix and $\mathbf{0}_{\tilde{n}}$ the $\tilde{n} \times \tilde{n}$ zero matrix, and the $2\tilde{n} \times 2\tilde{n}$ matrix:

$$F = \begin{bmatrix} B & C \\ I_n & \mathbf{0}_n \end{bmatrix}$$

Let Q and Z be unitary matrices (ie, $Q^H Q = Z^H Z = I_{2\tilde{n}}$ where H is the complex Hermitian transposition operator). Let Φ and Σ be upper triangular matrices with diagonal elements ϕ_{ii} and σ_{ii}. Then, we find the generalized Schur decomposition (QZ) of D and F:

$$Q'\Sigma Z = D$$
$$Q'\Phi Z = F$$

such that Σ and Φ are diagonal and the ratios of diagonal elements $|\phi_{ii}/\sigma_{ii}|$ are in increasing order (there exists a QZ decomposition for every ordering of these ratios). In such a way, the stable (smaller than one) generalized eigenvalues of F with respect to D would come first and the unstable generalized eigenvalues (exceeding one and infinite) would come last. QZ decompositions are performed by standard numerical software such as Matlab and many programs exist to achieve the QZ decomposition with the desired ordering of ratios.

Then, if we partition:

$$Z = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}$$

where each submatrix Z_{ii} has a size $\tilde{n} \times \tilde{n}$, we can find:

$$P = -Z_{21}^{-1} Z_{22}.$$

If the number of ratios of diagonal elements with absolute value less than 1 (ie, we have enough stable generalized eigenvalues of F with respect to D), then we can select a P such that $P^m x \rightarrow \mathbf{0}$ as $m \rightarrow \infty$ for any $\tilde{n}$-dimensional vector. If the number of ratios of diagonal elements with absolute value less than 1 is larger than $\tilde{n}$, there may be more than one possible choice of P such that $P^m x \rightarrow \mathbf{0}$ as $m \rightarrow \infty$ for any $\tilde{n}$-dimensional vector.

The reason a quadratic system appears is that, in general, we will have multiple possible paths for the endogenous variables of the model that would satisfy the equilibrium conditions (Uhlig, 1999 and Galor, 2007). Some of these paths (the stable manifolds) will be stable and satisfy appropriate transversality conditions (although they might imply limit cycles). The other paths (the unstable manifolds) will not. We will need to select the right eigenvalues that induce stability. For many DSGE models, we will have exactly $\tilde{n}$ stable generalized eigenvalues and the stable solution would also be unique. If we have too few stable generalized eigenvalues, the equilibrium dynamics will be inherently unstable. If we have too many, we can have sunspots (Lubik and Schorfheide, 2003). Suffice it to note here that all these issues would depend only on the first-order approximation and that going to higher-order approximations would not change the issues at hand. If we have uniqueness of equilibrium in the first-order approximation, we will also have uniqueness in the second-order approximation. And if we have multiplicity of equilibria in the first-order approximation, we will also have multiplicity in the second-order approximation.

Remark 12 (Partitioning the quadratic system) The quadratic system (20) can be further divided into two parts to get a recursive solution. The system:

$$\begin{gathered}[F_x(\bar{\mathbf{x}};0)]^i_j = [\mathcal{H}_{y'}]^i_\alpha[\mathbf{g}_x]^\alpha_\beta[\mathbf{h}_x]^\beta_j + [\mathcal{H}_y]^i_\alpha[\mathbf{g}_x]^\alpha_j + [\mathcal{H}_{x'}]^i_\beta[\mathbf{h}_x]^\beta_j + [\mathcal{H}_x]^i_j = \mathbf{0}; \\ i = 1,\ldots,n;\;\; j,\beta = 1,\ldots,\tilde{n};\;\; \alpha = 1,\ldots,n_y.\end{gathered} \tag{21}$$

only involves the $\tilde{n} \times n_y$ elements of $\mathbf{g}_\mathbf{x}(\bar{\mathbf{x}};0)$ and the $\tilde{n} \times n_x$ elements of $\mathbf{h}_\mathbf{x}(\bar{\mathbf{x}};0)$ related to the $\tilde{n}$ endogenous state variables $\mathbf{x}_1$. Once we have solved the $\tilde{n} \times \left(n_y + n_x\right)$ unknowns in this system, we can plug them into the system:

$$\begin{gathered}[F_x(\bar{\mathbf{x}};0)]^i_j = [\mathcal{H}_{y'}]^i_\alpha[\mathbf{g}_x]^\alpha_\beta[\mathbf{h}_x]^\beta_j + [\mathcal{H}_y]^i_\alpha[\mathbf{g}_x]^\alpha_j + [\mathcal{H}_{x'}]^i_\beta[\mathbf{h}_x]^\beta_j + [\mathcal{H}_x]^i_j = \mathbf{0}; \\ i = 1,\ldots,n;\;\; j,\beta = \tilde{n}+1,\ldots,n_x;\;\; \alpha = 1,\ldots,n_y.\end{gathered} \tag{22}$$

and solve for the $n_\epsilon \times n_y$ elements of $\mathbf{g}_\mathbf{x}(\bar{\mathbf{x}};0)$ and the $n_\epsilon \times n_x$ elements of $\mathbf{h}_\mathbf{x}((\bar{\mathbf{x}};0)$ related to the n_ϵ stochastic variables $\mathbf{x}_2$.

This recursive solution has three advantages. The first, and most obvious, is that it simplifies computations. The system (20) has $n_x \times \left(n_y + n_x\right)$ unknowns, while the system (21) has $\tilde{n} \times \left(n_y + n_x\right)$. The difference, $n_\epsilon \times \left(n_y + n_x\right)$, makes the second system considerably smaller. Think, for instance, about the medium-scale New Keynesian model in Fernández-Villaverde and Rubio-Ramírez (2008). In the notation of this chapter, the model has $n_x = 20$, $n_y = 1$, and $n_\epsilon = 5$. Thus, by partitioning the system, we go from solving for 420 unknowns to solve a first system of 315 unknowns and, later, a second system of 105 unknowns. The second advantage, which is not obvious in our compact notation, is that system (22) is linear and, therefore, much faster to solve and with a unique solution. In the next subsection, with our worked-out example, we will see this more clearly. The third advantage is that, in some cases, we may only care about the coefficients associated with the $\tilde{n}$ endogenous state variables $\mathbf{x}_1$. This occurs, for example, when we are interested in computing the deterministic transitional path of the model toward a steady state given some initial conditions or when we are plotting impulse response functions generated by the first-order approximation.

The coefficients $\mathbf{g}_\sigma(\bar{\mathbf{x}};0)$ and $\mathbf{h}_\sigma(\bar{\mathbf{x}};0)$ are the solution to the n equations:

$$\begin{gathered}[F_\sigma(\bar{\mathbf{x}};0)]^i = \mathbb{E}_t\{[\mathcal{H}_{y'}]^i_\alpha[\mathbf{g}_x]^\alpha_\beta[\mathbf{h}_\sigma]^\beta + [\mathcal{H}_{y'}]^i_\alpha[\mathbf{g}_x]^\alpha_\beta[\eta]^\beta_\phi[\epsilon']^\phi + [\mathcal{H}_{y'}]^i_\alpha[\mathbf{g}_\sigma]^\alpha \\ + [\mathcal{H}_y]^i_\alpha[\mathbf{g}_\sigma]^\alpha + [\mathcal{H}_{x'}]^i_\beta[\mathbf{h}_\sigma]^\beta + [\mathcal{H}_{x'}]^i_\beta[\eta]^\beta_\phi[\epsilon']^\phi\} \\ i = 1,\ldots,n;\;\; \alpha = 1,\ldots,n_y;\;\; \beta = 1,\ldots,n_x;\;\; \phi = 1,\ldots,n_\epsilon.\end{gathered}$$

Then:

$$[F_\sigma(\bar{\mathbf{x}};0)]^i = [\mathcal{H}_{y'}]^i_\alpha[\mathbf{g}_x]^\alpha_\beta[\mathbf{h}_\sigma]^\beta + [\mathcal{H}_{y'}]^i_\alpha[\mathbf{g}_\sigma]^\alpha + [\mathcal{H}_y]^i_\alpha[\mathbf{g}_\sigma]^\alpha + [f_{x'}]^i_\beta[\mathbf{h}_\sigma]^\beta = 0;$$
$$i=1,\ldots,n;\quad \alpha=1,\ldots,n_y;\quad \beta=1,\ldots,n_x;\quad \phi=1,\ldots,n_\epsilon.$$

Inspection of the previous equations shows that they are linear and homogeneous equations in $\mathbf{g}_\sigma$ and $\mathbf{h}_\sigma$. Thus, if a unique solution exists, it satisfies:

$$\mathbf{g}_\sigma = \mathbf{0}$$
$$\mathbf{h}_\sigma = \mathbf{0}$$

In other words, the coefficients associated with the perturbation parameter are zero and the first-order approximation is

$$\mathbf{g}(\mathbf{x};\sigma) - \bar{\mathbf{y}} = \mathbf{g}_\mathbf{x}(\bar{\mathbf{x}};0)(\mathbf{x}-\bar{\mathbf{x}})'$$
$$\mathbf{h}(\mathbf{x};\sigma) - \bar{\mathbf{x}} = \mathbf{h}_\mathbf{x}(\bar{\mathbf{x}};0)(\mathbf{x}-\bar{\mathbf{x}})'.$$

These equations embody certainty equivalence as defined by Simon (1956) and Theil (1957). Under certainty equivalence, the solution of the model, up to first-order, is identical to the solution of the same model under perfect foresight (or under the assumption that $\sigma = 0$). Certainty equivalence does not preclude the realization of the shock from appearing in the decision rule. What certainty equivalence precludes is that the standard deviation of it appears as an argument by itself, regardless of the realization of the shock.

The intuition for the presence of certainty equivalence is simple. Risk-aversion depends on the second derivative of the utility function (concave utility). However, Leland (1968) and Sandmo (1970) showed that precautionary behavior depends on the third derivative of the utility function. But a first-order perturbation involves the equilibrium conditions of the model (which includes first derivatives of the utility function, for example, in the Euler equation that equates marginal utilities over time) and first derivatives of these equilibrium conditions (and, therefore, second derivatives of the utility function), but not higher-order derivatives.

Certainty equivalence has several drawbacks. First, it makes it difficult to talk about the welfare effects of uncertainty. Although the dynamics of the model are still partially driven by the variance of the innovations (the realizations of the innovations depend on it), the agents in the model do not take any precautionary behavior to protect themselves from that variance, biasing any welfare computation. Second, related to the first point, the approximated solution generated under certainty equivalence cannot generate any risk premia for assets, a strongly counterfactual prediction.[h] Third, certainty equivalence prevents researchers from analyzing the consequences of changes in volatility.

[h] In general equilibrium, there is an intimate link between welfare computations and asset pricing. An exercise on the former is always implicitly an exercise on the latter (see Alvarez and Jermann, 2004).

Remark 13 (Perturbation and LQ approximations) Kydland and Prescott (1982)—and many papers after them—took a different route to solving DSGE models. Imagine that we have an optimal control problem that depends on n_x states $\mathbf{x}_t$ and n_u control variables $\mathbf{u}_t$. To save on notation, let us also define the column vector $\mathbf{w}_t = [\mathbf{x}_t, \mathbf{u}_t]'$ of dimension $n_w = n_x + n_u$. Then, we can write the optimal control problem as:

$$\max \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t r(\mathbf{w}_t)$$
$$\text{s.t. } \mathbf{x}_{t+1} = A(\mathbf{w}_t, \varepsilon_t)$$

where r is a return function, ε_t a vector of n_ε innovations with zero mean and finite variance, and A summarizes all the constraints and laws of motion of the economy. By appropriately enlarging the state space, this notation can accommodate the innovations having an impact on the period return function and some variables being both controls and states.

In the case where the return function r is quadratic, ie,

$$r(\mathbf{w}_t) = B_0 + B_1 \mathbf{w}_t + \mathbf{w}_t' Q \mathbf{w}_t$$

(where B_0 is a constant, B_1 a row vector $1 \times n_w$, and B_2 is an $n_w \times n_w$ matrix) and the function A is linear:

$$\mathbf{x}_{t+1} = B_3 \mathbf{w}_t + B_4 \varepsilon_t$$

(where B_3 is an $n_x \times n_w$ matrix and B_4 is an $n_x \times n_\varepsilon$ matrix), we are facing a stochastic discounted linear-quadratic regulator (LQR) problem. There is a large and well-developed research area on LQR problems. This literature is summarized by Anderson et al. (1996) and Hansen and Sargent (2013). In particular, we know that the optimal decision rule in this environment is a linear function of the states and the innovations:

$$\mathbf{u}_t = F_w \mathbf{w}_t + F_\varepsilon \varepsilon_t$$

where F_w can be found by solving a Ricatti equation Anderson et al. (1996, pp. 182–183) and F_ε by solving a Sylvester equation Anderson et al. (1996, pp. 202–205). Interestingly, F_w is independent of the variance of ε_t. That is, if ε_t has a zero variance, then the optimal decision rule is simply:

$$\mathbf{u}_t = F_w \mathbf{w}_t.$$

This neat separation between the computation of F_w and of F_ε allows the researcher to deal with large problems with ease. However, it also implies certainty equivalence.

Kydland and Prescott (1982) setup the social planner's problem of their economy, which fits into an optimal regulator problem, and they were able to write a function A that was linear in $\mathbf{w}_t$, but they did not have a quadratic return function. Instead, they took a quadratic approximation to the objective function of the social planner. Most of the literature that followed them used a Taylor series approximation of the objective function around the deterministic steady state, sometimes called the approximated LQR problem (Kydland and Prescott also employed a slightly different point of approximation that attempted to control for uncertainty; this did not make much quantitative difference). Furthermore, Kydland and Prescott worked with the value function representation of the problem. See Díaz-Giménez (1999) for an explanation of how to deal with the LQ approximation to the value function.

The result of solving the approximated LQR when the function A is linear is equivalent to the result of a first-order perturbation of the equilibrium conditions of the model. The intuition is simple. Derivatives are unique, and since both approaches search for a linear approximation to the solution of the model, they have to yield identical results.

However, approximated LQR have lost their popularity for three reasons. First, it is often hard to write the function A in a linear form. Second, it is challenging to set up a social planner's problem when the economy is not Pareto efficient. And even when it is possible to have a modified social planner's problem that incorporates additional constraints that incorporate non-optimalities (see, for instance, Benigno and Woodford, 2004), the same task is usually easier to accomplish by perturbing the equilibrium conditions of the model. Third, and perhaps most important, perturbations can easily go to higher-order terms and incorporate nonlinearities that break certainty equivalence.

4.2.5 Second-Order Perturbation

Once we have finished the first-order perturbation, we can iterate on the steps before to generate higher-order solutions. More concretely, the second-order approximations to $\mathbf{g}$ around $(\mathbf{x};\sigma)=(\bar{\mathbf{x}};0)$ are:

$$\begin{aligned}
[\mathbf{g}(\mathbf{x};\sigma)]^i = {} & [\mathbf{g}(\bar{\mathbf{x}};0)]^i + [\mathbf{g}_x(\bar{\mathbf{x}};0)]^i_a[(\mathbf{x}-\mathbf{x})]_a + [\mathbf{g}_\sigma(\bar{\mathbf{x}};0)]^i[\sigma] \\
& + \frac{1}{2}[\mathbf{g}_{xx}(\bar{\mathbf{x}};0)]^i_{ab}[(\mathbf{x}-\bar{\mathbf{x}})]_a[(\mathbf{x}-\bar{\mathbf{x}})]_b \\
& + \frac{1}{2}[\mathbf{g}_{x\sigma}(\bar{\mathbf{x}};0)]^i_a[(\mathbf{x}-\bar{\mathbf{x}})]_a[\sigma] \\
& + \frac{1}{2}[\mathbf{g}_{\sigma x}(\bar{\mathbf{x}};0)]^i_a[(\mathbf{x}-\bar{\mathbf{x}})]_a[\sigma] \\
& + \frac{1}{2}[\mathbf{g}_{\sigma\sigma}(\bar{\mathbf{x}};0)]^i[\sigma][\sigma]
\end{aligned}$$

where $i=1,\ldots,n_y$, $a,b=1,\ldots,n_x$, and $j=1,\ldots,n_x$.

Similarly, the second-order approximations to $\mathbf{h}$ around $(\mathbf{x};\sigma)=(\bar{\mathbf{x}};0)$ are:

$$
\begin{aligned}
[\mathbf{h}(\mathbf{x};\sigma)]^{j} = {} & [\mathbf{h}(\bar{\mathbf{x}};0)]^{j} + [\mathbf{h}_{x}(\bar{\mathbf{x}};0)]^{j}_{a}[(\mathbf{x}-\mathbf{x})]_{a} + [\mathbf{h}_{\sigma}(\bar{\mathbf{x}};0)]^{j}[\sigma] \\
& + \frac{1}{2}[\mathbf{h}_{xx}(\bar{\mathbf{x}};0)]^{j}_{ab}[(\mathbf{x}-\bar{\mathbf{x}})]_{a}[(\mathbf{x}-\bar{\mathbf{x}})]_{b} \\
& + \frac{1}{2}[\mathbf{h}_{x\sigma}(\bar{\mathbf{x}};0)]^{j}_{a}[(\mathbf{x}-\bar{\mathbf{x}})]_{a}[\sigma] \\
& + \frac{1}{2}[\mathbf{h}_{\sigma x}(\bar{\mathbf{x}};0)]^{j}_{a}[(\mathbf{x}-\bar{\mathbf{x}})]_{a}[\sigma] \\
& + \frac{1}{2}[\mathbf{h}_{\sigma\sigma}(\bar{\mathbf{x}};0)]^{j}[\sigma][\sigma],
\end{aligned}
$$

where $i=1,\ldots,n_y$, $a,b=1,\ldots,n_x$, and $j=1,\ldots,n_x$.

The unknown coefficients in these approximations are $[\mathbf{g}_{xx}]^{i}_{ab}$, $[\mathbf{g}_{x\sigma}]^{i}_{a}$, $[\mathbf{g}_{\sigma x}]^{i}_{a}$, $[\mathbf{g}_{\sigma\sigma}]^{i}$, $[\mathbf{h}_{xx}]^{j}_{ab}$, $[\mathbf{h}_{x\sigma}]^{j}_{a}$, $[\mathbf{h}_{\sigma x}]^{j}_{a}$, $[\mathbf{h}_{\sigma\sigma}]^{j}$. As before, we solve for these coefficients by taking the second derivatives of $F(\mathbf{x};\sigma)$ with respect to x and σ, making them equal to zero, and evaluating them at $(\bar{\mathbf{x}};0)$.

How do we solve the system? First, we exploit $F_{xx}(\bar{\mathbf{x}};0)$ to solve for $\mathbf{g}_{xx}(\bar{\mathbf{x}};0)$ and $h_{xx}(\bar{\mathbf{x}};0)$:

$$
\begin{aligned}
& [F_{xx}(\bar{\mathbf{x}};0)]^{i}_{jk} = \\
& \left([\mathcal{H}_{y'y'}]^{i}_{\alpha\gamma}[\mathbf{g}_{x}]^{\gamma}_{\delta}[\mathbf{h}_{x}]^{\delta}_{k} + [\mathcal{H}_{y'y}]^{i}_{\alpha\gamma}[\mathbf{g}_{x}]^{\gamma}_{k} + [\mathcal{H}_{y'x'}]^{i}_{\alpha\delta}[\mathbf{h}_{x}]^{\delta}_{k} + [\mathcal{H}_{y'x}]^{i}_{\alpha k}\right)[\mathbf{g}_{x}]^{\alpha}_{\beta}[\mathbf{h}_{x}]^{\beta}_{j} \\
& +[\mathcal{H}_{y'}]^{i}_{\alpha}[\mathbf{g}_{xx}]^{\alpha}_{\beta\delta}[\mathbf{h}_{x}]^{\delta}_{k}[\mathbf{h}_{x}]^{\beta}_{j} + [\mathcal{H}_{y'}]^{i}_{\alpha}[\mathbf{g}_{x}]^{\alpha}_{\beta}[\mathbf{h}_{xx}]^{\beta}_{jk} \\
& +\left([\mathcal{H}_{yy'}]^{i}_{\alpha\gamma}[\mathbf{g}_{x}]^{\gamma}_{\delta}[\mathbf{h}_{x}]^{\delta}_{k} + [\mathcal{H}_{yy}]^{i}_{\alpha\gamma}[\mathbf{g}_{x}]^{\gamma}_{k} + [\mathcal{H}_{yx'}]^{i}_{\alpha\delta}[\mathbf{h}_{x}]^{\delta}_{k} + [\mathcal{H}_{yx}]^{i}_{\alpha k}\right)[\mathbf{g}_{x}]^{\alpha}_{j} + [\mathcal{H}_{y}]^{i}_{\alpha}[\mathbf{g}_{xx}]^{\alpha}_{jk} \\
& +\left([\mathcal{H}_{x'y'}]^{i}_{\beta\gamma}[\mathbf{g}_{x}]^{\gamma}_{\delta}[\mathbf{h}_{x}]^{\delta}_{k} + [\mathcal{H}_{x'y}]^{i}_{\beta\gamma}[\mathbf{g}_{x}]^{\gamma}_{k} + [\mathcal{H}_{x'x'}]^{i}_{\beta\delta}[\mathbf{h}_{x}]^{\delta}_{k} + [\mathcal{H}_{x'x}]^{i}_{\beta k}\right)[\mathbf{h}_{x}]^{\beta}_{j} + [\mathcal{H}_{x'}]^{i}_{\beta}[\mathbf{h}_{xx}]^{\beta}_{jk} \\
& +[\mathcal{H}_{xy'}]^{i}_{j\gamma}[\mathbf{g}_{x}]^{\gamma}_{\delta}[\mathbf{h}_{x}]^{\delta}_{k} + [\mathcal{H}_{xy}]^{i}_{j\gamma}[\mathbf{g}_{x}]^{\gamma}_{k} + [\mathcal{H}_{xx'}]^{i}_{j\delta}[\mathbf{h}_{x}]^{\delta}_{k} + [\mathcal{H}_{xx}]^{i}_{jk} = 0; \\
& \qquad i=1,\ldots n, \quad j,k,\beta,\delta=1,\ldots n_x; \quad \alpha,\gamma=1,\ldots n_y.
\end{aligned}
$$

But we know the derivatives of $\mathcal{H}$. We also know the first derivatives of **g** and **h** evaluated at $(\bar{\mathbf{x}}, 0)$. Hence, the above expression is a system of $n \times n_x \times n_x$ linear equations in the $n \times n_x \times n_x$ unknown elements of $\mathbf{g}_{xx}$ and $\mathbf{h}_{xx}$. This point is crucial: linear solvers are fast and efficient. In the first-order approximation we had to solve a quadratic system to select between stable and unstable solutions. But once we are already in the stable manifold, there are no further additional solutions that we need to rule out. These quadratic terms involve the endogenous state vector x_1. Those terms capture nonlinear behavior and induce nonsymmetries. We will discuss those in more detail in our worked-out example below.

The coefficients in $\mathbf{g}_{\sigma\sigma}$ and $\mathbf{h}_{\sigma\sigma}$ come from solving the system of n linear equations in the n unknowns:

$$\begin{aligned}[F_{\sigma\sigma}(\bar{\mathbf{x}};0)]^i = {} & [\mathcal{H}_{y'}]^i_\alpha[\mathbf{g}_x]^\alpha_\beta[\mathbf{h}_{\sigma\sigma}]^\beta \\ & + [\mathcal{H}_{y'y'}]^i_{\alpha\gamma}[\mathbf{g}_x]^\gamma_\delta[\eta]^\delta_\xi[\mathbf{g}_x]^\alpha_\beta[\eta]^\beta_\phi[I]^\phi_\xi + [\mathcal{H}_{y'x'}]^i_{\alpha\delta}[\eta]^\delta_\xi[\mathbf{g}_x]^\alpha_\beta[\eta]^\beta_\phi[I]^\phi_\xi \\ & + [\mathcal{H}_{y'}]^i_\alpha[\mathbf{g}_{xx}]^\alpha_{\beta\delta}[\eta]^\delta_\xi[\eta]^\beta_\phi[I]^\phi_\xi + [\mathcal{H}_{y'}]^i_\alpha[\mathbf{g}_{\sigma\sigma}]^\alpha \\ & + [\mathcal{H}_y]^i_\alpha[\mathbf{g}_{\sigma\sigma}]^\alpha + [\mathcal{H}_{x'}]^i_\beta[\mathbf{h}_{\sigma\sigma}]^\beta \\ & + [\mathcal{H}_{x'y'}]^i_{\beta\gamma}[\mathbf{g}_x]^\gamma_\delta[\eta]^\delta_\xi[\eta]^\beta_\phi[I]^\phi_\xi + [\mathcal{H}_{x'x'}]^i_{\beta\delta}[\eta]^\delta_\xi[\eta]^\beta_\phi[I]^\phi_\xi = 0; \\ & = 1,\ldots,n; \alpha,\gamma = 1,\ldots,n_y; \beta,\ \delta = 1,\ldots,n_x; \phi,\ \xi = 1,\ldots,n_\epsilon.\end{aligned}$$

The coefficients $\mathbf{g}_{\sigma\sigma}$ and $\mathbf{h}_{\sigma\sigma}$ capture the correction for risk that breaks certainty equivalence. In addition, the cross derivatives $\mathbf{g}_{x\sigma}$ and $\mathbf{h}_{x\sigma}$ are zero when evaluated at $(\bar{\mathbf{x}};0)$. To see this, write the system $F_{\sigma x}(\bar{\mathbf{x}};0)=0$, taking into account that all terms containing either g_σ or $\mathbf{h}_\sigma$ are zero at $(\bar{\mathbf{x}};0)$. Then, we have a homogeneous system of $n \times n_x$ equations in the $n \times n_x$ elements of $\mathbf{g}_{\sigma x}$ and $\mathbf{h}_{\sigma x}$:

$$\begin{gathered}[F_{\sigma x}(\bar{\mathbf{x}};0)]^i_j = [\mathcal{H}_{y'}]^i_\alpha[\mathbf{g}_x]^\alpha_\beta[\mathbf{h}_{\sigma x}]^\beta_j + [\mathcal{H}_{y'}]^i_\alpha[\mathbf{g}_{\sigma x}]^\alpha_\gamma[\mathbf{h}_x]^\gamma_j + [\mathcal{H}_y]^i_\alpha[\mathbf{g}_{\sigma x}]^\alpha_j + [\mathcal{H}_{x'}]^i_\beta[\mathbf{h}_{\sigma x}]^\beta_j = 0; \\ i = 1,\ldots n; \quad \alpha = 1,\ldots,n_y; \quad \beta,\gamma,j = 1,\ldots,n_x.\end{gathered}$$

Hence, the last component of the second-order perturbation is given by:[i]

$$\begin{aligned}\mathbf{g}_{\sigma x} &= \mathbf{0} \\ \mathbf{h}_{\sigma x} &= \mathbf{0}.\end{aligned}$$

4.2.6 Higher-Order Perturbations

We can iterate the previous procedure (taking higher-order derivatives, plugging in the already found terms, and solving for the remaining ones) as many times as we want to obtain n-th order approximations. All the associated systems of equations that we would need to solve are linear, which keeps the computational complexity manageable. The only additional point to remember is that we will need to make assumptions about the higher moments of the innovations, as we will have expectational terms involving these higher moments.

If the functions $\mathbf{g}$ and $\mathbf{h}$ are analytic in a neighborhood of $\bar{\mathbf{x}}$, then the series we are building by taking higher-order approximations has an infinite number of terms and is convergent. Convergence will occur in a radius of convergence centered around $\bar{\mathbf{x}}$, (ie, the r such that for all state values with a distance with respect to $\bar{\mathbf{x}}$ smaller then r). This radius can be infinite. In that case, the series is guaranteed to converge uniformly everywhere. However, the radius can also be finite and there exist a nonremovable

[i] We conjecture (and we have checked up to as high an order of a perturbation as computer memory allows) that all terms involving odd derivatives of σ are zero. Unfortunately, we do not have a formal proof.

singularity on its boundary. Disappointingly, for most DSGE models, the radius of convergence is unknown (for more details and examples, see Swanson et al., 2006 and Aldrich and Kung, 2011). More research on this topic is sorely needed. Also, even when the series is convergent, there are two potential problems. First, at a j-th order approximation, we may lose the "right" shape of $\mathbf{g}$ and $\mathbf{h}$. For example, Aruoba et al. (2006) document how the decision rules for consumption and capital of the stochastic neoclassical growth model approximated with a fifth-order perturbation are no longer globally concave, as implied by economic theory. Instead, the approximated functions present oscillating patterns. Second, the convergence to the exact solution may not be monotone: it is easy to build examples where the errors a bit away from $\bar{\mathbf{x}}$ are worse for a $j+1$-th order approximation than for a j-th order approximation. Neither of these two problems is fatal, but the researcher needs to be aware of them and undertake the necessary tests to minimize their impact (for instance, checking the solution for different approximation orders).

Later, we will discuss how to gauge the accuracy of a solution and how to decide whether a higher-order approximation is required. For example, to deal with models with time-varying volatility, we would need at least a third-order approximation. Levintal (2015a) has argued that to approximate well models with disaster risk, we need a fifth-order approximation. The drawback of higher-order approximations is that we will run into problems of computational cost and memory use.

4.3 A Worked-Out Example

The previous derivations were somewhat abstract and the notation, even using tensors, burdensome. Consequently, it is useful to show how perturbation works in a concrete example. For that, we come back to our example of the neoclassical growth model defined by Eqs. (2)–(4), except that, to make the algebra easier, we assume $u(c) = \log c$ and $\delta = 1$.

The equilibrium conditions of the model are then:

$$\frac{1}{c_t} = \beta \mathbb{E}_t \frac{\alpha e^{z_{t+1}} k_{t+1}^{\alpha-1}}{c_{t+1}}$$

$$c_t + k_{t+1} = e^{z_t} k_t^{\alpha}$$

$$z_t = \rho z_{t-1} + \eta \varepsilon_t$$

While this parameterization is unrealistic for periods of time such as a quarter or a year typically employed in business cycle analysis, it has the enormous advantage of implying that the model has a closed-form solution. With $\delta = 1$, the income and the substitution effect from a productivity shock cancel each other, and consumption and investment are constant fractions of income:

$$c_t = (1 - \alpha\beta) e^{z_t} k_t^{\alpha}$$

$$k_{t+1} = \alpha\beta e^{z_t} k_t^{\alpha}$$

(these optimal decision rules can be verified by plugging them into the equilibrium conditions and checking that indeed these conditions are satisfied).

Imagine, however, that we do not know this exact solution and that we are searching a decision rule for consumption:

$$c_t = c(k_t, z_t)$$

and another one for capital:

$$k_{t+1} = k(k_t, z_t)$$

In our general notation, d would just be the stack of $c(k_t, z_t)$ and $k(k_t, z_t)$. We substitute these decision rules in the equilibrium conditions above (and, to reduce the dimensionality of the problem, we substitute out the budget constraint and the law of motion for technology) to get:

$$\frac{1}{c(k_t, z_t)} = \beta \mathbb{E}_t \frac{\alpha e^{\rho z_t + \sigma \varepsilon_{t+1}} k(k_t, z_t)^{\alpha-1}}{c(k(k_t, z_t), \rho z_t + \eta \varepsilon_{t+1})} \tag{23}$$

$$c(k_t, z_t) + k(k_t, z_t) = e^{z_t} k_t^{\alpha} \tag{24}$$

The decision rules are approximated by perturbation solutions on the two state variables plus the perturbation parameter σ:

$$\begin{aligned} c_t &= c(k_t, z_t; \sigma) \\ k_{t+1} &= k(k_t, z_t; \sigma). \end{aligned}$$

We introduce σ in the law of motion for technology:

$$z_t = \rho z_{t-1} + \sigma \eta \varepsilon_t.$$

In that way, if we set $\sigma = 0$, we recover a deterministic model. If $z_t = 0$ (either because $z_0 = 0$ or because t is sufficiently large such that $z_t \rightarrow 0$), we can find the steady state k by solving the system of equilibrium conditions:

$$\begin{aligned} \frac{1}{c} &= \beta \frac{\alpha k^{\alpha-1}}{c} \\ c + k &= k^{\alpha} \end{aligned}$$

which has a unique solution $k = k(k, 0; 0) = (\alpha\beta)^{\frac{1}{1-\alpha}}$ and $c = c(k, 0; 0) = (\alpha\beta)^{\frac{\alpha}{1-\alpha}} - (\alpha\beta)^{\frac{1}{1-\alpha}}$.

The second-order expansion for the consumption decision rule is given by:

$$\begin{aligned} c_t = {} & c + c_k(k_t - k) + c_z z_t + c_\sigma \sigma \\ & + \frac{1}{2} c_{kk}(k_t - k)^2 + c_{kz}(k_t - k) z_t + c_{k\sigma}(k_t - k)\sigma \\ & + \frac{1}{2} c_{zz} z_t^2 + c_{z\sigma} z_t \sigma + \frac{1}{2} c_{\sigma^2} \sigma^2 \end{aligned} \tag{25}$$

and for the capital decision rule:

$$\begin{aligned}k_{t+1} = k + k_k(k_t - k) + k_z z_t + k_\sigma \sigma \\ + \frac{1}{2}k_{kk}(k_t - k)^2 + k_{kz}(k_t - k)z_t + k_{k\sigma}(k_t - k)\sigma \\ + \frac{1}{2}k_{zz}z_t^2 + \frac{1}{2}k_{\sigma z}\sigma z_t + \frac{1}{2}k_{\sigma^2}\sigma^2\end{aligned} \tag{26}$$

(where we have already used the symmetry of second derivatives and assumed that all terms are evaluated at $(k, 0; 0)$). Higher-order approximations can be written in a similar way, but, for this example, a second-order approximation is all we need.

Beyond the correction for risk $\frac{1}{2}c_{\sigma^2}\sigma^2$ and $\frac{1}{2}k_{\sigma^2}\sigma^2$ that we discussed above, the additional terms in Eqs. (25) and (26) introduce dynamics that cannot be captured by a first-order perturbation. In the linear solution, the terms $c_z z_t$ and $k_\sigma \sigma$ imply that the effects of positive and negative shocks are mirrors of each other. That is why, for instance, researchers using linearized models only report impulse response functions to a positive or a negative shock: the other impulse response functions are the same but inverted. In comparison, in the second-order perturbation, the terms $\frac{1}{2}c_{zz}z_t^2$ and $\frac{1}{2}k_{zz}z_t^2$ mean that positive and negative shocks have divergent effects: z_t^2 is always positive and the impulse response functions are asymmetric. The terms $c_{kz}(k_t - k)z_t$ and $k_{kz}(k_t - k)z_t$ cause the effect of a shock to also depend on how much capital the economy has at period t, a mechanism missed in the first-order approximation since z_t enters linearly. This might be of importance in many applications. For example, the effects of a financial shock may depend on the household asset level.

To find the unknown coefficients in Eqs. (25) and (26), we come back to the equilibrium conditions (23) and (24), we substitute the decision rules with the approximated decision rules $c(k_t, z_t; \sigma)$ and $k(k_t, z_t; \sigma)$, and we rearrange terms to get:

$$F(k_t, z_t; \sigma) = \mathbb{E}_t \begin{bmatrix} \frac{1}{c(k_t, z_t; \sigma)} - \beta \frac{\alpha e^{\rho z_t + \sigma\eta\varepsilon_{t+1}} k(k_t, z_t; \sigma)^{\alpha - 1}}{c(k(k_t, z_t; \sigma), \rho z_t + \sigma\eta\varepsilon_{t+1}; \sigma)} \\ c(k_t, z_t; \sigma) + k(k_t, z_t; \sigma) - e^{z_t}k_t^\alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

More compactly:

$$F(k_t, z_t; \sigma) = \mathcal{H}(c(k_t, z_t; \sigma), c(k(k_t, z_t; \sigma), z_{t+1}; \sigma), k_t, k(k_t, z_t; \sigma), z_t; \sigma)$$

We will use $\mathcal{H}_i$ to represent the partial derivative of $\mathcal{H}$ with respect to the i component and drop the evaluation at the steady state of the functions when we do not need it.

We start with the first-order terms. We take derivatives of $F(k_t, z_t; \sigma)$ with respect to k_t, z_t, and σ and we equate them to zero:

$$\begin{aligned}F_k &= \mathcal{H}_1 c_k + \mathcal{H}_2 c_k k_k + \mathcal{H}_3 + \mathcal{H}_4 k_k = \mathbf{0}\\ F_z &= \mathcal{H}_1 c_z + \mathcal{H}_2 (c_k k_z + c_k \rho) + \mathcal{H}_4 k_z + \mathcal{H}_5 = \mathbf{0}\\ F_\sigma &= \mathcal{H}_1 c_\sigma + \mathcal{H}_2 (c_k k_\sigma + c_\sigma) + \mathcal{H}_4 k_\sigma + \mathcal{H}_6 = \mathbf{0}\end{aligned}$$

Note that:

$$\begin{aligned}F_k &= \mathcal{H}_1 c_k + \mathcal{H}_2 c_k k_k + \mathcal{H}_3 + \mathcal{H}_4 k_k = \mathbf{0}\\ F_z &= \mathcal{H}_1 c_z + \mathcal{H}_2 (c_k k_z + c_k \rho) + \mathcal{H}_4 k_z + \mathcal{H}_5 = \mathbf{0}\end{aligned}$$

is a quadratic system of four equations in four unknowns: c_k, c_z, k_k, and k_z (the operator F has two dimensions). As we mentioned above, the system can be solved recursively. The first two equations:

$$F_k = \mathcal{H}_1 c_k + \mathcal{H}_2 c_k k_k + \mathcal{H}_3 + \mathcal{H}_4 k_k = \mathbf{0}$$

only involve c_k and k_k (the terms affecting the deterministic variables).

Remark 14 (Quadratic problem, again) The first two equations:

$$F_k = \mathcal{H}_1 c_k + \mathcal{H}_2 c_k k_k + \mathcal{H}_3 + \mathcal{H}_4 k_k = \mathbf{0}$$

can easily be written in the form of a quadratic matrix system as follows. First, we write the two equations as:

$$\begin{pmatrix}\mathcal{H}_1^1\\ \mathcal{H}_1^2\end{pmatrix} c_k + \begin{pmatrix}\mathcal{H}_2^1\\ \mathcal{H}_2^2\end{pmatrix} c_k k_k + \begin{pmatrix}\mathcal{H}_3^1\\ \mathcal{H}_3^2\end{pmatrix} + \begin{pmatrix}\mathcal{H}_4^1\\ \mathcal{H}_4^2\end{pmatrix} k_k = \begin{pmatrix}0\\ 0\end{pmatrix}$$

where $\mathcal{H}_i^j$ is the j-th dimension of $\mathcal{H}_i$. But $\mathcal{H}_2^2 = 0$ and $\mathcal{H}_3^1 = 0$, then

$$\begin{pmatrix}\mathcal{H}_1^1\\ \mathcal{H}_1^2\end{pmatrix} c_k + \begin{pmatrix}\mathcal{H}_2^1\\ 0\end{pmatrix} c_k k_k + \begin{pmatrix}0\\ \mathcal{H}_3^2\end{pmatrix} + \begin{pmatrix}\mathcal{H}_4^1\\ \mathcal{H}_4^2\end{pmatrix} k_k = \begin{pmatrix}0\\ 0\end{pmatrix}.$$

We can use the second equation to eliminate c_k from the first equation. Then, rearranging the terms and calling $P = k_k$ we have the equation:

$$AP^2 - BP - C = 0$$

that we presented in the previous subsection. Note that, in this example, instead of a complicated matrix equation, we have a much simpler quadratic scalar equation.

Our quadratic system will have two solutions. One solution will imply that $k_k > 1$ and the other solution $k_k < 1$. The first solution is unstable. Remember that the first elements of the decision rule are

$$k_{t+1} = k + k_k (k_t - k) + \dots$$

If $k_k > 1$, a deviation of k_t with respect to k will imply an even bigger deviation of k_{t+1} with respect to k, leading to explosive behavior. In comparison, when $k_k < 1$, deviations of k_t with respect to k will, in the absence of additional shocks, dissipate over time. Once we know c_k and k_k, we can come back to

$$F_z = \mathcal{H}_1 c_z + \mathcal{H}_2(c_k k_z + c_k \rho) + \mathcal{H}_4 k_z + \mathcal{H}_5 = 0$$

and solve for c_z and k_z. As emphasized in Remark 12, this system is linear.

Finally, as in the general case, the last two equations

$$F_\sigma = \mathcal{H}_1 c_\sigma + \mathcal{H}_2(c_k k_\sigma + c_\sigma) + \mathcal{H}_4 k_\sigma + \mathcal{H}_6 = 0$$

form a linear, and homogeneous system in c_σ and k_σ. Hence, $c_\sigma = k_\sigma = 0$ and we obtain the certainty equivalence of first-order approximations.

To find the second-order approximation, we take second derivatives of $F(k_t, z_t; \sigma)$ around k, 0, and 0:

$$\begin{aligned} F_{kk} &= 0 \\ F_{kz} &= 0 \\ F_{k\sigma} &= 0 \\ F_{zz} &= 0 \\ F_{z\sigma} &= 0 \\ F_{\sigma\sigma} &= 0 \end{aligned}$$

(where we have already eliminated symmetric second derivatives). We substitute the coefficients that we already know from the first-order approximation and we get a linear system of 12 equations in 12 unknowns. Again, we get that all cross-terms on $k\sigma$ and $z\sigma$ are zero.

Imposing the results concerning the coefficients that are equal to zero, we can rewrite Eqs. (25) and (26) up to second-order as:

$$\begin{aligned} c_t = {} & c + c_k(k_t - k) + c_z z_t \\ & + \frac{1}{2} c_{kk}(k_t - k)^2 + c_{kz}(k_t - k) z_t + \frac{1}{2} c_{zz} z_t^2 + \frac{1}{2} c_{\sigma^2} \sigma^2 \end{aligned} \tag{27}$$

and

$$\begin{aligned} k_{t+1} = {} & k + k_k(k_t - k) + k_z z_t \\ & + \frac{1}{2} k_{kk}(k_t - k)^2 + k_{kz}(k_t - k) z_t + \frac{1}{2} k_{zz} z_t^2 + \frac{1}{2} k_{\sigma^2} \sigma^2. \end{aligned} \tag{28}$$

Since even with this simple neoclassical growth model the previous systems of equations are too involved to be written explicitly, we illustrate the procedure numerically. In Table 1, we summarize the parameter values for the four parameters of the model. We do not pretend to be selecting a realistic calibration (our choice of $\delta = 1$ precludes any attempt at matching observed data). Instead, we pick standard parameter values in the

Table 1 Calibration

Parameter	**Value**
β	0.99
α	0.33
ρ	0.95
η	0.01

literature. The discount factor, β, is 0.99, the elasticity of output with respect to capital, α, is 0.33, the persistence of the autoregressive process, ρ, is 0.95, and the standard deviation of the innovation, η, is 0.01. With this calibration, the steady state is $c = 0.388$ and $k = 0.188$.

The first-order components of the solution are (already selecting the stable solution):

$$\begin{aligned} c_k &= 0.680 \quad c_z = 0.388 \\ k_k &= 0.330 \quad k_z = 0.188 \end{aligned}$$

and the second-order components:

$$\begin{aligned} c_{kk} &= -2.420 \quad c_{kz} = 0.680 \quad c_{zz} = 0.388 \quad c_{\sigma\sigma} = 0 \\ k_{kk} &= -1.174 \quad k_{kz} = 0.330 \quad k_{zz} = 0.188 \quad k_{\sigma\sigma} = 0 \end{aligned}$$

In addition, recall that we have the theoretical results: $c_\sigma = k_\sigma = c_{k\sigma} = k_{k\sigma} = c_{z\sigma} = k_{z\sigma} = 0$. Thus, we get our second-order approximated solutions for the consumption decision rule:

$$\begin{aligned} c_t = {} & 0.388 + 0.680(k_t - 0.188) + 0.388 z_t \\ & - 1.210(k_t - 0.188)^2 + 0.680(k_t - 0.188) z_t + 0.194 z_t^2 \end{aligned}$$

and for the capital decision rule:

$$\begin{aligned} k_{t+1} = {} & 0.188 + 0.330\,(k_t - 0.188) + 0.188 z_t \\ & - 0.587(k_t - 0.188)^2 + 0.330\,(k_t - 0.188)\, z_t + 0.094 z_t^2. \end{aligned}$$

In this case, the correction for risk is zero. This should not be a surprise. In the neoclassical growth model, risk is production risk driven by technology shocks. This production risk is brought about by capital: the more capital the representative household accumulates, the more it exposes itself to production risk. At the same time, the only asset available for net saving in this economy is capital. Thus, any increment in risk (ie, a rise in the standard deviation of the technology shock) generates two counterbalancing mechanisms: a desire to accumulate more capital to buffer future negative shocks and a desire to accumulate less capital to avoid the additional production risk. For low values of risk aversion, both mechanisms nearly cancel each other (with a log utility function, they perfectly compensate each other: in the exact solution, the standard deviation of the innovation to the shock does not appear, only the realization of z_t). For higher values of risk aversion or for models with different assets (for instance, a model where the representative household can save in the form of an international bond whose payments are not perfectly correlated with the productivity shock within the country), the correction for risk can be quite different from zero.

The next step is to compare the exact and the approximated decision rules. With our calibration, the exact solution is given by:

$$c_t = 0.673 e^{z_t} k_t^{0.33}$$
$$k_{t+1} = 0.327 e^{z_t} k_t^{0.33}.$$

To gauge how close these two solutions are, we plot in Fig. 1 the exact decision rule for capital (continuous line in the top and bottom panels), the first-order approximation (discontinuous line in the top panel), and the second-order approximation (discontinuous line in the bottom panel). In both panels, we plot the decision rule for capital when $z_t = 0$ and for values of capital that are ± 25% of the value of capital in the steady state. The first-order approximation is nearly identical to the exact solution close to the steady state. Only farther away, do both solutions diverge. At the start of the grid (with $k = 0.1412$),

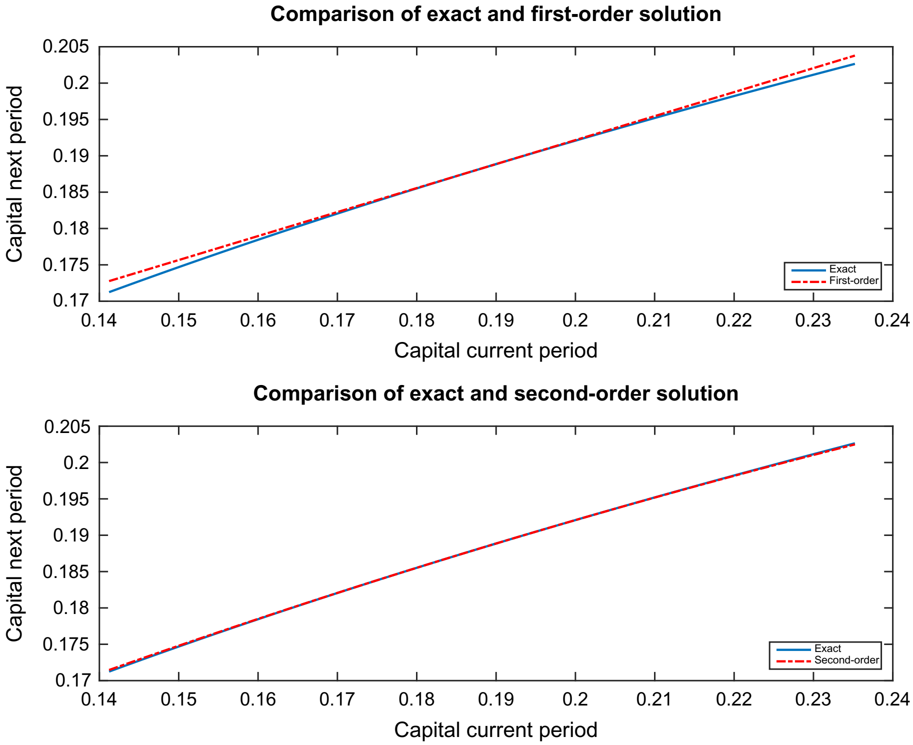


Fig. 1 Comparison of exact and perturbation solution.

the exact decision rule and the first-order approximation diverge by nearly 1%. The second-order approximation, in comparison, is more accurate along the whole range of values of capital. Even at $k = 0.1412$, the difference between both solutions is only 0.13%. This result shows the often good global properties of perturbation solutions.

We will revisit below how to assess the accuracy of a solution. Suffice it to say at this moment that whether 0.13% is too large or accurate enough is application dependent. For instance, for the computation of business cycle moments, we often need less accuracy than for welfare evaluations. The reason is that while errors in the approximation of a moment of the model, such as the mean or variance of consumption, tend to cancel each other, welfare is a nonlinear function of the allocation and small errors in computing an allocation can translate into large errors in computing welfare.

4.4 Pruning

Although the higher-order perturbations that we described are intuitive and straightforward to compute, they often generate explosive sample paths even when the corresponding linear approximation is stable. These explosive sample paths arise because the higher-order terms induce additional fixed points for the system, around which the approximated solution is unstable (see Kim et al., 2008 and Den Haan and De Wind, 2012). A simple example clarifies this point. Imagine that we have an approximated decision rule for capital (where, for simplicity, we have eliminated the persistence on the productivity process z_t) that has the form:

$$k_{t+1} = a_0 + a_1 k_t + a_2 k_t^2 + \ldots + b_1 \varepsilon_t + \ldots$$

If we substitute recursively, we find:

$$k_{t+1} = a_1 k_t + a_2 \left(a_1 k_{t-1} + a_2 k_{t-1}^2\right)^2 + \ldots + b_1 \varepsilon_t + \ldots,$$

an expression that involves terms in k_{t-1}^3 and k_{t-1}^4. If the support of ε_t is not bounded, sooner or later, we will have, in a simulation, an innovation large enough such that k_{t+1} is far away from its steady-state value. As the simulation progresses over time, that value of k_{t+1} will be raised to cubic and higher-order powers, and trigger an explosive path. The presence of this explosive behavior complicates any model evaluation because no unconditional moments would exist based on this approximation. It also means that any unconditional moment-matching estimation methods, such as the generalized method of moments (GMM) or the simulated method of moments (SMM), are inapplicable in this context as they rely on finite moments from stationary and ergodic probability distributions.

For second-order approximations, Kim et al. (2008) propose pruning the approximation. Loosely speaking, pruning means to eliminate, in the recursions, all terms that are of a higher order than the order of the solution (ie, if we are dealing with a second-order perturbation, all terms involving states or innovations raised to powers higher than 2). Kim et al. (2008) prove that the pruned approximation does not explode.

Andreasen et al. (2013) extend Kim et al. (2008)'s approach by showing how to apply pruning to an approximation of any arbitrary order by exploiting what the authors refer to as the pruned state-space system. Under general technical conditions, Andreasen et al. (2013) show that first and second unconditional moments for a pruned state-space system exist. Then, they provide closed-form expressions for first and second unconditional moments and impulse response functions. This is important because these expressions let researchers avoid the use of numerical simulations to compute these moments. These numerical simulations have often been shown to be unreliable, in particular, when solving for the generalized impulse response functions of DSGE models (for the definition of generalized impulse response functions, see Koop et al., 1996). Andreasen et al. (2013) also derive conditions for the existence of higher unconditional moments, such as skewness and kurtosis.

4.5 Change of Variables

In Remark 4, we discussed the possibility of performing the perturbation of a DSGE model in logs of the variables of interest, instead of doing it in levels. In a creative contribution, Judd (2003) argues that loglinearization is a particular case of the more general idea of a change of variables and shows how this technique could be efficiently implemented. In this subsection, we explain Judd's contribution by following Fernández-Villaverde and Rubio-Ramírez (2006).

The point of departure is to note that if we have a Taylor expansion of a variable x around a point a:

$$d(x) \simeq d(a) + \frac{\partial d(a)}{\partial a}(x-a) + H.O.T.,$$

(where $H.O.T.$ stands for higher-order terms), we can rewrite the expansion in terms of a transformed variable $Y(x)$:

$$g(y) = h(d(X(y))) = g(b) + \frac{\partial g(b)}{\partial b}(Y(x)-b) + H.O.T.$$

where $b = Y(a)$ and $X(y)$ is the inverse of $Y(x)$. Since with a perturbation we find a Taylor series approximation of the unknown function d that solves the operator $\mathcal{H}(\cdot)$ as a function of the states x, the change of variables means we can find an alternative Taylor series in terms of $Y(x)$.

Why do we want to perform this change of variables? The famous British meteorologist Eric Eady (1915–1966) remarked once that: "It is not the process of linearization that limits insight. It is the nature of the state that we choose to linearize about."

By picking the right change of variables, we can reshape a highly nonlinear problem into a much more linear one and, therefore, significantly increase the accuracy of the perturbation.[j]

4.5.1 A Simple Example

Imagine that our aim is to approximate the decision rule for capital in our workhorse stochastic neoclassical growth model with a first-order perturbation (the same ideas would apply if we are trying to approximate other decision rules, expectations, value functions, etc.). Remember that we derived that such an approximation had the form:

$$k_{t+1} = k + a_1(k_t - k) + b_1 z_t$$

where a and b are the coefficients that we find by taking derivatives of $F(k_t, z_t; \sigma)$ and k is the steady-state value of capital. In this section, it is more convenient to rewrite the decision rule as:

$$(k_{t+1} - k) = a_1(k_t - k) + b_1 z_t.$$

Analogously a loglinear approximation of the policy function will take the form:

$$\log k_{t+1} - \log k = a_2(\log k_t - \log k) + b_2 z$$

or in equivalent notation:

$$\hat{k}_{t+1} = a_2 \hat{k}_t + b_2 z_t$$

where $\hat{x} = \log x - \log x_0$ is the percentage deviation of the variable x with respect to its steady state.

How do we go from one approximation to the second one? First, we write the linear system in levels as:

$$k_{t+1} = d(k_t, z_t; \sigma) = d(k, 0; 0) + d_1(k, 0; 0)(k_t - k) + d_2(k, 0; 0) z_t$$

where $d(k, 0; 0) = k$, $d_1(k, 0; 0) = a_1$, $d_2(k, 0; 0) = b_1$. Second, we propose the changes of variables $h = \log d$, where $Y(x) = \log x$ and $Y(x) = \log x$. Third, we apply Judd (2003)'s formulae for this example:

$$\log k_{t+1} - \log k = d_1(k, 0, 0)(\log k_t - \log k) + \frac{1}{k} d_2(k, 0, 0) z$$

Finally, by equating coefficients, we obtain a simple closed-form relation between the parameters of both representations: $a_2 = a_1$ and $b_2 = \frac{1}{k} b_1$.

[j] An idea related to the change of variables is the use of gauges, where the perturbation is undertaken not in terms of powers of the perturbation parameter, σ, but of a series of gauge functions $\{\delta_n(\sigma)\}_{n=1}^{\infty}$ such that: $\lim_{n \to \infty} \frac{\delta_{n+1}(\sigma)}{\delta_n(\sigma)} = 0$. See Judd (1998) for details.

Three points are important. First, moving from a_1 and b_1 to a_2 and b_2 is an operation that only involves k, a value that we already know from the computation of the first-order perturbation in levels. Therefore, once the researcher has access to the linear solution, obtaining the loglinear one is immediate.[k] Second, we have not used any assumption on the utility or production functions except that they satisfy the general technical conditions of the stochastic neoclassical growth model. Third, the change of variables can be applied to a perturbation of an arbitrary order. We only presented the case for a first-order approximation to keep the exposition succinct.

4.5.2 A More General Case

We can now present a more general case of change of variables. The first-order solution of a model is:

$$d(x) \simeq d(a) + \frac{\partial d(a)}{\partial a}(x-a).$$

If we expand $g(y) = h(d(X(y)))$ around $b = Y(a)$, where $X(y)$ is the inverse of $Y(x)$, we can write:

$$g(y) = h(d(X(y))) = g(b) + g_\alpha(b)(Y^\alpha(x) - b^\alpha)$$

where $g_\alpha = h_A d_i^A X_\alpha^i$ comes from the application of the chain rule.

Following Judd (2003), we use this approach to encompass any power function approximation of the form:

$$k_{t+1}(k,z;\gamma,\zeta,\varphi)^\gamma - k^\gamma = a_3\left(k_t^\zeta - k^\zeta\right) + b_3 z^\varphi$$

where we impose $\varphi \geq 1$ to ensure that we have real values for the power z^φ. Power functions are attractive because, with only three free parameters (γ,ζ,φ), we can capture many nonlinear structures and nest the log transformation as the limit case when the coefficients γ and ζ tend to zero and $\varphi = 1$. The changes of variables for this family of functions are given by $h = d^\gamma$, $Y = x^\zeta$, and $X = y^{\frac{1}{\zeta}}$. Following the same reasoning as before, we derive:

$$k_{t+1}(k,z;\gamma,\zeta,\varphi)^\gamma - k^\gamma = \frac{\gamma}{\zeta}k^{\gamma-\zeta}a_1\left(k_t^\zeta - k^\zeta\right) + \frac{\gamma}{\varphi}k^{\gamma-1}b_1 z^\varphi.$$

The relation between the new and the old coefficients is again easy to compute: $a_3 = \frac{\gamma}{\zeta}k^{\gamma-\zeta}a_1$ and $b_3 = \frac{\gamma}{\varphi}k^{\gamma-1}b_1$.

[k] A heuristic argument that delivers the same result takes: $(k_{t+1} - k) = a_1(k_t - k) + b_1 z_t$ and divides on both sides by k: $\frac{k_{t+1}-k}{k} = a_1\frac{k_t - k}{k} + \frac{1}{k}b_1 z$. Noticing that $\frac{x_t - x}{x} \simeq \log x_t - \log x$, we get back the same relation as the one above. Our argument in the main text is more general and does not depend on an additional approximation.

A slightly more restrictive case is to impose that $\gamma = \zeta$ and $z = 1$. Then, we get a power function with only one free parameter γ:

$$k_{t+1}(k, z; \gamma)^{\gamma} - k^{\gamma} = a_4\left(k_t^{\zeta} - k^{\zeta}\right) + b_4 z$$

or, by defining $\widetilde{k}_t = k_t^{\gamma} - k^{\gamma}$, we get:

$$\widetilde{k}_{t+1} = a_4 \widetilde{k}_t + b_4 z$$

with $a_4 = a_1$ and $b_4 = k^{\gamma - 1} b_1$. This representation has the enormous advantage of being a linear system, which makes it suitable for analytic study and, as we will see in Section 10, for estimation with a Kalman filter.

4.5.3 The Optimal Change of Variables

The previous subsection showed how to go from a first-order approximation to the solution of a DSGE model to a more general representation indexed by some parameters. The remaining question is how to select the optimal value of these parameters.[1]

Fernández-Villaverde and Rubio-Ramírez (2006) argue that a reasonable criterion (and part of the motivation for the change of variables) is to select these parameters to improve the accuracy of the solution of the model. More concretely, the authors propose to minimize the Euler error function with respect to some metric. Since we have not introduced the measures of accuracy of the solution to a DSGE model, we will skip the details of how to do so. Suffice it to say that Fernández-Villaverde and Rubio-Ramírez (2006) find that the optimal change of variables improves the average accuracy of the solution by a factor of around three. This improvement makes a first-order approximation competitive in terms of accuracy with much more involved methods. Fernández-Villaverde and Rubio-Ramírez (2006) also report that the optimal parameter values depend on the standard deviation of the exogenous shocks to the economy. This is a significant result: the change of variables corrects by the level of uncertainty existing in the economy and breaks certainty equivalence.

Remark 15 (Loglinearization v. lognormal–loglinear approximation) A different solution technique, called lognormal–loglinear approximation, is popular in finance. Its relation with standard loglinearization (as a particular case of first-order perturbation with a change of variables in logs) often causes confusion among researchers and students. Thus, once we have understood the change of variables technique, it is worthwhile to dedicate this remark to clarifying the similarities and differences between the first-order

[1] We do not even need to find the optimal value of these parameters. It may be the case that a direct but not optimal choice of parameter values already delivers substantial improvements in accuracy at a very low computational cost. When one is maximizing, for example, a likelihood function, being at the true maximum matters. When one is finding parameters that improve accuracy, optimality is desirable but not essential, and it can be traded off against computational cost.

perturbation in logs and the lognormal–loglinear approximation. The best way to illustrate this point is with a concrete example. Imagine that we have a household with utility function

$$\max \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \log C_t$$

and budget constraint:

$$W_{t+1} = R_{t+1}(W_t - C_t)$$

where W_t is total wealth and W_0 is given. Then, the optimality conditions are:

$$1 = \beta \mathbb{E}_t \frac{C_t}{C_{t+1}} R_{t+1}$$
$$W_{t+1} = R_{t+1}(W_t - C_t)$$

with steady state $R = \frac{1}{\beta}$ and $W = R(W - C)$.

Under a standard first-order perturbation in logs (loglinearization) around the previous steady state, and after some algebra:

$$\mathbb{E}_t \Delta \hat{c}_{t+1} = \mathbb{E}_t \hat{r}_{t+1}$$
$$\hat{w}_{t+1} = \hat{r}_{t+1} + \frac{1}{\rho}\hat{w}_t + \left(1 - \frac{1}{\rho}\right)\hat{c}_t$$

where, for a variable X_t,

$$\hat{x}_t = x_t - x = \log X_t - \log X$$

and $\rho = \frac{W - C}{W}$. Subtracting $\hat{w}_t$ from the second equation:

$$\Delta \hat{w}_{t+1} = \hat{r}_{t+1} + \left(1 - \frac{1}{\rho}\right)(\hat{c}_t - \hat{w}_t)$$

If we want to express these two equations in logs, instead of log-deviations (and using the fact that $r = -\log\beta$):

$$\mathbb{E}_t \Delta c_{t+1} = \log\beta + \mathbb{E}_t r_{t+1} \tag{29}$$

$$\Delta w_{t+1} = r_{t+1} + k + \left(1 - \frac{1}{\rho}\right)(c_t - w_t) \tag{30}$$

where

$$k = -r - \left(1 - \frac{1}{\rho}\right)(c - w).$$

In comparison, a lognormal–loglinearization still uses the approximation of the budget constraint (30), but it assumes that $\frac{C_t}{C_{t+1}}R_{t+1}$ is distributed as a lognormal random variable. Since, for an arbitrary variable:

$$\log \mathbb{E}_t X_t = \mathbb{E}_t \log X_t + \frac{1}{2} Var_t \log X_t,$$

we can go back to the Euler equation

$$1 = \beta \mathbb{E}_t \frac{C_t}{C_{t+1}} R_{t+1}$$

and rewrite it as:

$$\begin{aligned} 0 &= \log\beta + \log \mathbb{E}_t \frac{C_t}{C_{t+1}} R_{t+1} \\ &= \log\beta + \mathbb{E}_t \log \frac{C_t}{C_{t+1}} R_{t+1} + \frac{1}{2} Var_t \log \frac{C_t}{C_{t+1}} R_{t+1} \end{aligned}$$

or, rearranging terms:

$$\mathbb{E}_t \Delta c_{t+1} = \log\beta + \mathbb{E}_t r_{t+1} + \frac{1}{2}\left[Var_t \Delta c_{t+1} + Var_t r_{t+1} - 2cov_t(\Delta c_{t+1}, r_{t+1})\right] \tag{31}$$

More in general, in a lognormal–loglinearization, we approximate the nonexpectational equations with a standard loglinearization and we develop the expectational ones (or at least the ones with returns on them) using a lognormal assumption. In particular, we do not approximate the Euler equation. Once we have assumed that $\frac{C_t}{C_{t+1}}R_{t+1}$ is lognormal, all the results are exact.

If we compare the two equations for the first difference of consumption, (29) and (31), we see that the lognormal–loglinear approximation introduces an additional term

$$\frac{1}{2}\left[Var_t \Delta c_{t+1} + Var_t r_{t+1} - 2cov_t(\Delta c_{t+1}, r_{t+1})\right]$$

that breaks certainty equivalence. This novel feature has important advantages. For example, for a pricing kernel M_t and an asset i, we have the pricing equation:

$$1 = \mathbb{E}_t M_{t+1} R_{i,t+1}.$$

Then:

$$0 = \mathbb{E}_t \log M_{t+1} R_{i,t+1} + \frac{1}{2} Var_t \log M_{t+1} R_{i,t+1}$$

or:

$$\mathbb{E}_t r_{i,t+1} = -\mathbb{E}_t m_{t+1} - \frac{1}{2} Var_t m_{t+1} - \frac{1}{2} Var_t r_{i,t+1} - cov_t\left(m_{t+1}, r_{i,t+1}\right)$$

If we look at the same expression for the risk-free bond:

$$1 = \mathbb{E}_t M_{t+1} R_{f,t+1}$$

we get:

$$r_{f,t+1} = -\mathbb{E}_t m_{t+1} - \frac{1}{2} Var_t m_{t+1}$$

and we can find that the excess return is:

$$\mathbb{E}_t r_{i,t+1} - r_{f,t+1} = -\frac{1}{2} Var_t r_{i,t+1} - cov_t\left(m_{t+1}, r_{i,t+1}\right),$$

an expression that it is easy to interpret.

On the other hand, this expression also embodies several problems. First, it is often unclear to what extent, in a general equilibrium economy, $\frac{C_t}{C_{t+1}} R_{t+1}$ is close to lognormality. Second, in lognormal–loglinear approximation, we are mixing two approaches, a lognormal assumption with a loglinearization. This is not necessarily coherent from the perspective of perturbation theory and we may lack theoretical foundations for the approach (including an absence of convergence theorems). Third, in the loglinearization, we can compute all the coefficients by solving a quadratic matrix system. In the lognormal–loglinear approximation, we need to compute second moments and, in many applications, how to do so may not be straightforward. Finally, it is not obvious how to get higher-order approximations with the lognormal–loglinear approximation, while perturbation theory can easily handle higher-order solutions.

4.6 Perturbing the Value Function

In some applications, it is necessary to perturb the value function of a DSGE model, for example, when we are dealing with recursive preferences or when we want to evaluate welfare. Furthermore, a perturbed value function can be an outstanding initial guess for value function iteration, making it possible to deal with high-dimensional problems that could be otherwise too slow to converge. Given the importance of perturbing the value function, this section illustrates in some detail how to do so.

Since all that we learned in the general case subsection will still apply by just changing the operator $\mathcal{H}$ from the equilibrium conditions to the Bellman operator, we can go directly to a concrete application. Consider a value function problem (following the same notation as above).

$$V(k_t, z_t) = \max_{c_t} \left[(1-\beta)\log c_t + \beta \mathbb{E}_t V(k_{t+1}, z_{t+1})\right]$$
$$\text{s.t. } c_t + k_{t+1} = e^{z_t} k_t^{\alpha} + (1-\delta)k_t$$
$$z_t = \rho z_{t-1} + \eta \varepsilon_t,\ \varepsilon_t \sim N(0,1)$$

where we have "normalized" $\log c_t$ by $(1-\beta)$ to make the value function and the utility function have the same order of magnitude (thanks to normalization, $V_{ss} = \log c$, where V_{ss} is the steady-state value function and c is the steady-state consumption).

We can rewrite the problem in terms of a perturbation parameter σ:

$$V(k_t, z_t; \sigma) = \max_{c_t} \left[\log c_t + \beta \mathbb{E}_t V\left(e^{z_t} k_t^{\alpha} + (1-\delta)k_t - c_t, \rho z_t + \sigma \eta \varepsilon_{t+1}; \sigma\right)\right].$$

Note that we have made explicit the dependencies in the next period states from the current period state. The perturbation solution of this problem is a value function $V(k_t, z_t; \sigma)$ and a policy function for consumption $c(k_t, z_t; \sigma)$. For example, the second-order Taylor approximation of the value function around the deterministic steady state $(k, 0; 0)$ is:

$$\begin{aligned} V(k_t, z_t; \sigma) = {} & V_{ss} + V_{1,ss}\,(k_t - k) + V_{2,ss} z_t + V_{3,ss}\sigma \\ & + \frac{1}{2} V_{11,ss}(k_t - k)^2 + V_{12,ss}(k_t - k) z_t + V_{13,ss}(k_t - k)\sigma \\ & + \frac{1}{2} V_{22,ss} z_t^2 + V_{23,ss} z_t \sigma + \frac{1}{2} V_{33,ss}\sigma^2 \end{aligned}$$

where:

$$V_{ss} = V(k, 0; 0)$$
$$V_{i,ss} = V_i(k, 0; 0) \text{ for } i = \{1,2,3\}$$
$$V_{ij,ss} = V_{ij}(k, 0; 0) \text{ for } i, j = \{1,2,3\}$$

By certainty equivalence:

$$V_{3,ss} = V_{13,ss} = V_{23,ss} = 0$$

and then:

$$\begin{aligned} V(k_t, z_t; 1) = {} & V_{ss} + V_{1,ss}\,(k_t - k) + V_{2,ss} z_t \\ & + \frac{1}{2} V_{11,ss}\,(k_t - k)^2 + \frac{1}{2} V_{22,ss} z_{tt}^2 + V_{12,ss}\,(k_t - k) z + \frac{1}{2} V_{33,ss}\sigma^2 \end{aligned}$$

Note that $V_{33,ss} \neq s0$, a difference from the LQ approximation to the utility function that we discussed in Remark 13.

Similarly, the policy function for consumption can be expanded as:

$$c_t = c(k_t, z_t; \sigma) = c_{ss} + c_{1,ss}(k_t - k) + c_{2,ss} z_t + c_{3,ss}\sigma$$

where $c_{i,ss} = c_1(k, 0; 0)$ for $i = \{1,2,3\}$. Since the first derivatives of the consumption function only depend on the first and second derivatives of the value function, we must have that $c_{3,ss} = 0$ (remember that precautionary consumption depends on the third derivative of the value function; Kimball, 1990).

To find the linear components of our approximation to the value function, we take derivatives of the value function with respect to controls (c_t), states (k_t, z_t), and the perturbation parameter σ and solve the associated system of equations when $\sigma = 0$. We can find the quadratic components of the value function by taking second derivatives, plugging in the known components from the previous step, and solving the system when $\sigma = 0$.

We are ready now to show some of the advantages of perturbing the value function. First, we have an evaluation of the welfare cost of business cycle fluctuations readily available. At the deterministic steady state $k_t = k$ and $z_t = 0$, we have:

$$V(k, 0; \sigma) = V_{ss} + \frac{1}{2} V_{33,ss}\sigma^2.$$

Hence $\frac{1}{2} V_{33,ss}\sigma^2$ is a measure of the welfare cost of the business cycle: it is the difference, up to second-order, between the value function evaluated at the steady-state value of the state variables $(k, 0)$ and the steady-state value function (where not only are we at the steady state, but where we know that in future periods we will be at that point as well). Note that this last quantity is not necessarily negative. Indeed, it may well be positive in many models, such as in a stochastic neoclassical growth model with leisure choice. For an explanation and quantitative evidence, see Cho et al. (2015).[m]

It is easier to interpret $V_{33,ss}$ if we can transform it into consumption units. To do so, we compute the decrease in consumption τ that will make the household indifferent between consuming $(1-\tau)c$ units per period with certainty or c_t units with uncertainty. That is, τ satisfies:

$$\log(1-\tau)c = \log c + \frac{1}{2} V_{33,ss}\sigma^2$$

where we have used $V_{ss} = \log c$. Then,

$$\tau = 1 - e^{\frac{1}{2} V_{33,ss}\sigma^2}.$$

[m] In his classical calculation about the welfare cost of the business cycle, Lucas Jr. (1987) assumed an endowment economy, where the representative household faces the same consumption process as the one observed for the US economy. Thus, for any utility function with risk aversion, the welfare cost of the business cycle must be positive (although Lucas' point, of course, was that it was rather small). When consumption and labor supply are endogenous, agents can take advantage of uncertainty to increase their welfare. A direct utility function that is concave in allocations can generate a convex indirect utility function on prices and those prices change in general equilibrium as a consequence of the agents' responses to uncertainty.

We close this section with a numerical application. To do so, we pick the same calibration as in Table 1. We get:

$$V(k_t, z_t; 1) = -0.54000 + 0.026(k_t - 0.188) + 0.250z_t - 0.069(k_t - 0.188)^2 \tag{32}$$

(where, for this calibration, $V_{kz} = V_{z^2} = V_{\sigma^2} = 0$) and:

$$c(k_t, z_t; \chi) = 0.388 + 0.680(k_t - 0.188) + 0.388z_t,$$

which is the same approximation to the consumption decision rule we found when we tackled the equilibrium conditions of the model. For this calibration, the welfare cost of the business cycle is zero.[n]

We can also use Eq. (32) as an initial guess for value function iteration. Thanks to it, instead of having to iterate hundreds of times, as if we were starting from a blind initial guess, value function iteration can converge after only a few dozen interactions.

Finally, a mixed strategy is to stack both the equilibrium conditions of the model and the value function evaluated at the optimal decision rules:

$$V(k_t, z_t) = (1-\beta)\log c_t + \beta\mathbb{E}_t V(k_{t+1}, z_{t+1}).$$

in the operator $\mathcal{H}$. This strategy delivers an approximation to the value function and the decision rules with a trivial cost.[o]

5. PROJECTION

Projection methods (also known as weighted residual methods) handle DSGE models by building a function indexed by some coefficients that approximately solves the operator $\mathcal{H}$. The coefficients are selected to minimize a residual function that evaluates how far away the solution is from generating a zero in $\mathcal{H}$. More concretely, projection methods solve:

$$\mathcal{H}(d) = \mathbf{0}$$

[n] Recall that the exact consumption decision rule is $c_t = 0.673e^{z_t}k_t^{0.33}$. Since the utility function is log, the period utility from this decision rule is $\log c_t = z_t + \log 0.673 + 0.33\log k_t$. The unconditional mean of z_t is 0 and the capital decision rule is certainty equivalent in logs. Thus, there is no (unconditional) welfare cost of changing the variance of z_t.

[o] We could also stack derivatives of the value function, such as:

$$(1-\beta)c_t^{-1} - \beta\mathbb{E}_t V_{1,t+1} = 0$$

and find the perturbation approximation to the derivative of the value function (which can be of interest in itself or employed in finding higher-order approximations of the value function).

by specifying a linear combination:

$$d^j(\mathbf{x}|\theta) = \sum_{i=0}^{j} \theta_i \Psi_i(\mathbf{x}) \tag{33}$$

of basis function $\Psi_i(\mathbf{x})$ given coefficients $\theta = \{\theta_0, \ldots, \theta_j\}$. Then, we define a residual function:

$$R(\mathbf{x}|\theta) = \mathcal{H}\left(d^j(\mathbf{x}|\theta)\right)$$

and we select the values of the coefficients θ that minimize the residual given some metric. This last step is known as "projecting" $\mathcal{H}$ against that basis to find the components of θ (and hence the name of the method).

Inspection of Eq. (33) reveals that to build the function $d^j(\mathbf{x}|\theta)$, we need to pick a basis $\{\Psi_i(\mathbf{x})\}_{i=0}^{\infty}$ and decide which inner product we will use to "project" $\mathcal{H}$ against that basis to compute θ. Different choices of bases and of the projection algorithm will imply different projection methods. These alternative projections are often called in the literature by their own particular names, which can be sometimes bewildering.

Projection theory, which has been applied in *ad hoc* ways by economists over the years, was popularized as a rigorous approach in economics by Judd (1992) and Gaspar and Judd (1997) and, as in the case of perturbation, it has been authoritatively presented by Judd (1998).[p]

Remark 16 (Linear v. nonlinear combinations) Instead of linear combinations of basis functions, we could deal with more general nonlinear combinations:

$$d^j(\mathbf{x}|\theta) = f\left(\{\Psi_i(\mathbf{x})\}_{i=0}^{j} | \theta\right)$$

for a known function f. However, the theory for nonlinear combinations is less well developed, and we can already capture a full range of nonlinearities in d^j with the appropriate choice of basis functions Ψ_i. In any case, it is more pedagogical to start with the linear combination case. Most of the ideas in the next pages carry over the case of nonlinear combinations. The fact that we are working with linear combinations of basis functions also means that, in general, we will have the same number of coefficients θ as the number of basis functions Ψ_i times the dimensionality of d^j.

[p] Projection theory is more modern than perturbation. Nevertheless, projection methods have been used for many decades in the natural sciences and engineering. Spectral methods go back, at least, to Lanczos (1938). Alexander Hrennikoff and Richard Courant developed the finite elements method in the 1940s, although the method was christened by Clough (1960), who made pioneering contributions while working at Boeing. See Clough and Wilson (1999) for a history of the early research on finite elements.

5.1 A Basic Projection Algorithm

Conceptually, projection is easier to present than perturbation (although its computational implementation is harder). We can start directly by outlining a projection algorithm:

Algorithm 1 (Projection Algorithm)

1. Define $j+1$ known linearly independent functions $\psi_i : \Omega \rightarrow \mathbb{R}$ where $j < \infty$. We call the $\psi_0, \psi_1, \ldots, \psi_j$ the *basis functions*. These basis functions depend on the vector of state variables x.
2. Define a vector of coefficients $\theta^l = [\theta^l_0, \theta^l_1, \ldots, \theta^l_j]$ for $l = 1, \ldots, m$ (where recall that m is the dimension that the function d of interest maps into). Stack all coefficients on a $m \times (j+1)$ matrix $\theta = [\theta^1; \theta^2; \ldots; \theta^l]$.
3. Define a combination of the basis functions and the θ's:

$$d^{l,j}(\cdot|\theta^l) = \sum_{i=0}^{j} \theta^l_i \psi_i(\cdot)$$

for $l = 1, \ldots, m$. Then:

$$d^j(\cdot|\theta) = \left[d^{1,j}(\cdot|\theta^l); d^{2,j}(\cdot|\theta^l); \ldots; d^{m,j}(\cdot|\theta^l)\right].$$

4. Plug $d^j(\cdot|\theta)$ into the operator H(·) to find the *residual equation*:

$$R(\cdot|\theta) = \mathcal{H}\left(d^j(\cdot|\theta)\right).$$

5. Find the value of θˆthat makes the residual equation as close to $\mathbf{0}$ as possible given some objective function $\rho : J^2 \times J^2 \rightarrow \mathbb{R}$:

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^{m \times (j+1)}} \rho(R(\cdot|\theta), \mathbf{0}).$$

To ease notation, we have made two simplifications on the previous algorithm. First, we assumed that, along each dimension of d, we used the same basis functions ψ_i and the same number $j+1$ of them. Nothing forces us to do so. At the mere cost of cumbersome notation, we could have different basis functions for each dimension and a different number of them (ie, different j's). While the former is not too common in practice, the latter is standard, since some variables' influence on the function d can be harder to approximate than others'.[q]

We specify a metric function ρ to gauge how close the residual function is to zero over the domain of the state variables. For example, in Fig. 2, we plot two different residual

[q] For the nonlinear combination case, $f\left(\{\Psi_i(\mathbf{x})\}_{i=0}^{j}|\theta\right)$, we would just write the residual function:

$$R(\cdot|\theta) = \mathcal{H}\left(f\left(\{\Psi_i(\mathbf{x})\}_{i=0}^{j}|\theta\right)\right)$$

and find the θ's that minimize a given metric. Besides the possible computational complexities of dealing with arbitrary functions $f\left(\{\Psi_i(\mathbf{x})\}_{i=0}^{j}|\theta\right)$, the conceptual steps are the same.

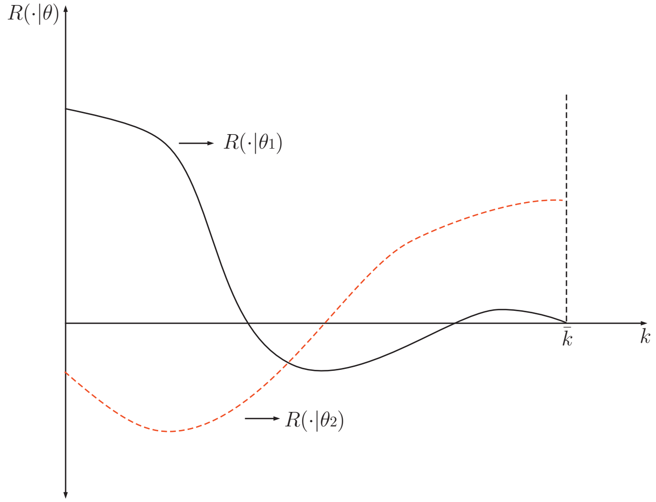


Fig. 2 Residual functions.

functions for a problem with only one state variable k_t (think, for instance, of a deterministic neoclassical growth model) that belongs to the interval $[0, \bar{k}]$, one for coefficients θ_1 (continuous line) and one for coefficients θ_2 (discontinuous line). $R(\cdot|\theta_1)$ has large values for low values of k_t, but has small values for high levels of k_t. $R(\cdot|\theta_2)$ has larger values on average, but it never gets as large as $R(\cdot|\theta_1)$. Which of the two residual functions is closer to zero over the interval? Obviously, different choices of ρ will yield different answers. We will discuss below how to select a good ρ.

A small example illustrates the previous steps. Remember that we had, for the stochastic neoclassical growth model, the system built by the Euler equation and the resource constraint of the economy:

$$\mathcal{H}(d) = \left\{ \begin{array}{c} u'\left(d^1(k_t, z_t)\right) \\ -\beta \mathbb{E}_t \left[u'\left(d^1\left(d^2(k_t, z_t), z_{t+1}\right)\right)\left(\alpha e^{\rho z_t + \sigma \varepsilon_{t+1}}\left(d^2(k_t, z_t)\right)^{\alpha-1} + 1 - \delta\right)\right] \\ d^1(k_t, z_t) + d^2(k_t, z_t) - e^{z_t} k_t^{\alpha} - (1-\delta)k_t \end{array} \right. = \mathbf{0},$$

for all k_t and z_t and where:

$$c_t = d^1(k_t, z_t)$$
$$k_{t+1} = d^2(k_t, z_t)$$

and we have already recursively substituted k_{t+1} in the decision rule of consumption evaluated at $t + 1$. Then, we can define

$$c_t = d^{1,j}\left(k_t, z_t | \theta^1\right) = \sum_{i=0}^{j} \theta_i^1 \psi_i(k_t, z_t)$$

and

$$k_{t+1} = d^{2,j}\left(k_t, z_t|\theta^2\right) = \sum_{i=0}^{j} \theta_i^2 \psi_i(k_t, z_t)$$

for some $\psi_0(k_t,z_t), \psi_1(k_t,z_t), \ldots, \psi_j(k_t,z_t)$. Below we will discuss which basis functions we can select for this role.

The next step is to write the residual function:

$$R(k_t, z_t|\theta) = \left\{ \begin{array}{c} u'\left(\sum_{i=0}^{j} \theta_i^1 \psi_i(k_t, z_t)\right) \\ -\beta \mathbb{E}_t \left[\begin{array}{c} u'\left(\sum_{i=0}^{j} \theta_i^1 \psi_i\left(\sum_{i=0}^{j} \theta_i^2 \psi_n(k_t, z_t), \rho z_t + \sigma \varepsilon_{t+1}\right)\right) \\ \times \left(\alpha e^{\rho z_t + \sigma \varepsilon_{t+1}} \left(\sum_{i=0}^{j} \theta_i^2 \psi_i(k_t, z_t)\right)^{\alpha - 1} + 1 - \delta\right) \end{array} \right] \\ \sum_{i=0}^{j} \theta_i^1 \psi_i(k_t, z_t) + \sum_{i=0}^{j} \theta_i^2 \psi_i(k_t, z_t) - e^{z_t} k_t^{\alpha} - (1-\delta) k_t \end{array} \right. ,$$

for all k_t and z_t, $\theta = \left[\theta^1; \theta^2\right]$.

The final step is to find $\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^{m \times (j+1)}} \rho(R(\cdot|\theta), \mathbf{0})$. Again, we will discuss these choices below in detail, but just for concreteness, let us imagine that we pick $m \times (j+1)$ points (k_l, z_l) and select the metric function to be zero at each of these $m \times (j+1)$ points and one everywhere else. Such a metric is trivially minimized if we make the residual function equal to zero exactly on those points. This is equivalent to solving the system of $m \times (j+1)$ equations:

$$R(k_l, z_l|\theta) = \mathbf{0}, \text{ for } l = 1, \ldots, m \times (j+1)$$

with $m \times (j+1)$ unknowns (we avoid here the discussion about the existence and uniqueness of such a solution).

Remark 17 (Relation to econometrics) Many readers will be familiar with the use of the word "projection" in econometrics. This is not a coincidence. A common way to present linear regression is to think about the problem of searching for the unknown conditional expectation function:

$$\mathbb{E}(Y|X)$$

for some variables Y and X. Given that this conditional expectation is unknown, we can approximate it with the first two monomials on X, 1 (a constant) and X (a linear function), and associated coefficients θ_0 and θ_1:

$$\mathbb{E}(Y|X) \simeq \theta_0 + \theta_1 X.$$

These two monomials are the first two elements of a basis composed by the monomials (and also of the Chebyshev polynomials, a basis of choice later in this section). The residual function is then:

$$R(Y,X|\theta_0,\theta_1) = Y - \theta_0 - \theta_1 X.$$

The most common metric in statistical work is to minimize the square of this residual:

$$R(Y,X|\theta_{0},\theta_1)^2$$

by plugging in the observed series $\{Y,X\}_{t=1:T}$. The difference, thus, between ordinary least squares and the projection algorithm is that while in the former we use observed data, in the latter we use the operator $\mathcal{H}(d)$ imposed by economic theory. This link is even clearer when we study the econometrics of semi-nonparametric methods, such as sieves (Chen, 2007), which look for flexible basis functions indexed by a low number of coefficients and that, nevertheless, impose fewer restrictions than a linear regression.

Remark 18 (Comparison with other methods) From our short description of projection methods, we can already see that other algorithms in economics are particular cases of it. Think, for example, about the parameterized expectations approach (Marcet and Lorenzoni, 1999). This approach consists of four steps.

First, the conditional expectations that appear in the equilibrium conditions of the model are written as a flexible function of the state variables of the model and some coefficients. Second, the coefficients are initialized at an arbitrary value. Third, the values of the coefficients are updated by running a nonlinear regression that minimizes the distance between the conditional expectations forecasted by the function guessed in step 1 and the actual realization of the model along a sufficiently long simulation. Step 3 is repeated until the coefficient values used to simulate the model and the coefficient values that come out of the nonlinear regression are close enough.

Step 1 is the same as in any other projection method: the function of interest (in this case the conditional expectation) is approximated by a flexible combination of basis functions. Often the parameterized expectations approach relies on monomials to do so (or functions of the monomials), which, as we will argue below, is rarely an optimal choice. But this is not an inherent property of the approach. Christiano and Fisher (2000) propose to use functions of Chebyshev polynomials, which will yield better results. More important is the iterative procedure outlined by steps 2–4. Finding the fixed point of the values of the coefficients by simulation and a quadratic distances is rarely the best option. Even if, under certain technical conditions (Marcet and Marshall, 1994) the algorithm converges, such convergence can be slow and fragile. In the main text, we will explain that a collocation approach can achieve the same goal much more efficiently and without having to resort to simulation (although there may be concrete cases where simulation is a superior strategy).

Value function iteration and policy function iteration can also be understood as particular forms of projection, where the basis functions are linear functions (or higher-order interpolating functions such as splines). Since in this chapter we are not dealing with these methods, we skip further details.

5.2 Choice of Basis and Metric Functions

The previous subsection highlighted the two issues ahead of us: how to decide which basis $\psi_0, \psi_1, \ldots, \psi_j$ to select and which metric function ρ to use. Different choices in each of these issues will result in slightly different projection methods, each with its weaknesses and strengths.

Regarding the first issue, we can pick a global basis (ie, basis functions that are nonzero and smooth for most of the domain of the state variable Ω) or a local basis (ie, basis functions that are zero for most of the domain of the state variable, and nonzero and smooth for only a small portion of the domain Ω). Projection methods with a global basis are often known as spectral methods. Projection methods with a local basis are also known as finite elements methods.

5.3 Spectral Bases

Spectral techniques were introduced in economics by Judd (1992). The main advantage of this class of global basis functions is their simplicity: building and working with the approximation will be straightforward. The main disadvantage of spectral bases is that they have a hard time dealing with local behavior. Think, for instance, about Fig. 3, which plots the decision rule $k_{t+1} = d(k_t)$ that determines capital tomorrow given capital today for some model that implies a nonmonotone, local behavior represented by the hump in the middle of the capital range (perhaps due to a complicated incentive constraint). The change in the coefficients θ required to capture that local shape of d would leak into the approximation for the whole domain Ω. Similar local behavior appears when we deal with occasionally binding constraints, kinks, or singularities.

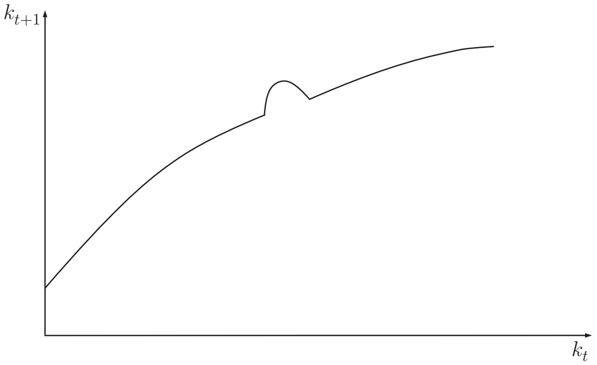


Fig. 3 Decision rule for capital.

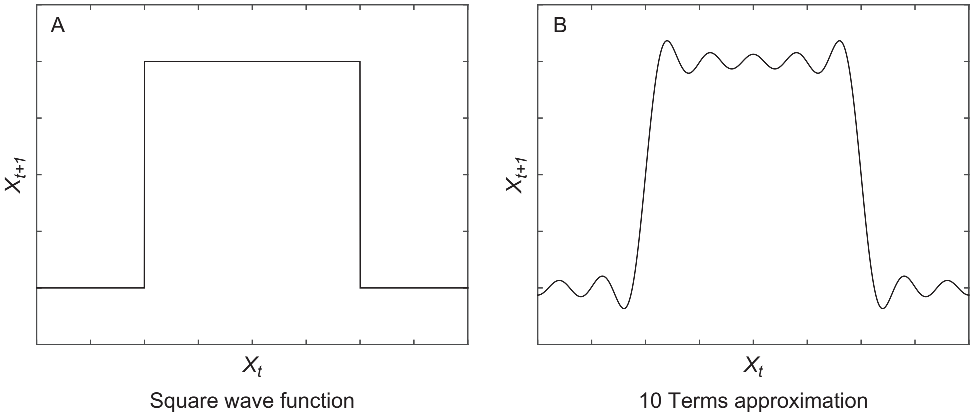


Fig. 4 Gibbs phenomenon.

A well-known example of this problem is the Gibbs phenomenon. Imagine that we are trying to approximate a piecewise continuously differentiable periodic function with a jump discontinuity, such as a square wave function (Fig. 4, panel A):

$$f(x) = \begin{cases} \frac{\pi}{4}, \text{ if } x \in [2j\pi, 2(j+1)\pi] \text{ and for } \forall j \in \mathbb{N} \\ -\frac{\pi}{4}, \text{ otherwise.} \end{cases}$$

Given that the function is periodic, a sensible choice for a basis is a trigonometric series $\sin(x)$, $\sin(2x)$, $\sin(3x)$, … The optimal approximation is:

$$\sin(x) + \frac{1}{3}\sin(3x) + \frac{1}{5}\sin(5x) + \ldots$$

The approximation behaves poorly at a jump discontinuity. As shown in Fig. 4, panel B, even after using 10 terms, the approximation shows large fluctuations around all the discontinuity points $2j\pi$ and $2(j+1)\pi$. These fluctuations will exist even if we keep adding many more terms to the approximation. In fact, the rate of convergence to the true solution as $n \to \infty$ is only $O(n)$.

5.3.1 Unidimensional Bases

We will introduce in this subsection some of the most common spectral bases. First, we will deal with the unidimensional case where there is only one state variable. This will allow us to present most of the relevant information in a succinct fashion. It would be important to remember, however, that our exposition of unidimensional bases cannot be exhaustive (for instance, in the interest of space, we will skip splines) and that the researcher may find herself tackling a problem that requires a specific basis. One of the great advantages of projection methods is their flexibility to accommodate unexpected requirements. In the next subsection, we will deal with the case of an arbitrary number of state variables and we will discuss how to address the biggest challenge of projection methods: the curse of dimensionality.

5.3.1.1 Monomials

A first basis is the monomials 1, x, x^2, x^3, … Monomials are simple and intuitive. Furthermore, even if this basis is not composed by orthogonal functions, if J_1 is the space of bounded measurable functions on a compact set, the Stone–Weierstrass theorem tells us that we can uniformly approximate any continuous function defined on a closed interval with linear combinations of these monomials.

Rudin (1976, p. 162) provides a formal statement of the theorem:

Theorem 1 (Stone–Weierstrass) *Let $\mathcal{A}$ be an algebra of real continuous functions on a compact set* K. *If $\mathcal{A}$ separates points on* K *and if $\mathcal{A}$ vanishes at no point of* K, *then the uniform closure $\mathcal{B}$ of $\mathcal{A}$ consists of all real continuous functions on* K.

A consequence of this theorem is that if we have a real function f that is continuous on K, we can find another function $h \in \mathcal{B}$ such that for $\varepsilon > 0$:

$$|f(x) - h(x)| < \varepsilon,$$

for all $x \in K$.

Unfortunately, monomials suffer from two severe problems. First, monomials are (nearly) multicollinear. Fig. 5 plots the graphs of x^{10} (continuous line) and x^{11} (discontinuous line) for $x \in [0.5, 1.5]$. Both functions have a very similar shape. As we add higher monomials, the new components of the solution do not allow the distance between the exact function we want to approximate and the computed approximation to diminish sufficiently fast.[r]

Second, monomials vary considerably in size, leading to scaling problems and the accumulation of numerical errors. We can also see this point in Fig. 5: x^{11} goes from $4.8828e^{-04}$ to 86.4976 just by moving x from 0.5 to 1.5.

The challenges presented by the use of monomials motivate the search for an orthogonal basis in a natural inner product that has a bounded variation in range. Orthogonality will imply that when we add more one element of the basis (ie, when we go from order j

[r] A sharp case of this problem is when $\mathcal{H}(\cdot)$ is linear. In that situation, the solution of the projection involves the inversion of matrices. When the basis functions are similar, the condition numbers of these matrices (the ratio of the largest and smallest absolute eigenvalues) are too high. Just the first six monomials can generate condition numbers of 10^{10}. In fact, the matrix of the least squares problem of fitting a polynomial of degree 6 to a function (the *Hilbert Matrix*) is a popular test of numerical accuracy since it maximizes rounding errors. The problem of the multicollinearity of monomials is also well appreciated in econometrics.

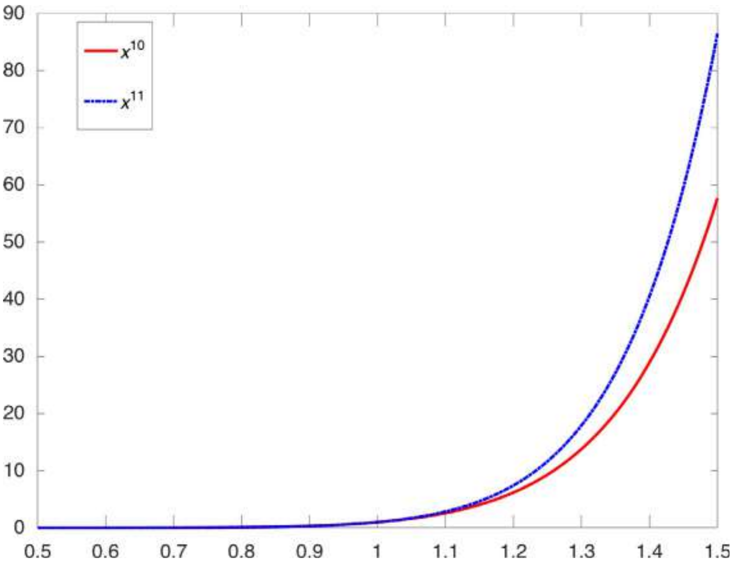


Fig. 5 Graphs of x^{10} and x^{11}.

to order $j + 1$), the newest element brings a sufficiently different behavior so as to capture features of the unknown function d not well approximated by the previous elements of the basis.

5.3.1.2 Trigonometric Series

A second basis is a trigonometric series

$$1/(2\pi)^{0.5}, \cos x/(2\pi)^{0.5}, \sin x/(2\pi)^{0.5}, \ldots,$$
$$\cos kx/(2\pi)^{0.5}, \sin kx/(2\pi)^{0.5}, \ldots$$

Trigonometric series are well suited to approximate periodic functions (recall our example before of the square wave function). Trigonometric series are, therefore, quite popular in the natural sciences and engineering, where periodic problems are common. Furthermore, they are easy to manipulate as we have plenty of results involving the transformation of trigonometric functions and we can bring to the table the powerful tools of Fourier analysis. Sadly, economic problems are rarely periodic (except in the frequency analysis of time series) and periodic approximations to nonperiodic functions are highly inefficient.

5.3.1.3 Orthogonal Polynomials of the Jacobi Type

We motivated before the need to use a basis of orthogonal functions. Orthogonal polynomials of the Jacobi (also known as hypergeometric) type are a flexible class of polynomials well suited for our needs.

The Jacobi polynomial of degree n, $P_n^{\alpha,\beta}(x)$ for $\alpha, \beta > -1$, is defined by the orthogonality condition:

$$\int_{-1}^{1} (1-x)^{\alpha}(1+x)^{\beta} P_n^{\alpha,\beta}(x) P_m^{\alpha,\beta}(x) dx = 0 \text{ for } m \neq n$$

One advantage of this class of polynomials is that we have a large number of alternative expressions for them. The orthogonality condition implies, with the normalizations:

$$P_n^{\alpha,\beta}(1) = \binom{n+\alpha}{n},$$

that the general n term is given by:

$$2^{-n} \sum_{k=0}^{n} \binom{n+\alpha}{k} \binom{n+\beta}{n-k} (x-1)^{n-k}(x+1)^{k}$$

Recursively:

$$\begin{gathered} 2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)P_{n+1} = \\ \left(\begin{gathered} (2n+\alpha+\beta+1)\left(\alpha^2-\beta^2\right) \\ +(2n+\alpha+\beta)(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)x \end{gathered} \right) P_n \\ -2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)P_{n-1} \end{gathered}$$

Two important cases of Jacobi polynomials are the Legendre polynomials, where $\alpha = \beta = -\frac{1}{2}$, and the Chebyshev polynomials, where $\alpha = \beta = 0$. There is a generalization of Legendre and Chebyshev polynomials, still within the Jacobi family, known as the Gegenbauer polynomials, which set $\alpha = \beta = \upsilon - \frac{1}{2}$ for a parameter υ.

Boyd and Petschek (2014) compare the performance of Gegenbauer, Legendre, and Chebyshev polynomials. Their table 1 is particularly informative. We read it as suggesting that, except for some exceptions that we find of less relevance in the solution of DSGE models, Chebyshev polynomials are the most convenient of the three classes of polynomials. Thus, from now on, we focus on Chebyshev polynomials.

5.3.1.4 Chebyshev Polynomials

Chebyshev polynomials are one of the most common tools of applied mathematics. See, for example, Boyd (2000) and Fornberg (1996) for references and background material. The popularity of Chebyshev polynomials is easily explained if we consider some of their advantages.

First, numerous simple closed-form expressions for the Chebyshev polynomials are available. Thus, the researcher can easily move from one representation to another according to her convenience. Second, the change between the coefficients of a Chebyshev expansion of a function and the values of the function at the Chebyshev nodes is quickly performed by the cosine transform. Third, Chebyshev polynomials are more robust than their alternatives for interpolation. Fourth, Chebyshev polynomials are smooth and bounded between $[-1,1]$. Finally, several theorems bound the errors for Chebyshev polynomials' interpolations.

The most common definition of the Chebyshev polynomials is recursive, with $T_0(x)=1$, $T_1(x)=x$, and the general $n+1$-th order polynomial given by:

$$T_{n+1}(x)=2xT_n(x)-T_{n-1}(x)$$

Applying this recursive definition, the first few polynomials are 1, x, $2x^2-1$, $4x^3-3x$, $8x^4-8x^2+1$,... Thus, the approximation of a function with Chebyshev polynomials is not different from an approximation with monomials (and, thus, we can rely on appropriate versions of the Stone–Weierstrass theorem), except that the orthogonality properties of how Chebyshev polynomials group the monomials make the approximation better conditioned.

Fig. 6 plots the Chebyshev polynomials of order 0–5. The first two polynomials coincide with the first two monomials, a constant and the 45-degree line. The Chebyshev polynomial of order two is a parabola. Higher-order Chebyshev polynomials accumulate several waves. Fig. 6 shows that the Chebyshev polynomials of order n has n zeros, given by

$$x_k=\cos\left(\frac{2k-1}{2n}\pi\right), k=1,\ldots,n.$$

This property will be useful when we describe collocation in a few pages. Also, these zeros are quadratically clustered toward ±1.

Other explicit and equivalent definitions for the Chebyshev polynomials include

$$\begin{aligned}
T_n(x) &= \cos(n \arccos x)\\
&= \frac{1}{2}\left(z^n+\frac{1}{z^n}\right) \text{ where } \frac{1}{2}\left(z+\frac{1}{z}\right)=x\\
&= \frac{1}{2}\left(\left(x+\left(x^2-1\right)^{0.5}\right)^n+\left(x-\left(x^2-1\right)^{0.5}\right)^n\right)\\
&= \frac{1}{2}\sum_{k=0}^{[n/2]}(-1)^k\frac{(n-k-1)!}{k!(n-2k)!}(2x)^{n-2k}\\
&= \frac{(-1)^n\pi^{0.5}}{2^n\Gamma\left(n+\frac{1}{2}\right)}\left(1-x^2\right)^{0.5}\frac{d^n}{dx^n}\left(\left(1-x^2\right)^{n-\frac{1}{2}}\right).
\end{aligned}$$

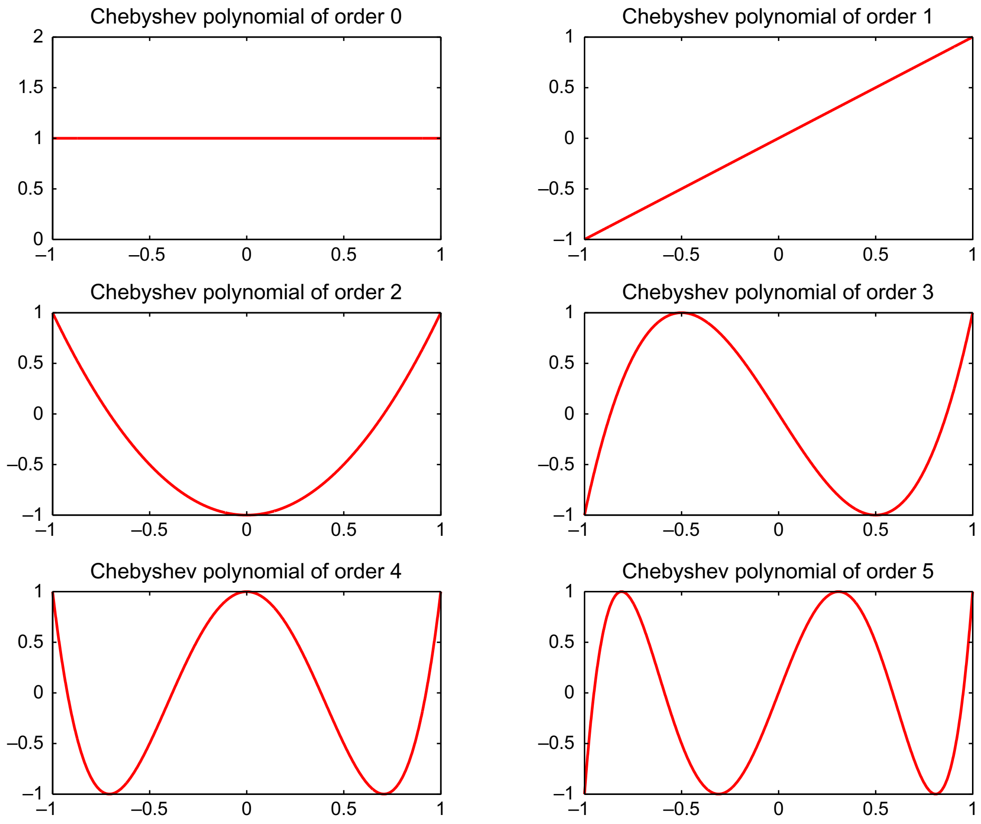


Fig. 6 First six Chebyshev polynomials.

Perhaps the most interesting of these definitions is the first one, since it tells us that Chebyshev polynomials are a trigonometric series in disguise (Boyd, 2000).

A few additional facts about Chebyshev polynomials deserve to be highlighted. First, the $n + 1$ extrema of the polynomial $T_n(x_k)$ $(n > 0)$ are given by:

$$x_k = cos\left(\frac{k}{n}\pi\right), \quad k = 0, \ldots, n. \tag{34}$$

All these extrema are either -1 or 1. Furthermore, two of the extrema are at the endpoints of the domain: $T_n(-1) = (-1)^n$ and $T_n(1) = 1$. Second, the domain of the Chebyshev polynomials is $[-1, 1]$. Since the domain of a state variable x in a DSGE model would be, in general, different from $[-1, 1]$, we can use a linear translation from $[a, b]$ into $[-1, 1]$:

$$2\frac{x-a}{b-a} - 1.$$

Third, the Chebyshev polynomials are orthogonal with respect to the weight function:

$$w(x) = \frac{1}{(1-x^2)^{0.5}}.$$

We conclude the presentation of Chebyshev polynomials with two remarkable results, which we will use below. The first result, due to Erdös and Turán (1937),[s] tells us that if an approximating function is exact at the roots of the n_1^{th} order Chebyshev polynomial, then, as $n_1 \to \infty$, the approximation error becomes arbitrarily small. The *Chebyshev interpolation theorem* will motivate, in a few pages, the use of orthogonal collocation where we pick as collocation points the zeros of a Chebyshev polynomial (there are also related, less used, results when the extrema of the polynomials are chosen instead of the zeros).

Theorem 2 (Chebyshev interpolation theorem) *If $d(x) \in \mathcal{C}[a,b]$, if $\{\phi_i(x), i=0,\ldots\}$ is a system of polynomials (where $\phi_i(x)$ is of exact degree* i*) orthogonal to with respect to $w(x)$ on $[a,b]$ and if $p_j = \sum_{i=0}^{j} \theta_i \phi_i(x)$ interpolates $f(x)$ in the zeros of $\phi_{n+1}(x)$, then:*

$$\lim_{j\to\infty} \left(\left\| d - p_j \right\|_2 \right)^2 = \lim_{n\to\infty} \int_a^b w(x) \left(d(x) - p_j \right)^2 dx = 0$$

We stated a version of the theorem that shows $\mathcal{L}_2$ convergence (a natural norm in economics), but the result holds for $\mathcal{L}_p$ convergence for any $p > 1$. Even if we called this result the *Chebyshev interpolation theorem*, its statement is more general, as it will apply to other polynomials that satisfy an orthogonality condition. The reason we used *Chebyshev* in the theorem's name is that the results are even stronger if the function $d(x)$ satisfies a Dini–Lipschitz condition and the polynomials $\phi_i(x)$ are Chebyshev to uniform convergence, a much more reassuring finding.[t]

But the previous result requires that $j \to \infty$, which is impossible in real applications. The second result will give a sense of how big is the error we are accepting by truncating the approximation of $d(\cdot)$ after a finite (and often relatively low) j.

[s] We reproduce the statement of the theorem, with only minor notational changes, from Mason and Handscomb (2003), chapter 3, where the interested reader can find related results and all the relevant details. This class of theorems is usually derived in the context of interpolating functions.

[t] A function f satisfies a Dini–Lipschitz condition if

$$\lim_{\delta\to 0^+} \omega(\delta) \log \delta = 0$$

where $\omega(\delta)$ is a modulus of continuity of f with respect to δ such that:

$$|f(x+\delta) - f(x)| \leq \omega(\delta).$$

Theorem 3 (Chebyshev truncation theorem, Boyd (Boyd (2000), p. 47)) *The error in approximating* d *is bounded by the sum of the absolute values of all the neglected coefficients. In other words, if we have*

$$d^j(\cdot|\theta) = \sum_{i=0}^{j} \theta_i \psi_i(\cdot)$$

then

$$\left|d(x) - d^j(x|\theta)\right| \leq \sum_{i=j+1}^{\infty} |\theta_i|$$

for any $x \in [-1, 1]$ *and any* j.

We can make the last result even stronger. Under certain technical conditions, we will have a geometric convergence of the Chebyshev approximation to the exact unknown function.[u] And when we have geometric convergence,

$$\left|d(x) - d^j(x|\theta)\right| \sim O(\theta_j)$$

that is, the truncation error created by stopping at the polynomial j is of the same order of magnitude as the coefficient θ_j of the last polynomial. This result also provides us with a simple numerical test: we can check the coefficient θ_j from our approximation: if θ_j is not close enough to zero, we probably need to increase j. We will revisit the evaluation of the accuracy of an approximation in Section 7.

Remark 19 (Change of variables) We mentioned above that, since a state variable x_t in a DSGE model would have, in general, a domain different from $[-1, 1]$, we can use a linear translation from $[a, b]$ into $[-1, 1]$:

$$2\frac{x_t - a}{b - a} - 1.$$

This transformation points to a more general idea: the change of variables as a way to improve the accuracy of an approximation (see also Section 4.5 for the application of the same idea in perturbation). Imagine that we are solving the stochastic neoclassical growth model. Instead of searching for

$$c_t = d^1(k_t, z_t)$$

and

$$k_{t+1} = d^2(k_t, z_t),$$

[u] Convergence of the coefficients is geometric if $\lim_{j\to\infty} \log(|\theta_j|)/j = \text{constant}$. If the lim is infinity, convergence is supergeometric; if the lim is zero, convergence is subgeometric.

we could, instead, search for

$$\log c_t = d^1(\log k_t, z_t)$$

and

$$\log k_{t+1} = d^2(\log k_t, z_t),$$

by defining

$$\log c_t = d^{1,j}\left(\log k_t, z_t|\theta^1\right) = \sum_{i=0}^{j} \theta_i^1 \psi_i(\log k_t, z_t)$$

and

$$\log k_{t+1} = d^{2,j}\left(\log k_t, z_t|\theta^2\right) = \sum_{i=0}^{j} \theta_i^2 \psi_i(\log k_t, z_t).$$

In fact, even in the basic projection example above, we already have a taste of this idea, as we used z_t as a state variable, despite the fact that it appears in the production function as e^{z_t}. An alternative yet equivalent reparameterization writes $A_t = e^{z_t}$ and $z_t = \log A_t$. The researcher can use her a priori knowledge of the model (or preliminary computational results) to search for an appropriate change of variables in her problem. We have changed both state and control variables, but nothing forced us to do so: we could have just changed one variable but not the other or employed different changes of variables.

Remark 20 (Boyd's moral principle)

All of the conveniences of Chebyshev polynomials we just presented are not just theoretical. Decades of real-life applications have repeatedly shown how well Chebyshev polynomials work in a wide variety of applications. In the case of DSGE models, the outstanding performance of Chebyshev polynomial has been shown by Aruoba et al. (2006) and Caldara et al. (2012). Boyd (2000, p. 10), only half-jokingly, has summarized these decades of experience in what he has named his Moral Principle 1:

1. When in doubt, use Chebyshev polynomials unless the solution is spatially periodic, in which case an ordinary Fourier series is better.
2. Unless you are sure another set of basis functions is better, use Chebyshev polynomials.
3. Unless you are really, really sure another set of basis functions is better, use Chebyshev polynomials.

5.3.2 Multidimensional Bases

All of the previous discussion presented unidimensional basis functions. This was useful to introduce the topic. However, most problems in economics are multidimensional: nearly all DSGE models involve several state variables. How do we generalize our basis functions?

The answer to this question is surprisingly important. Projection methods suffer from an acute curse of dimensionality. While solving DSGE models with one or two state variables and projection methods is relatively straightforward, solving DSGE models with 20 state variables and projection methods is a challenging task due to the curse of dimensionality. The key to tackling this class of problems is to intelligently select the multidimensional basis.

5.3.2.1 Discrete State Variables

The idea that the state variables are continuous was implicit in our previous discussion. However, there are many DSGE models where either some state variable is discrete (ie, the government can be in default or not, as in Bocola (2015), or monetary policy can be either active or passive in the sense of Leeper, 1991) or where we can discretize one continuous state variable without losing much accuracy. The best example of the latter is the discretization of exogenous stochastic processes for productivity or preference shocks. Such discretization can be done with the procedures proposed by Tauchen (1986) or Kopecky and Suen (2010), who find a finite state Markov chain that generates the same population moments as the continuous process. Experience suggests that, in most applications, a Markov chain with 5 or 7 states suffices to capture nearly all the implications of the stochastic process for quantitative analysis.

A problem with discrete state variables can be thought of as one where we search for a different decision rule for each value of that state variable. For instance, in the stochastic neoclassical growth model with state variables k_t and z_t, we can discretize the productivity level z_t into a Markov chain with n points

$$z_t \in \{z_1, .., z_n\}$$

and transition matrix:

$$P_{z,z'} = \left\{ \begin{array}{ccc} p_{11} & \cdots & p_{1n} \\ \vdots\ddots & \vdots & \\ p_{n1} & \cdots & p_{nn} \end{array} \right\} \tag{35}$$

where entry p_{ij} is the probability that the chain will move from position i in the current period to position j in the next period.

Remark 21 (Discretization methods) Tauchen (1986) procedure to discretize an AR(1) stochastic process

$$z_t = \rho z_{t-1} + \epsilon_t$$

with stationary distribution $N(0, \sigma_z^2)$, where $\sigma_z = \frac{\sigma_\epsilon}{\sqrt{1-\rho^2}}$, works as follows:

Algorithm 2 (AR(1) Discretization)

1. Set n, the number of potential realizations of the process z.
2. Set the upper ($\bar{z}$) and lower ($\underline{z}$) bounds for the process. An intuitive way to set the bounds is to pick m such that:

$$\bar{z} = m\sigma_z$$
$$\underline{z} = -m\sigma_z$$

   The latter alternative is appealing given the symmetry of the normal distribution around 0. Usual values of m are between 2 and 3.
3. Set $\{z_i\}_{i=1}^{n}$ such that:

$$z_i = \underline{z} + \frac{\bar{z} - \underline{z}}{n-1}(i-1)$$

   and construct the midpoints $\{\tilde{z}_i\}_{i=1}^{n-1}$, which are given by:

$$\tilde{z}_i = \frac{z_{i+1} + z_i}{2}$$

4. The transition probability $p_{ij} \in P_{z,z'}$ (the probability of going to state z_j conditional on being on state z_i), is computed according to:

$$p_{ij} = \Phi\left(\frac{\tilde{z}_j - \rho z_i}{\sigma}\right) - \Phi\left(\frac{\tilde{z}_{j-1} - \rho z_i}{\sigma}\right) \qquad j = 2,3,\ldots,n-1$$

$$p_{i1} = \Phi\left(\frac{\tilde{z}_1 - \rho z_i}{\sigma}\right)$$

$$p_{in} = 1 - \Phi\left(\frac{\tilde{z}_{n-1} - \rho z_i}{\sigma}\right)$$

   where $\Phi(\cdot)$ denotes a CDF of a $N(0,1)$.

To illustrate Tauchen's procedure, let us assume we have a stochastic process:

$$z_t = 0.95 z_{t-1} + \epsilon_t$$

with $N(0,0.007^2)$ (this is a standard quarterly calibration for the productivity process for the US economy; using data after 1984 the standard deviation is around 0.0035) and we want to approximate it with a 5-point Markov chain and $m = 3$. Tauchen's procedure gives us:

$$z_t \in \{-0.0673, -0.03360, 0.0336, 0.0673\} \tag{36}$$

and transition matrix:

$$P_{z,z'} = \left\{ \begin{array}{ccccc} 0.9727 & 0.0273 & 0 & 0 & 0 \\ 0.0041 & 0.9806 & 0.0153 & 0 & 0 \\ 0 & 0.0082 & 0.9837 & 0.0082 & 0 \\ 0 & 0 & 0.0153 & 0.9806 & 0.0041 \\ 0 & 0 & 0 & 0.0273 & 0.9727 \end{array} \right\} \tag{37}$$

Note how the entries in the diagonal are close to 1 (the persistence of the continuous stochastic process is high) and that the probability of moving two or more positions is zero. It would take at least 4 quarters for the Markov chain to travel from z_1 to z_5 (and vice versa).

Tauchen's procedure can be extended to VAR processes instead of an AR process. This is convenient because we can always rewrite a general ARMA(p,q) process as a VAR(1) (and a VAR(p) as a VAR(1)) by changing the definition of the state variables. Furthermore, open source implementations of the procedure exist for all major programming languages.

Kopecky and Suen (2010) show that an alternative procedure proposed by Rouwenhorst (1995) is superior to Tauchen's method when ρ, the persistence of the stochastic process, is close to 1. The steps of Rouwenhorst (1995)'s procedure are:

Algorithm 3 (Alternative AR(1) Discretization)

1. Set n, the number of potential realizations of the process z.

2. Set the upper ($\bar{z}$) and lower ($\underline{z}$) bounds for the process. Let $\underline{z} = -\lambda$ and $\bar{z} = \lambda$. λ can be set to be $\lambda = \sqrt{n-1}\sigma_z$.

3. Set $\{z_i\}_{i=1}^{n}$ such that:

$$z_i = \underline{z} + \frac{\bar{z} - \underline{z}}{n-1}(i-1)$$

4. When $n = 2$, let P_2 be given by:

$$P_2 = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix}$$

p, q can be set to be $p = q = \dfrac{1+\rho}{2}$.

5. For $n \geq 3$, construct recursively the transition matrix:

$$P_n = p\begin{bmatrix} P_{n-1} & \mathbf{0} \\ \mathbf{0}' & 0 \end{bmatrix} + (1-p)\begin{bmatrix} \mathbf{0} & P_{n-1} \\ 0 & \mathbf{0}' \end{bmatrix} + (1-q)\begin{bmatrix} \mathbf{0}' & 0 \\ P_{n-1} & \mathbf{0} \end{bmatrix} + q\begin{bmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & P_{n-1} \end{bmatrix}$$

where $\mathbf{0}$ is an $(n-1) \times 1$ column vector of zeros. Divide all but the top and bottom rows by 2 so that the sum of the elements of each row is equal to 1. The final outcome is $P_{z,z'}$.

Once productivity has been discretized, we can search for

$$\begin{aligned} c(k,z_m) &= d^{c,m,j}(k|\theta^{m,c}) = \sum_{i=0}^{j} \theta_i^{m,c}\psi_i(k) \\ k(k,z_m) &= d^{k,m,j}\left(k|\theta^{m,k}\right) = \sum_{i=0}^{j} \theta_i^{m,k}\psi_i(k) \end{aligned}$$

where $m = 1,\ldots,n$. That is, we search for decision rules for capital and consumption when productivity is z_1 today, decision rules for capital and consumption when productivity is z_2 today, and so on, for a total of $2 \times n$ decision rules. Since n is usually a small number (we mentioned above 5 or 7), the complexity of the problem is not exploding.

Note that since we substitute these decision rules in the Euler equation:

$$u'(c_t) = \beta\mathbb{E}_t\left[u'(c_{t+1})\left(\alpha e^{z_{t+1}}k_{t+1}^{\alpha-1} + 1 - \delta\right)\right]. \tag{38}$$

to get:

$$u'\left(d^{c,m,j}(k|\theta^{m,c})\right) = \\ \beta\sum_{l=0}^{n} p_{ml}\left[u'\left(d^{c,l,j}\left(d^{k,m,j}\left(k|\theta^{m,k}\right)\big|\theta^{l,c}\right)\right)\left(\alpha e^{z_{t+1}}\left(d^{k,m,j}\left(k|\theta^{m,k}\right)\right)_{t+1}^{\alpha-1} + 1 - \delta\right)\right]$$

we are still taking account of the fact that productivity can change in the next period (and hence, consumption and capital accumulation will be determined by the decision rule for the next period level of productivity). Also, since now the stochastic process is discrete, we can substitute the integral on the right-hand side of Eq. (38) for the much simpler sum operator with the probabilities from the transition matrix (35). Otherwise, we would need to use a quadrature method to evaluate the integral (see Judd, 1998 for the relevant formulae and the proposal in Judd et al., 2011a).

Thus, discretization of state variables such as the productivity shock is more often than not an excellent strategy to deal with multidimensional problems: simple, transparent, and not too burdensome computationally. Furthermore, we can discretize some of the state variables and apply the methods in the next paragraphs to deal with the remaining continuous state variables. In computation, mixing of strategies is often welcomed.

5.3.2.2 Tensors

Tensors build multidimensional basis functions by finding the Kronecker product of all unidimensional basis functions.[v] Imagine, for example, that we have two state variables,

[v] One should not confuse the tensors presented here with the tensor notation used for perturbation methods. While both situations deal with closely related mathematical objects, the key when we were dealing with perturbation was the convenience that tensor notation offered.

physical capital k_t and human capital h_t. We have three Chebyshev polynomials for each of these two state variables:

$$\psi_0^k(k_t), \psi_1^k(k_t), \text{ and } \psi_2^k(k_t)$$

and

$$\psi_0^h(h_t), \psi_1^h(h_t), \text{ and } \psi_2^h(h_t).$$

Then, the tensor is given by:

$$\begin{aligned}&\psi_0^k(k_t)\psi_0^h(h_t), \psi_0^k(k_t)\psi_1^h(h_t), \psi_0^k(k_t)\psi_2^h(h_t),\\&\psi_1^k(k_t)\psi_0^h(h_t), \psi_1^k(k_t)\psi_1^h(h_t), \psi_1^k(k_t)\psi_2^h(h_t),\\&\psi_2^k(k_t)\psi_2^h(h_t), \psi_2^k(k_t)\psi_1^h(h_t), \text{ and } \psi_2^k(k_t)\psi_2^h(h_t).\end{aligned}$$

More formally, imagine that we want to approximate a function of n state variables $d:[-1,1]^n \rightarrow \mathbb{R}$ with Chebyshev polynomial of degree j. We build the sum:

$$d^j(\cdot|\theta) = \sum_{i_1=0}^{j} \cdots \sum_{i_n=0}^{j} \theta_{i_1,\ldots,i_n} \psi_{i_1}^1(\cdot) * \ldots * \psi_{i_n}^n(\cdot)$$

where $\psi_{i_\kappa}^\kappa$ is the Chebyshev polynomials of degree i_κ on the state variable κ and θ is the vector of coefficients $\theta_{i_1,\ldots,i_n}$. To make the presentation concise, we have made three simplifying assumptions. First, we are dealing with the case that d is one dimensional. Second, we are using the same number of Chebyshev polynomials for each state variable. Three, the functions $\psi_{i_\kappa}^\kappa$ could be different from the Chebyshev polynomials and belong to any basis we want (there can even be a different basis for each state variable). Eliminating these simplifications is straightforward, but notationally cumbersome.

There are two main advantages of a tensor basis. First, it is trivial to build. Second, if the one-dimensional basis is orthogonal, then the tensor basis is orthogonal in the product norm. The main disadvantage is the exponential growth in the number of coefficients $\theta_{i_1,\ldots,i_n}$: $(j+1)^n$. In the example above, even using only three Chebyshev polynomials (ie, $j = 2$) for each of these two state variables, we end up having to solve for nine coefficients. This curse of dimensionality is acute: with five state variables and three Chebyshev polynomials, we end up with 243 coefficients. With ten Chebyshev polynomials, we end up with 100,000 coefficients.

5.3.2.3 Complete Polynomials

In practice, it is infeasible to use tensors when we are dealing with models with more than 3 continuous state variables and a moderate j. A solution is to eliminate some elements of the tensor in a way that avoids much numerical degradation. In particular, Gaspar and Judd (1997) propose using the complete polynomials:

$$\mathcal{P}_\kappa^n \equiv \left\{\psi_{i_1}^1 * \ldots * \psi_{i_n}^n \text{ with } |\mathbf{i}| \leq \kappa\right\}$$

where

$$|\mathbf{i}| = \sum_{l=1}^{n} i_l, 0 \leq i_1, \ldots, i_n.$$

Complete polynomials, instead of employing all the elements of the tensor, keep only those such that the sum of the order of the basis functions is less than a prefixed κ. The intuition is that the elements of the tensor $\psi_{i_1}^{1} * \ldots * \psi_{i_n}^{n}, |\mathbf{i}| > \kappa$ add little additional information to the basis: most of the flexibility required to capture the behavior of d is already in the complete polynomials. For instance, if we are dealing with three state variables and Chebyshev polynomials $j = 4$, we can keep the complete polynomials of order 6:

$$\mathcal{P}_6^3 \equiv \left\{ \psi_{i_1}^{1} * \ldots * \psi_{i_n}^{n} \text{ with } |\mathbf{i}| \leq 6 \right\}.$$

Complete polynomials eliminate many coefficients: in our example, instead of $(4 + 1)^3 = 125$ coefficients of the tensor, when $\kappa = 6$ we only need to approximate 87 coefficients. Unfortunately, we still need too many coefficients. In Section 5.7, we will present an alternative: Smolyak's algorithm. However, since the method requires the introduction of a fair amount of new notation and the presentation of the notion of interpolating polynomials, we postpone the discussion and, instead, start analyzing the finite element methods.

5.4 Finite Elements

Finite elements techniques, based on local basis functions, were popularized in economics by McGrattan (1996) (see, also, Hughes (2000), for more background, and Brenner and Scott (2008), for all the mathematical details that we are forced to skip in a handbook chapter). The main advantage of this class of basis functions is they can easily capture local behavior and achieve a tremendous level of accuracy even in the most challenging problems. That is why finite element methods are often used in mission-critical design in industry, such as in aerospace or nuclear power plant engineering. The main disadvantage of finite elements methods is that they are hard to code and expensive to compute. Therefore, we should choose this strategy when accuracy is more important than speed of computation or when we are dealing with complicated, irregular problems.

Finite elements start by bounding the domain $\mathbf{\Omega}$ of the state variables. Some of the bounds would be natural (ie, $k_t > 0$). Other bounds are not ($k_t < \bar{k}$) and we need some care in picking them. For example, we can guess a $\bar{k}$ sufficiently large such that, in the simulations of the model, k_t never reaches $\bar{k}$. This needs, however, to be verified and some iterative fine-tuning may be required.[w]

[w] Even if the simulation rarely reaches $\bar{k}$, it may be useful to repeat the computation with a slightly higher bound $\omega\bar{k}$, with $\omega > 1$, to check that we still do not get to $\bar{k}$. In some rare cases, the first simulation might not have reached $\bar{k}$ because the approximation of the function $d(\cdot)$ precluded traveling into that region.

The second step in the finite elements method is to partition Ω into small, nonintersecting elements. These small sections are called elements (hence the name, "finite elements"). The boundaries of the elements are called nodes. The researcher enjoys a fantastic laxity in selecting the partition. One natural partition is to divide Ω into equal elements: simple and direct. But elements can be of unequal size. More concretely, we can have small elements in the areas of Ω where the economy will spend most of the time, while just a few large elements will cover areas of Ω infrequently visited (these areas can be guessed based on the theoretical properties of the model, or they can be verified by an iterative procedure of element partition; we will come back to this point below). Or we can have small elements in the areas of Ω where the function $d(\cdot)$ we are looking for changes quickly in shape, while we reserve large elements for areas of Ω where the function d is close to linear. Thanks to this flexibility in the element partition, we can handle kinks or constraints, which are harder to tackle with spectral methods (or next to impossible to do with perturbation, as they violate differentiability conditions).[x]

An illustration of such capability appears in Fig. 7, where we plot the domain Ω of a dynamic model of a firm with two state variables, bonds b_t on the x-axis (values to the right denote positive bond holdings by the firm and values to the left negative bond holdings), and capital k_t on the y-axis. The domain Ω does not include an area in the lower left corner, of combinations of negative bond holdings (ie, debt) and low capital. This area is excluded because of a financial constraint: firms cannot take large amounts of debt when

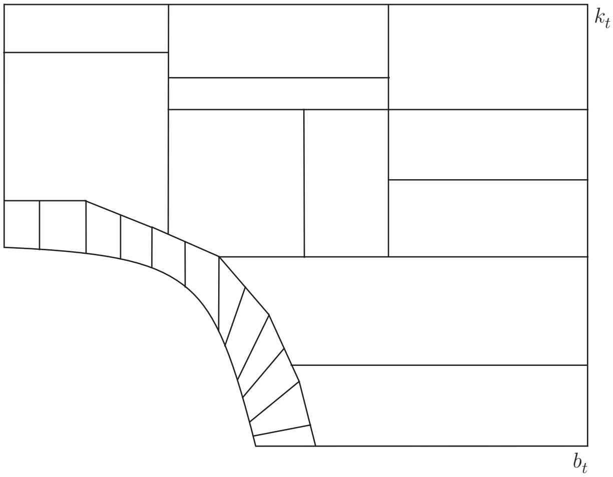


Fig. 7 Two-dimensional element grid.

[x] This flexibility in the definition of the elements is a main reason why finite elements methods are appreciated in industry, where applications often do not conform to the regularity technical conditions required by perturbation or spectral techniques.

they do not have enough capital to use as collateral (the concrete details of this financial constraint or why the shape of the restricted area is the one we draw are immaterial for the argument). In Fig. 7, the researcher has divided the domain Ω into unequal elements: there are many of them, of small size, close to the lower left corner boundary. One can suspect that the decision rule for the firm for b_t and k_t may change rapidly close to the frontier or, simply, the researcher wants to ensure the accuracy of the solution in that area. Farther away from the frontier, elements become larger. But even in those other regions, the researcher can partition the domain Ω with very different elements, some smaller (high levels of debt and k_t), some larger (high levels of b_t and k_t), depending on what the researcher knows about the shape of the decision rule.

There is a whole area of research concentrated on the optimal generation of an element grid that we do not have space to review. The interested reader can check Thompson et al. (1985). For a concrete application of unequal finite elements to the stochastic neoclassical growth model to reduce computational time, see Fernández-Villaverde and Rubio-Ramírez (2004).

The third step in the finite elements method is to choose a basis for the policy functions in each element. Since the elements of the partition of Ω are usually small, a linear basis is often good enough. For instance, letting $\{k_0, k_1, \ldots, k_j\}$ be the nodes of a partition of Ω into elements, we can define the tent functions for $i \in \{1, j-1\}$

$$\psi_i(k) = \begin{cases} \dfrac{k - k_{i-1}}{k_i - k_{i-1}}, \text{ if } x \in [k_{i-1}, k_i] \\ \dfrac{k_{i+1} - k}{k_{i+1} - k_i}, \text{ if } k \in [k_i, k_{i+1}] \\ 0 \quad \text{elsewhere} \end{cases}$$

and the corresponding adjustments for the first function:

$$\psi_0(k) = \begin{cases} \dfrac{k_0 - k}{k_1 - k_0}, \text{ if } x \in [k_0, k_1] \\ 0 \quad \text{elsewhere} \end{cases}$$

and the last one

$$\psi_j(k) = \begin{cases} \dfrac{k - k_{j-1}}{k_j - k_{j-1}}, \quad \text{if } k \in [k_i, k_{i+1}] \\ 0 \quad \text{elsewhere.} \end{cases}$$

We plot examples of these tent functions in Fig. 8.

We can extend this basis to higher dimensions by either discretizing some of the state variables (as we did when we talked about spectral bases) or by building tensors of them. Below, we will also see how to use Smolyak's algorithm with finite elements.

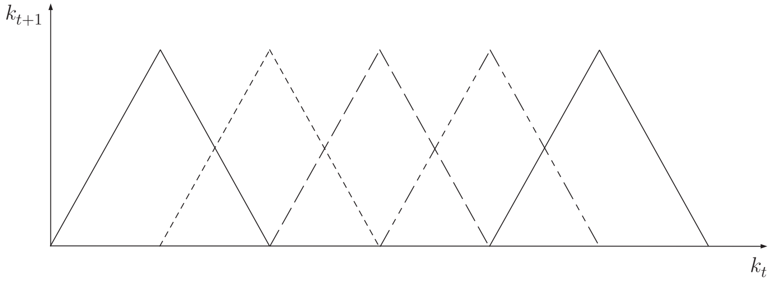


Fig. 8 Five basis functions.

The fourth step in the finite elements method is the same as for any other projection method: we build

$$d^{n,j}(\cdot|\theta^n) = \sum_{i=0}^{j} \theta_i^n \psi_i(\cdot)$$

and we plug them into the operator $\mathcal{H}$. Then, we find the unknown coefficients as we would do with Chebyshev polynomials.

By construction, the different parts of the approximating function will be pasted together to ensure continuity. For example, in our Fig. 8, there are two basis functions in the element defined by the nodes k_i and k_{i+1}

$$\psi_i(k) = \frac{k_{i+1} - k}{k_{i+1} - k_i}$$

$$\psi_{i+1}(k) = \frac{k - k_i}{k_{i+1} - k_i}$$

and their linear combination (ie, the value of $d^{n,j}(\cdot|\theta^n)$ in that element) is:

$$\hat{d}\left(k|k_{i+1}, k_i, \theta_{i+1}^n, \theta_i^n\right) = \theta_i^n \frac{k_{i+1} - k}{k_{i+1} - k_i} + \theta_{i+1}^n \frac{k - k_i}{k_{i+1} - k_i} = \frac{\left(\theta_{i+1}^n - \theta_i^n\right)k + \theta_i^n k_{i+1} - \theta_{i+1}^n k_i}{k_{i+1} - k_i},$$

which is a linear function, with positive or negative slope depending on the sign of $\theta_{i+1}^n - \theta_i^n$. Also note that the value of $d^{n,j}(\cdot|\theta^n)$ in the previous element is the linear function:

$$\hat{d}\left(k|k_i, k_{i-1}, \theta_i^n, \theta_{i-1}^n\right) = \frac{\left(\theta_i^n - \theta_{i-1}^n\right)k + \theta_{i-1}^n k_i - \theta_i^n k_{i-1}}{k_i - k_{i-1}}.$$

When we evaluate both linear functions at k_i

$$\hat{d}\left(k_i|k_i, k_{i-1}, \theta_i^n, \theta_{i-1}^n\right) = \theta_i^n$$

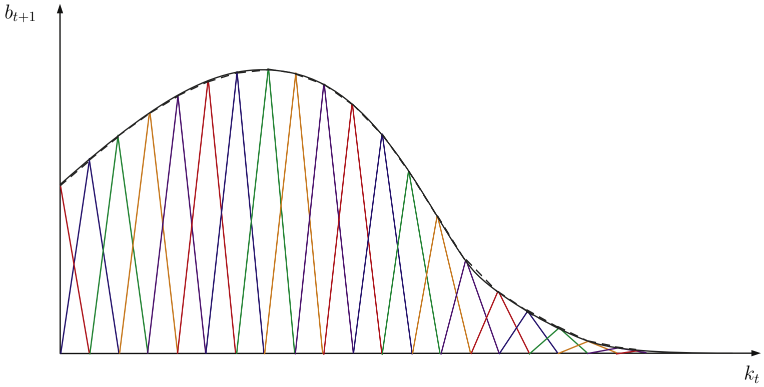


Fig. 9 Finite element approximation.

and

$$\hat{d}\left(k_i|k_{i+1},k_i,\theta^n_{i+1},\theta^n_i\right)=\theta^n_i$$

that is, both functions have the same value equal to the coefficient θ^n_i, which ensures continuity (although, with only tent functions, we cannot deliver differentiability).

The previous derivation also shows why finite elements are a smart strategy. Imagine that our metric ρ is such that we want to make the residual function equal to zero in the nodes of the elements (below we will present a metric like this one). With our tent functions, this amounts to picking, at each k_i, the coefficient θ^n_i such that the approximating and exact function coincide:

$$d^{n,j}(\cdot|\theta^n)=d^n(\cdot).$$

This implies that the value of d^n outside k_i are irrelevant for our choice of θ^n_i. An example of such piecewise linear approximation to a decision rule for the level of debt tomorrow, b_{t+1}, given capital today, k_t, in a model of financial frictions, is drawn in Fig. 9. The discontinuous line is the approximated decision rule and the continuous line the exact one. The tent functions are multiplied by the coefficients to make the approximation and the exact solution equal at the node points. We can appreciate an already high level of accuracy. As the elements become smaller and smaller, the approximation will become even more accurate (ie, smooth functions are locally linear).

This is a stark example of a more general point: the large system of nonlinear equations that we will need to solve in a finite element method will be sparse, a property that can be suitably exploited by modern nonlinear solvers.

Remark 22 (Finite elements method refinements) An advantage of the finite elements method is that we can refine the solution that we obtain as much as we desire (with only the constraints of computational time and memory). The literature distinguishes among three different refinements. First, we have the *h-refinement*. This scheme subdivides each element into smaller elements to improve resolution uniformly over the domain. That is, once we have obtained a first solution, we check whether this solution achieves the desired level of accuracy. If it does not, we go back to our partition, and we subdivide the elements. We can iterate in this procedure as often as we need. Second, we have *r-refinement*: This scheme subdivides each element only in those regions where there are high nonlinearities. Third, we have the *p-refinement*: This scheme increases the order of the approximation in each element, that is, it adds more basis functions (for example, several Chebyshev polynomials). If the order of the expansion is high enough, we generate a hybrid of finite and spectral methods known as spectral elements. This approach has gained much popularity in the natural sciences and engineering. See, for example, Solín et al. (2004).

Sometimes, *h-refinements* and *p-refinements* are mixed in what is known as the hp-finite element method, which delivers exponential convergence to the exact solution. Although difficult to code and computationally expensive, an hp-finite element method is, perhaps, the most powerful solution technique available for DSGE models, as it can tackle even the most challenging problems.[y]

The three refinements can be automatically implemented: we can code the finite element algorithm to identify the regions of Ω where, according to some goal of interest (for example, how tightly a Euler equation is satisfied), we refine the approximation without further input from the researcher. See Demkowicz (2007).

5.5 Objective Functions

Our second choice is to select a metric function ρ to determine how we "project." The most common answer to this question is given by a *weighted residual*: we select θ to get the residual close to 0 in the weighted integral sense. Since we did not impose much structure on the operator $\mathcal{H}$ and therefore, on the residual function $R(\cdot|\theta)$, we will deal with the simplest case where $R(\cdot|\theta)$ is unidimensional. More general cases can be dealt with at the cost of heavier notation. Given some weight functions $\phi_i : \Omega \to \mathbb{R}$, we define the metric:

$$\rho(R(\cdot|\theta),0)=\begin{cases}0 & \text{if } \int_\Omega \phi_i(\mathbf{x})R(\cdot|\theta)d\mathbf{x}=0,\, i=1,..,j+1\\ 1 & \text{otherwise}\end{cases}$$

[y] An additional, new refinement is the extended finite element method (x-fem), which adds to the basis discontinuous functions that can help in capturing irregularities in the solution. We are not aware of applications of the x-fem in economics.

Hence, the problem is to choose the θ that solves the system of integral equations:

$$\int_{\Omega} \phi_i(\mathbf{x}) R(\cdot|\theta) d\mathbf{x} = 0, i = 1,..,j+1. \tag{39}$$

Note that, for the system to have a solution, we need $j + 1$ weight functions. Thanks to the combination of approximating the function d by basis functions ψ_i and the definition of weight functions ϕ_i, we have transformed a rather intractable functional equation problem into a standard nonlinear equations system. The solution of this system can be found using standard methods, such as a Newton algorithm for small problems or a Levenberg–Marquardt method for bigger ones.

However, the system (39) may have no solution or it may have multiple ones. We know very little about the theoretical properties of projection methods in economic applications. The literature in applied mathematics was developed for the natural sciences and engineering and many of the technical conditions required for existence and convergence theorems to work do not easily travel across disciplines. In fact, some care must be put into ensuring that the solution of the system (39) satisfies the transversality conditions of the DSGE model (ie, we are picking the stable manifold). This can usually be achieved with the right choice of an initial guess θ_0 or by adding boundary conditions to the solver.

As was the case with the bases, we will have plenty of choices for our weight functions. Instead of reviewing all possible alternatives, we will focus on the most popular ones in economics.

5.5.1 Weight Function I: Least Squares

Least squares use as weight functions the derivatives of the residual function:

$$\phi_i(\mathbf{x}) = \frac{\partial R(\mathbf{x}|\theta)}{\partial \theta_{i-1}}$$

for all $i \in 1,..,j + 1$. This choice is motivated by the variational problem:

$$\min_{\theta} \int_{\Omega} R^2(\cdot|\theta) d\mathbf{x}$$

with first-order condition:

$$\int_{\Omega} \frac{\partial R(\mathbf{x}|\theta)}{\partial \theta_{i-1}} R(\cdot|\theta) d\mathbf{x} = 0, i = 1,..,j+1.$$

This variational problem is mathematically equivalent to a standard regression problem in econometrics.

While least squares are intuitive and there are algorithms that exploit some of their structure to increase speed and decrease memory requirements, they require the computation of the derivative of the residual, which can be costly. Also, least squares problems are often ill-conditioned and complicated to solve numerically.

5.5.2 Weight Function II: Subdomain

The subdomain approach divides the domain $\boldsymbol{\Omega}$ into 1,..,j + 1 subdomains $\boldsymbol{\Omega}_i$ and define the j + 1 step functions:

$$\phi_i(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \boldsymbol{\Omega}_i \\ 0 & \text{otherwise} \end{cases}$$

This choice is equivalent to solving the system:

$$\int_{\Omega_i} R(\cdot|\theta)d\mathbf{x} = 0, i = 1, .., j+1.$$

The researcher has plenty of flexibility to pick her subdomains as to satisfy her criteria of interest.

5.5.3 Weight Function III: Collocation

This method is also known as pseudospectral or the method of selected points. It defines the weight function as:

$$\phi_i(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{x}_i)$$

where δ is the Dirac delta function and $\mathbf{x}_i$ are the j + 1 collocation points selected by the researcher.

This method implies that the residual function is zero at the n collocation points. Thus, instead of having to compute complicated integrals, we only need to solve the system:

$$R(\mathbf{x}_i|\theta) = 0, i = 1, .., j+1.$$

This is attractive when the operator $\mathcal{H}$ generates large nonlinearities.

A systematic way to pick collocation points is to use the zeros of the $(j+1)$-th-order Chebyshev polynomial in each dimension of the state variable (or the corresponding polynomials, if we are using different approximation orders along each dimension). This approach is known as orthogonal collocation. The Chebyshev interpolation theorem tells us that, with this choice of collocation points, we can achieve $\mathcal{L}_p$ convergence and sometimes even uniform convergence to the unknown function d. Another possibility is to pick, as collocation points, the extrema of the j-th-order Chebyshev polynomial in each dimension. Experience shows a surprisingly good performance of orthogonal collocation methods and it is one of our recommended approaches.

5.5.4 Weight Function IV: Galerkin or Rayleigh–Ritz

The last weight function we consider is the Galerkin (also called Rayleigh–Ritz when it satisfies some additional properties of less importance for economists). This approach takes as the weight function the basis functions used in the approximation:

$$\phi_i(\mathbf{x}) = \psi_{i-1}(\mathbf{x}).$$

Then we have:

$$\int_{\Omega} \psi_i(\mathbf{x}) R(\cdot|\theta) d\mathbf{x} = 0, i = 1, .., j+1.$$

The interpretation is that the residual has to be orthogonal to each of the basis functions.

The Galerkin approach is highly accurate and robust, but difficult to code. If the basis functions are complete over J_1 (they are indeed a basis of the space), then the Galerkin solution will converge pointwise to the true solution as n goes to infinity:

$$\lim_{j \to \infty} d^j(\cdot|\theta) = d(\cdot)$$

Also, practical experience suggests that a Galerkin approximation of order j is as accurate as a pseudospectral $j+1$ or $j+2$ expansion.

In the next two remarks, we provide some hints for a faster and more robust solution of the system of nonlinear equations:

$$\int_{\Omega} \phi_i(\mathbf{x}) R(\cdot|\theta) d\mathbf{x} = 0, i = 1, .., j+1, \tag{40}$$

a task that can be difficult if the number of coefficients is large and the researcher does not have a good initial guess θ_0 for the solver.

Remark 23 (Transformations of the problem) A bottleneck for the solution of (39) can be the presence of strong nonlinearities. Fortunately, it is often the case that simple changes in the problem can reduce these nonlinearities. For example, Judd (1992) proposes that if we have an Euler equation:

$$\frac{1}{c_t} = \beta \mathbb{E}_t \left\{ \frac{1}{c_{t+1}} R_{t+1} \right\}$$

where R_{t+1} is the gross return rate of capital, we can take its inverse:

$$\beta c_t = \left(\mathbb{E}_t \left\{ \frac{1}{c_{t+1}} R_{t+1} \right\} \right)^{-1},$$

which now is linear on the left-hand side and much closer to linear on the right-hand side. Thus, instead of computing the residual for some state variable x_t

$$R(\cdot|\theta) = \frac{1}{c(x_t|\theta)} - \beta \mathbb{E}_t \left\{ \frac{1}{c(x_t|\theta)} R_{t+1}(x_t|\theta) \right\},$$

we compute:

$$\widetilde{R}(\cdot|\theta) = \beta c(x_t|\theta) - \left(\mathbb{E}_t \left\{ \frac{1}{c(x_t|\theta)} R_{t+1}(x_t|\theta) \right\} \right)^{-1}.$$

Similar algebraic manipulations are possible in many DSGE models.

Remark 24 (Multistep schemes) The system (39) can involve a large number of coefficients. A natural strategy is to solve first a smaller system and to use that solution as an input for a larger system. This strategy, called a multistep scheme, often delivers excellent results, in particular when dealing with orthogonal bases such as Chebyshev polynomials.

More concretely, instead of solving the system for an approximation with $j+1$ basis functions, we can start by solving the system with only $j'+1 \ll j+1$ basis functions and use the solution to this first problem as a guess for the more complicated problem. For example, if we are searching for a solution with 10 Chebyshev polynomials and m dimensions, we first find the approximation with only 3 Chebyshev polynomials. Therefore, instead of solving a system of $10 \times m$ equations, we solve a system of $3 \times m$. Once we have the solution θ^3, we build the initial guess for the problem with 10 Chebyshev polynomials as:

$$\theta_0 = \left[\theta^3, \mathbf{0}_{1\times m}, \ldots, \mathbf{0}_{1\times m}\right],$$

that is, we use θ^3 for the first coefficients and zero for the additional new coefficients. Since the additional polynomials are orthogonal to the previous ones, the final values of the coefficients associated with the three first polynomials will change little with the addition of 7 more polynomials: the initial guess θ^3 is, thus, most splendid. Also, given the fast convergence of Chebyshev polynomials, the coefficients associated with higher-order polynomials will be close to zero. Therefore, our initial guess for those coefficients is also informative.

The researcher can use as many steps as she needs. By judiciously coding the projection solver, the researcher can write the program as depending on an abstract number of Chebyshev polynomials. Then, she can call the solver inside a loop and iteratively increase the level of approximation from j' to j as slow or as fast as required.

5.6 A Worked-Out Example

We present now a worked-out example of how to implement a projection method in a DSGE model. In particular, we will use Chebyshev polynomials and orthogonal collocation to solve the stochastic neoclassical growth model with endogenous labor supply.

In this economy, there is a representative household, whose preferences over consumption, c_t, and leisure, $1 - l_t$, are representable by the utility function:

$$\mathbb{E}_0 \sum_{t=1}^{\infty} \beta^{t-1} \frac{\left(c_t^{\tau}(1-l_t)^{1-\tau}\right)^{1-\eta}}{1-\eta}$$

where $\beta \in (0,1)$ is the discount factor, η controls the elasticity of intertemporal substitution and risk aversion, τ controls labor supply, and $\mathbb{E}_0$ is the conditional expectation operator.

There is one good in the economy, produced according to the aggregate production function:

$$y_t = e^{z_t} k_t^{\alpha} l_t^{1-\alpha}$$

where k_t is the aggregate capital stock, l_t is aggregate labor, and z_t is a stochastic process for technology:

$$z_t = \rho z_{t-1} + \epsilon_t$$

with $|\rho| < 1$ and $\epsilon_t \sim N(0, \sigma^2)$. Capital evolves according to:

$$k_{t+1} = (1-\delta)k_t + i_t$$

and the economy must satisfy the resource constraint $y_t = c_t + i_t$.

Since both welfare theorems hold in this economy, we solve directly for the social planner's problem:

$$\begin{gathered} V(k_t, z_t) = \max_{c_t, l_t} \frac{\left(c_t^{\tau}(1-l_t)^{1-\tau}\right)^{1-\eta}}{1-\eta} + \beta \mathbb{E}_t V(k_{t+1}, z_{t+1}) \\ \text{s.t. } k_{t+1} = e^{z_t} k_t^{\alpha} l_t^{1-\alpha} + (1-\delta)k_t - c_t \\ z_t = \rho z_{t-1} + \epsilon_t \end{gathered}$$

given some initial conditions k_0 and z_0. Tackling the social planner's problem is only done for convenience, and we could also solve for the competitive equilibrium. In fact, one key advantage of projection methods is that they easily handle non-Pareto efficient economies.

We calibrate the model with standard parameter values to match US quarterly data (see Table 2). The only exception is η, for which we pick a value of 5, in the higher range of empirical estimates. Such high-risk aversion induces, through precautionary behavior, more curvature in the decision rules. This curvature would present a more challenging test bed for the projection method.

We discretize z_t into a 5-point Markov chain $\{z_1, \ldots, z_5\}$ using Tauchen's procedure and covering ±3 unconditional standard deviations of z_t (this is the same Markov chain as

Table 2 Calibration

Parameter	Value
β	0.991
η	5.000
τ	0.357
α	0.300
δ	.0196
ρ	0.950
σ	0.007

the example in Remark 21, see (36) and (37) for the concrete values of the discretization). We will use p_{mn} to denote the generic entry of the transition matrix $P_{z,z'}$ generated by Tauchen's procedure for z_m today moving to z_n next period.

Then, we approximate the value function $V^j(k_t)$ and the decision rule for labor, $l^j(k_t)$, for $j = 1,\ldots,5$ using 11 Chebyshev polynomials as:

$$V^j\left(k_t|\theta^{V,j}\right) = \sum_{i=0}^{10}\theta_i^{V,j}T_i(k_t) \tag{41}$$

$$l^j\left(k_t|\theta^{l,j}\right) = \sum_{i=0}^{10}\theta_i^{l,k}T_i(k_t) \tag{42}$$

Once we have the decision rule for labor, we can find output:

$$y^j(k_t) = e^{z_t}k_t^{\alpha}\left(l^j\left(k_t|\theta^{l,j}\right)\right)^{1-\alpha},$$

With output, from the first-order condition that relates the marginal utility consumption and the marginal productivity of labor, we can find consumption:

$$c^j(k_t) = \frac{\tau}{1-\tau}(1-\alpha)e^{z_t}k_t^{\alpha}\left(l^j\left(k_t|\theta^{l,j}\right)\right)^{-\alpha}\left(1-l^j\left(k_t|\theta^{l,j}\right)\right) \tag{43}$$

and, from the resource constraint, capital next period:

$$k^j(k_t) = e^{z_t}k_t^{\alpha}\left(l^j\left(k_t|\theta^{l,j}\right)\right)^{1-\alpha} + (1-\delta)k_t - c^j(k_t) \tag{44}$$

Our notations $y^j(k_t)$, $c^j(k_t)$, and $k^j(k_t)$ emphasize the exact dependence of these three variables on capital and the productivity level: once we have approximated $l^j\left(k_t|\theta^{l,j}\right)$, simple algebra with the equilibrium conditions allows us to avoid further approximation.

We decided to approximate the value function and the decision rule for labor and use them to derive the other variables of interest to illustrate how flexible projection methods are. We could, as well, have decided to approximate the decision rules for consumption and capital and find labor and the value function using the equilibrium conditions. The researcher should pick the approximating functions that are more convenient, either for algebraic reasons or her particular goals.

To solve for the unknown coefficients θ^V and θ^l, we plug the functions (41), (42), (43), and (44) into the Bellman equation to get:

$$\sum_{i=0}^{10}\theta_i^{V,j}T_i(k_t) = \frac{\left(\left(c^j(k_t)\right)^{\theta}\left(1-\sum_{i=0}^{10}\theta_i^l T_i(k_t)\right)^{1-\theta}\right)^{1-\tau}}{1-\tau} + \beta\sum_{m=1}^{5}p_{jm}\sum_{i=0}^{10}\theta_i^{V,j}T_i\left(k^j(k_t)\right) \tag{45}$$

where, since we are already using the optimal decision rules, we can drop the max operator. Also, we have substituted the expectation by the sum operator and the transition probabilities p_{jm}. We plug the same functions (41), (42), (43), and (44) into the Euler equation to get:

$$\frac{\left(c_t^{\theta}\left(1-\sum_{i=0}^{10}\theta_i^{l,k}T_i(k_t)\right)^{1-\theta}\right)^{1-\tau}}{c_t}=\beta\mathbb{E}_t\sum_{m=1}^{5}p_{jm}\sum_{i=0}^{10}\theta_i^{V,j}T_i\left(k^j(k_t)\right)\prime, \tag{46}$$

where $T_i(k^j(k_t))\prime$ is the derivative of the Chebyshev polynomial with respect to its argument.

The residual equation groups Eqs. (45) and (46):

$$R\left(k_t, z_j\middle|\theta\right)=\begin{cases}\sum_{i=0}^{10}\theta_i^{V,j}T_i(k_t)-\dfrac{\left(\left(c^j(k_t)\right)^{\theta}\left(1-\sum_{i=0}^{10}\theta_i^{l}T_i(k_t)\right)^{1-\theta}\right)^{1-\tau}}{1-\tau}\\ -\beta\sum_{m=1}^{5}p_{jm}\sum_{i=0}^{10}\theta_i^{V,j}T_i\left(k^j(k_t)\right)\\ \dfrac{\left(c_t^{\theta}\left(1-\sum_{i=0}^{10}\theta_i^{l,k}T_i(k_t)\right)^{1-\theta}\right)^{1-\tau}}{c_t}-\beta\mathbb{E}_t\sum_{m=1}^{5}p_{jm}\sum_{i=0}^{10}\theta_i^{V,j}T_i\left(k^j(k_t)\right)\prime\end{cases}$$

where θ stacks $\theta^{V,j}$ and $\theta^{l,k}$. Given that we use 11 Chebyshev polynomials for the value function and another 11 for the decision rule for labor for each of the 5 levels of z_j, θ has 110 elements ($110 = 11 \times 2 \times 5$). If we evaluate the residual function at each of the 11 zeros of the Chebyshev of order 11 for capital and the 5 levels of z_j, we will have the 110 equations required to solve for those 110 coefficients. A Newton solver can easily deal with this system (although, as explained in Remark 24, using a multistep approach simplifies the computation: we used 3 Chebyshev polynomials in the first step and 11 Chebyshev polynomials in the second one).

We plot the main components of the solution in Fig. 10. The top left panel draws the value function, with one line for each of the five values of productivity and capital on the x-axis. As predicted by theory, the value function is increasing and concave in both state variables, k_t and z_t. We follow the same convention for the decision rules for consumption (top right panel), labor supply (bottom left panel), and capital next period, k_{t+1} (bottom right panel). The most noticeable pattern is the near linearity of the capital decision rule. Once the researcher has found the value function and all the decision rules, she can easily simulate the model, compute impulse response functions, and evaluate welfare.

The accuracy of the solution is impressive, with Euler equation errors below -13 in the log_{10} scale. Section 7 discusses how to interpret these errors. Suffice it to say here that, for practical purposes, the solution plotted in Fig. 10 can be used instead of the exact solution of the stochastic neoclassical growth model with a discrete productivity level.

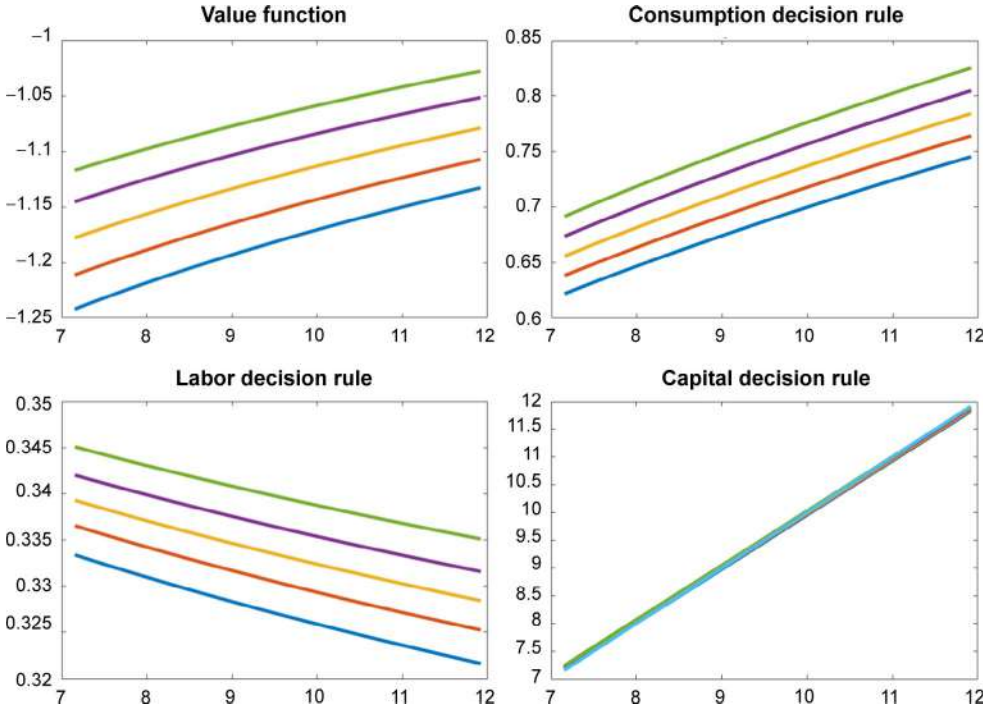


Fig. 10 Solution, stochastic neoclassical growth model.

5.7 Smolyak's Algorithm

An alternative to complete polynomials that can handle the curse of dimensionality better than other methods is Smolyak's algorithm. See Smolyak (1963), Delvos (1982), Barthelmann et al. (2000), and, especially, Bungartz and Griebel (2004) for a summary of the literature. Krüger and Kubler (2004) and Malin et al. (2011) introduced the algorithm in economics as a solution method for DSGE models. Subsequently, Smolyak's algorithm has been applied by many researchers. For example, Fernández-Villaverde et al. (2015a) rely on Smolyak's algorithm to solve a New Keynesian model with a ZLB (a model with 5 state variables), Fernández-Villaverde and Levintal (2016) exploit it to solve a New Keynesian model with big disasters risk (a model with 12 state variables), and Gordon (2011) uses it to solve a model with heterogeneous agents. Malin et al. (2011) can accurately compute a model with 20 continuous state variables and a considerable deal of curvature in the production and utility functions. In the next pages, we closely follow the explanations in Krüger and Kubler (2004) and Malin et al. (2011) and invite the reader to check those papers for further details.[z]

[z] There is also a promising line of research based on the use of ergodic sets to solve highly dimensional models (Judd et al., 2011b; Maliar et al., 2011; and Maliar and Maliar, 2015). Maliar and Maliar (2014) cover the material better than we could.

As before, we want to approximate a function (decision rule, value function, expectation, etc.) on n state variables, $d:[-1,1]^n \rightarrow \mathbb{R}$ (the generalization to the case $d:[-1,1]^n \rightarrow \mathbb{R}^m$ is straightforward, but tedious). The idea of Smolyak's algorithm is to find a grid of points $\mathbb{G}(q,n) \in [-1,1]^n$ where $q > n$ and an approximating function $d(x|\theta,\ q,\ n):[-1,1]^n \rightarrow \mathbb{R}$ indexed by some coefficients θ such that, at the points $x_i \in \mathbb{G}(q,n)$, the unknown function $d(\cdot)$ and $d(\cdot|\theta,\ q,\ n)$ are equal:

$$d(x_i) = d(x_i|\theta,\ q,\ n)$$

and, at the points $x_i \notin \mathbb{G}(q,n)$, $d(\cdot|\theta, q, n)$ is close to the unknown function $d(\cdot)$. In other words, at the points $x_i \in \mathbb{G}(q,\ n)$, the operator $\mathcal{H}(\cdot)$ would be exactly satisfied and, at other points, the residual function would be close to zero. The integer q indexes the size of the grid and, with it, the precision of the approximation.

The challenge is to judiciously select grid points $\mathbb{G}(q,n)$ in such a way that the number of coefficients θ does not explode with n. Smolyak's algorithm is (almost) optimal for that task within the set of polynomial approximations (Barthelmann et al., 2000). Also, the method is universal, that is, almost optimal for many different function spaces.

5.7.1 Implementing Smolyak's Algorithm

Our search of a grid of points $\mathbb{G}(q,n)$ and a function $d(x|\theta,q,n)$ will proceed in several steps.

5.7.1.1 First Step: Transform the Domain of the State Variables

For any state variable $\tilde{x}_l$, $l = 1,\ldots,n$ that has a domain $[a,b]$, we use a linear translation from $[a,b]$ into $[-1,1]$:

$$x_l = 2\frac{\tilde{x}_l - a}{b-a} - 1.$$

5.7.1.2 Second Step: Setting the Order of the Polynomial

We define $m_1 = 1$ and $m_i = 2^{i-1} + 1$, $i = 2,\ldots,$ where $m_i - 1$ will be the order of the polynomial that we will use to approximate $d(\cdot)$.

5.7.1.3 Third Step: Building the Gauss–Lobotto Nodes

We build the sets:

$$\mathcal{G}^i = \{\zeta_1^i,\ldots,\zeta_{m_i}^i\} \subset [-1,1]$$

that contain the Gauss–Lobotto nodes (also known as the Clenshaw–Curtis points), that is, the extrema of the Chebyshev polynomials:

$$\zeta_j^i = -cos\left(\frac{j-1}{m_i - 1}\pi\right), j = 1, \ldots, m_i$$

with the initial set $\mathcal{G}^1 = \{0\}$ (with a change of notation, this formula for the extrema is the same as the one in Eq. (34)). For instance, the first three sets are given by:

$$\mathcal{G}^1 = \{0\}, \text{ where } i = 1, m_1 = 1.$$
$$\mathcal{G}^2 = \{-1, 0, 1\}, \text{ where } i = 2, m_3 = 3.$$
$$\mathcal{G}^3 = \left\{-1, -\cos\left(\frac{\pi}{4}\right), 0, -\cos\left(\frac{3\pi}{4}\right), 1\right\}, \text{ where } i = 3, m_5 = 5.$$

Since, in the construction of the sets, we impose that $m_i = 2^{i-1} + 1$, we generate sets that are nested, that is, $\mathcal{G}^i \subset \mathcal{G}^{i+1}, \forall i = 1, 2, \ldots$. This result is crucial for the success of the algorithm.

5.7.1.4 Fourth Step: Building a Sparse Grid

For any integer q bigger than the number of state variables n, $q > n$, we define a sparse grid as the union of the Cartesian products:

$$\mathbb{G}(q, n) = \bigcup_{q-n+1 \le |\mathbf{i}| \le q} \left(\mathcal{G}^{i_1} \times \ldots \times \mathcal{G}^{i_n}\right),$$

where $|\mathbf{i}| = \sum_{l=1}^{n} i_l$.

To illustrate how this sparse grid works, imagine that we are dealing with a DSGE model with two continuous state variables. If we pick $q = 2 + 1 = 3$, we have the sparse grid

$$\begin{aligned}\mathbb{G}(3, 2) &= \bigcup_{2 \le |\mathbf{i}| \le 3} \left(\mathcal{G}^{i_1} \times \mathcal{G}^{i_2}\right) \\ &= \left(\mathcal{G}^1 \times \mathcal{G}^1\right) \cup \left(\mathcal{G}^1 \times \mathcal{G}^2\right) \cup \left(\mathcal{G}^2 \times \mathcal{G}^1\right) \\ &= \{(-1, 0), (0, 1), (0, 0), (0, -1), (1, 0)\}\end{aligned}$$

We plot this grid in the top left panel of Fig. 11, which reproduces fig. 1 in Krüger and Kubler (2004).

If we pick $q = 2 + 2 = 4$, we have the sparse grid

$$\begin{aligned}\mathbb{G}(4, 2) &= \bigcup_{3 \le |\mathbf{i}| \le 4} \left(\mathcal{G}^{i_1} \times \mathcal{G}^{i_2}\right) \\ &= \left(\mathcal{G}^1 \times \mathcal{G}^2\right) \cup \left(\mathcal{G}^1 \times \mathcal{G}^3\right) \cup \left(\mathcal{G}^2 \times \mathcal{G}^2\right) \cup \left(\mathcal{G}^3 \times \mathcal{G}^1\right) \\ &= \left\{\begin{array}{c} (-1, 1), (-1, 0), (-1, -1), \left(-\cos\left(\frac{\pi}{4}\right), 0\right), \\ (0, 1), \left(0, -\cos\left(\frac{3\pi}{4}\right)\right), (0, 0), \left(0, -\cos\left(\frac{\pi}{4}\right)\right), \\ (0, -1), \left(-\cos\left(\frac{3\pi}{4}\right), 0\right), (1, 1), (1, 0), (1, -1) \end{array}\right\}\end{aligned}$$

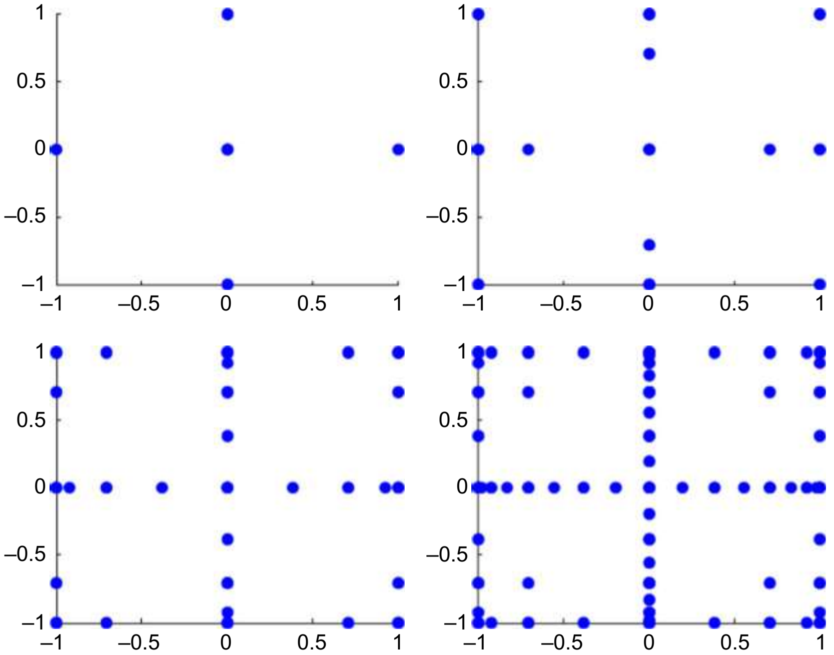


Fig. 11 Four sparse grids

We plot this grid in the top right panel of Fig. 11. Note that the sparse grids have a hierarchical structure, where $\mathbb{G}(3,2) \in \mathbb{G}(4,2)$ or, more generally, $\mathbb{G}(q,n) \in \mathbb{G}(q+1,n)$.

Following the same strategy, we can build $\mathbb{G}(5,2)$, plotted in the bottom left panel of Fig. 11, and $\mathbb{G}(6,2)$, plotted in the bottom right panel of Fig. 11 (in the interest of concision, we skip the explicit enumeration of the points of these two additional grids). In Fig. 12, we plot a grid for a problem with 3 state variables, $\mathbb{G}(5,3)$.

The sparse grid has two important properties. First, the grid points cluster around the corners of the domain of the Chebyshev polynomials and the central cross. Second, the number of points in a sparse grid when $q = n + 2$ is given by $1 + 4n + 2n(n - 1)$. The cardinality of this grid grows polynomially on n^2. Similar formulae hold for other $q > n$. For example, the cardinality of the grid grows polynomially on n^3 when $q = n + 3$. In fact, the computational burden of the method notably increases as we keep n fixed and a rise q. Fortunately, experience suggests that $q = n + 2$ and $q = n + 3$ are usually enough to deliver the desired accuracy in DSGE models.

The nestedness of the sets of the Gauss–Lobotto nodes plays a central role in controlling the cardinality of $\mathbb{G}(q,n)$. In comparison, the number of points in a rectangular grid is 5^n, an integer that grows exponentially on n. If $n = 2$, this would correspond, in the top right panel of Fig. 11, to having all possible tensors of

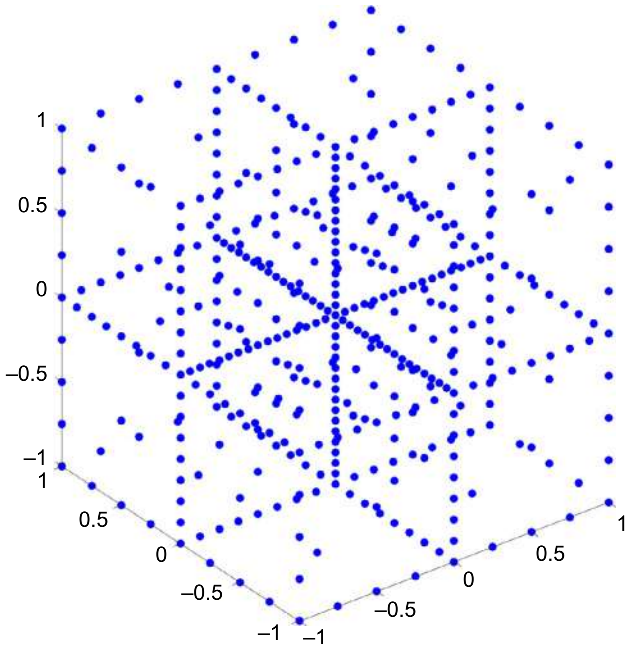


Fig. 12 A sparse grid, 3 state variables.

Table 3 Size of the grid for $q = n + 2$

n	$\mathbb{G}(q,n)$	5^n
2	13	25
3	25	125
4	41	625
5	61	3125
12	313	244,140,625

$\left\{-1, -\cos\left(\frac{\pi}{4}\right), 0, -\cos\left(\frac{3\pi}{4}\right), 1\right\}$ and $\left\{-1, -\cos\left(\frac{\pi}{4}\right), 0, -\cos\left(\frac{3\pi}{4}\right), 1\right\}$ covering the whole of the $[-1,1]^2$ square. Instead of keeping these 25 points, Smolyak's algorithm eliminates 12 of them and only keeps 13. To illustrate how dramatic is the difference between polynomial and exponential growth, Table 3 shows the cardinality of both grids as we move from 2 state variables to 12.

5.7.1.5 Fifth Step: Building Tensor Products

We use the Chebyshev polynomials $\psi_i(x_i) = T_{i-1}(x_i)$ to build the tensor-product multivariate polynomial:

$$p^{|\mathbf{i}|}(x|\theta) = \sum_{l_1=1}^{m_{i_1}} \dots \sum_{l_n=1}^{m_{i_n}} \theta_{l_1 \dots l_n} \psi_{l_1}(x_1) \dots \psi_{l_n}(x_n)$$

where $|\mathbf{i}| = \sum_{l=1}^{n} i_l$, $x_i \in [-1,1]$, $x = \{x_1, \dots, x_n\}$, and θ stacks all the coefficients $\theta_{l_1 \dots l_n}$. So, for example, for a DSGE model with two continuous state variables and $q = 3$, we will have:

$$\begin{aligned}
p^{\mathbf{1},\mathbf{1}}(x|\theta) &= \sum_{l_1=1}^{m_1} \sum_{l_n=1}^{m_1} \theta_{l_1 l_2} \psi_{l_1}(x_1) \psi_{l_2}(x_2) = \theta_{11} \\
p^{\mathbf{1},\mathbf{2}}(x|\theta) &= \sum_{l_1=1}^{m_1} \sum_{l_n=1}^{m_2} \theta_{l_1 l_2} \psi_{l_1}(x_1) \psi_{l_2}(x_2) = \theta_{11} + \theta_{12} T_1(x_2) + \theta_{13} T_2(x_2) \\
p^{2,\mathbf{1}}(x|\theta) &= \sum_{l_1=1}^{m_2} \sum_{l_n=1}^{m_1} \theta_{l_1 l_2} \psi_{l_1}(x_1) \psi_{l_2}(x_2) = \theta_{11} + \theta_{21} T_1(x_1) + \theta_{31} T_2(x_1)
\end{aligned}$$

where we have already used $T_0(x_i) = 1$. Therefore, for $x = \{x_1, x_2\}$:

$$\begin{aligned}
p^{|\mathbf{2}|}(x|\theta) &= p^{\mathbf{1},\mathbf{1}}(x|\theta) \\
p^{|\mathbf{3}|}(x|\theta) &= p^{\mathbf{1},\mathbf{2}}(x|\theta) + p^{2,\mathbf{1}}(x|\theta).
\end{aligned}$$

Most conveniently, for an arbitrary grid with points $k_1, \dots, k_n > 1$ along each dimension, these coefficients are given by:

$$\theta_{l_1 \dots l_n} = \frac{2^n}{(k_1 - 1) \dots (k_n - 1)} \frac{1}{c_{l_1} \dots c_{l_n}} \sum_{j_1=1}^{k_1} \dots \sum_{j_n=1}^{k_n} \frac{1}{c_{j_1} \dots c_{j_n}} \psi_{l_1}(\zeta_1) \dots \psi_{l_d}(\zeta_n) d(\zeta_1, \dots, \zeta_n) \quad (47)$$

where $c_j = 1$ for all j, except for the cases $c_1 = c_{k_d} = 2$, and $\zeta_k \in \mathcal{G}^i$ are the Gauss–Lobotto nodes. This approximation is exact in the Gauss–Lobotto nodes and interpolates among them.

There is nothing special about the use of Chebyshev polynomials as the basis functions $\psi_j(x)$ and we could rely, if required, on other basis functions. For instance, one can implement a finite element method with the Smolyak algorithm by partitioning $\mathbf{\Omega}$ into elements and defining local basis functions as in Nobile et al. (2008). We use Chebyshev polynomials just because they have been popular in the applications of the Smolyak algorithm in macroeconomics.

5.7.1.6 Sixth Step: Building the Interpolating Function in n Dimensions

The Smolyak function that interpolates on $\mathbb{G}(q,n)$ is:

$$d(x|\theta, q, n) = \sum_{\max(n, q-n+1) \leq |\mathbf{i}| \leq q} (-1)^{q-|\mathbf{i}|} \binom{n-1}{q-|\mathbf{i}|} p^{|\mathbf{i}|}(x|\theta),$$

which is nothing more than the weighted sum of the tensors. In our previous example, a DSGE model with two continuous state variables and $q=3$, we will have the sparse grid:

$$\mathbb{G}(3,2)=\{(-1,0),(0,1),(0,0),(0,-1),(1,0)\}$$

(this sparse grid was drawn in the top left panel of Fig. 11) and:

$$\begin{aligned} d(x|\theta,q,n) &= \sum_{2\leq|\mathbf{i}|\leq 3}(-1)^{3-|\mathbf{i}|}\binom{1}{3-|\mathbf{i}|}p^{|\mathbf{i}|}(x|\theta) \\ &= (-1)\binom{1}{1}p^{|\mathbf{2}|}(x|\theta)+(-1)^{0}\binom{1}{0}p^{|\mathbf{3}|}(x|\theta) \\ &= p^{\mathbf{1},\mathbf{2}}(x|\theta)+p^{2,\mathbf{1}}(x|\theta)-p^{\mathbf{1},\mathbf{1}}(x|\theta) \\ &= \theta_{11}+\theta_{21}T_1(x_1)+\theta_{31}T_2(x_1)+\theta_{12}T_1(x_2)+\theta_{13}T_2(x_2). \end{aligned}$$

Each of the coefficients in this approximation is given by the formula in Eq. (47):

$$\begin{aligned} \theta_{21} &= \frac{1}{2}(d(1,0)-d(-1,0)) \\ \theta_{12} &= \frac{1}{2}(d(0,1)-d(0,-1)) \\ \theta_{31} &= \frac{1}{4}(d(1,0)+d(-1,0))-\frac{1}{2}d(0,0) \\ \theta_{13} &= \frac{1}{4}(d(0,1)+d(0,-1))-\frac{1}{2}d(0,0) \end{aligned}$$

except the constant term:

$$\theta_{11}=\frac{1}{4}(d(0,1)+d(0,-1)+d(1,0)+d(-1,0)),$$

which instead ensures that the interpolating function satisfies $d(0,0)=d(x|\theta,q,n)$. It is easy to check that we indeed satisfy the condition that the approximating function equates the unknown function at the points of the sparse grid. For example, at $(-1,0)$:

$$\begin{aligned} d((-1,0)|\theta,q,n) &= \theta_{11}+\theta_{21}T_1(-1)+\theta_{31}T_2(-1)+\theta_{12}T_1(0)+\theta_{13}T_2(0) \\ &= \theta_{11}-\theta_{21}+\theta_{31}-\theta_{13} \\ &= \frac{1}{4}(d(0,1)+d(0,-1)+d(1,0)+d(-1,0)) \\ &\quad -\frac{1}{2}(d(1,0)-d(-1,0)) \\ &\quad +\frac{1}{4}(d(1,0)+d(-1,0))-\frac{1}{2}d(0,0) \\ &\quad -\frac{1}{4}(d(0,1)+d(0,-1))+\frac{1}{2}d(0,0) \\ &= d(-1,0). \end{aligned}$$

An interesting property of this construction of $d(x|\theta,q,n)$ is that the cardinality of $\mathbb{G}(q,n)$ and the number of coefficients on θ coincide. In our previous example, $\mathbb{G}(3,2)=5$ and $\theta=\{\theta_{11},\theta_{21},\theta_{31},\theta_{12},\theta_{13}\}$. A second relevant property is that $d(x|\theta,q,n)$ exactly replicates any polynomial function built with monomials of degree less than or equal to $q-n$.

5.7.1.7 Seventh Step: Solving for the Polynomial Coefficients

We plug $d(x|\theta,q,n)$ into the operator $\mathcal{H}(\cdot)$ for all $x_i\in\mathbb{G}(q,n)$. At this point the operator needs to be exactly zero:

$$\mathcal{H}(d(x_i|\theta,q,n))=0$$

and we solve for the unknown coefficients on θ. In our previous example, we had $\mathbb{G}(3,2)=\{(-1,0),(0,1),(0,0),(0,-1),(1,0)\}$ and, therefore:

$$d((-1,0)|\theta,q,n)=\theta_{11}+\theta_{21}T_1(-1)+\theta_{31}T_2(-1)+\theta_{12}T_1(0)+\theta_{13}T_2(0)=\theta_{11}-\theta_{21}+\theta_{31}-\theta_{13}$$

$$d((0,1)|\theta,q,n)=\theta_{11}+\theta_{21}T_1(0)+\theta_{31}T_2(0)+\theta_{12}T_1(1)+\theta_{13}T_2(1)=\theta_{11}-\theta_{31}+\theta_{12}+\theta_{13}$$

$$d((0,0)|\theta,q,n)=\theta_{11}+\theta_{21}T_1(0)+\theta_{31}T_2(0)+\theta_{12}T_1(0)+\theta_{13}T_2(0)=\theta_{11}-\theta_{31}-\theta_{13}$$

$$d((0,-1)|\theta,q,n)=\theta_{11}+\theta_{21}T_1(0)+\theta_{31}T_2(0)+\theta_{12}T_1(-1)+\theta_{13}T_2(-1)=\theta_{11}-\theta_{31}-\theta_{12}+\theta_{13}$$

$$d((1,0)|\theta,q,n)=\theta_{11}+\theta_{21}T_1(1)+\theta_{31}T_2(1)+\theta_{12}T_1(0)+\theta_{13}T_2(0)=\theta_{11}+\theta_{21}+\theta_{31}-\theta_{13}$$

The system of equations:

$$\mathcal{H}(d(x_i|\theta,q,n))=0,\, x_i\in\mathbb{G}(q,n)$$

can be solved with a standard nonlinear solver. Krüger and Kubler (2004) and Malin et al. (2011) suggest a time-iteration method that starts, as an initial guess, from the first-order perturbation of the model. This choice is, nevertheless, not essential to the method.

5.7.2 Extensions

Recently, Judd et al. (2014b) have proposed an important improvement of Smolyak's algorithm. More concretely, the authors first present a more efficient implementation of Smolyak's algorithm that uses disjoint-set generators that are equivalent to the sets $\mathcal{G}^i$. Second, the authors use a Lagrange interpolation scheme. Third, the authors build an anisotropic grid, which allows having a different number of grid points and basis functions for different state variables. This may be important to capture the fact that, often, it is harder to approximate the decision rules of agents along some dimensions than along others. Finally, the authors argue that it is much more efficient to employ a derivative-free fixed-point iteration method instead of the time-iteration scheme proposed by Krüger and Kubler (2004) and Malin et al. (2011).

In comparison, Brumm and Scheidegger (2015) keep a time-iteration procedure, but they embed on it an adaptive sparse grid. This grid is refined locally in an automatic fashion, which allows the capture of steep gradients and some nondifferentiabilities. The authors provide a fully hybrid parallel implementation of the method, which takes advantage of the fast improvements in massively parallel processing.

6. COMPARISON OF PERTURBATION AND PROJECTION METHODS

After our description of perturbation and projection methods, we can offer some brief comments on their relative strengths and weaknesses.

Perturbation methods have one great advantage: their computational efficiency. We can compute, using a standard laptop computer, a third-order approximation to DSGE models with dozens of state variables in a few seconds. Perturbation methods have one great disadvantage: they only provide a local solution. The Taylor series expansion is accurate around the point at which we perform the perturbation and deteriorates as we move away from that point. Although perturbation methods often yield good global results (see Aruoba et al., 2006; Caldara et al., 2012; and Swanson et al., 2006), such performance needs to be assessed in each concrete application and even a wide range of accuracy may not be sufficient for some quantitative experiments. Furthermore, perturbation relies on differentiability conditions that are often violated by models of interest, such as those that present kinks or occasionally binding constraints.[aa]

Projection methods are nearly the mirror image of perturbation. Projection methods have one great advantage: Chebyshev and finite elements produce solutions that are of high accuracy over the whole range of state variable values. See, again, Aruoba et al. (2006) and Caldara et al. (2012). And projection methods can attack even the most complex problems with occasionally binding constraints, irregular shapes, and local behavior. But power and flexibility come at a cost: computational effort. Projection methods are harder to code, take longer to run, and suffer, as we have repeatedly pointed out, from an acute curse of dimensionality.[ab]

Thus, which method to use in real life? The answer, not unsurprisingly, is "it depends." Solution methods for DSGE models provide a menu of options. If we are dealing, for example, with a standard middle-sized New Keynesian model with

[aa] Researchers have proposed getting around these problems with different devices, such as the use of penalty functions. See, for example, Preston and Roca (2007). In fact, the recent experience of several central banks pushing their target interest rates below zero suggests that many constraints such as the ZLB may be closer to such a penalty function than to a traditional kink.

[ab] The real bottleneck for most research projects involving DSGE models is coding time, not running time. Moving from a few seconds of running time with perturbation to a few minutes of running time with projection is a minuscule fraction of the cost of coding a finite elements method in comparison with the cost of employing Dynare to find a perturbation.

25 state variables, perturbation methods are likely to be the best option. The New Keynesian model is sufficiently well behaved that a local approximation would be good enough for most purposes. A first-order approximation will deliver accurate estimates of the business cycle statistics such as variances and covariances, and a second- or third-order approximation is likely to generate good welfare estimates (although one should always be careful when performing welfare evaluations). If we are dealing, in contrast, with a DSGE model with financial constraints, large risk aversion, and only a few state variables, a projection method is likely to be a superior option. An experienced researcher may even want to have two different solutions to check one against the other, perhaps of a simplified version of the model, and decide which one provides her with a superior compromise between coding time, running time, and accuracy.

Remark 25 (Hybrid methods) The stark comparison between perturbation and projection methods hints at the possibility of developing hybrid methods that combine the best of both approaches. Judd (1998, section 15.6), proposes the following hybrid algorithm:

Algorithm 4 (Hybrid algorithm)

1. Use perturbation to build a basis tailored to the DSGE model we need to solve.
2. Apply a Gram-Schmidt process to build an orthogonal basis from the basis obtained in 1.
3. Employ a projection method with the basis from 2.

While this algorithm is promising (see the example provided by Judd, 1998), we are unaware of further explorations of this proposal.

More recently, Levintal (2015b) and Fernández-Villaverde and Levintal (2016) have proposed the use of Taylor-based approximations that also have the flavor of a hybrid method. The latter paper shows the high accuracy of this hybrid method in comparison with pure perturbation and projection methods when computing a DSGE model with disaster risk and a dozen state variables. Other hybrid proposals include Maliar et al. (2013).

7. ERROR ANALYSIS

A final step in every numerical solution of a DSGE model is to assess the error created by the approximation, that is, the difference between the exact and the approximated solution. This may seem challenging since the exact solution of the model is unknown. However, the literature has presented different methods to evaluate the errors.[ac] We will concentrate on the two most popular procedures to assess error: χ^2 −test proposed by

[ac] Here we follow much of the presentation of Aruoba et al. (2006), where the interested reader can find more details.

Den Haan and Marcet (1994) and the Euler equation error proposed by Judd (1992). Throughout this section, we will use the superscript j to index the perturbation order, the number of basis functions, or another characteristic of the solution method. For example, $c^j(k_t, z_t)$ will be the approximation to the decision rule for consumption $c(k_t, z_t)$ in a model with state variables k_t and z_t.

Remark 26 (Theoretical bounds) There are (limited) theoretical results bounding the approximation errors and their consequences. Santos and Vigo-Aguiar (1998) derive upper bounds for the error in models computed with value function iteration. Santos and Rust (2004) extend the exercise for policy function iteration. Santos and Peralta-Alva (2005) propose regularity conditions under which the error from the simulated moments of the model converge to zero as the approximated equilibrium function approaches the exact, but unknown, equilibrium function. Fernández-Villaverde et al. (2006) explore similar conditions for likelihood functions and Stachurski and Martin (2008) perform related work for the computation of densities of ergodic distributions of variables of interest. Judd et al. (2014a) have argued for the importance of constructing lower bounds on the size of approximation errors and propose a methodology to do so. Kogan and Mitra (2014) have studied the information relaxation method of Brown et al. (2010) to measure the welfare cost of using approximated decision rules. Santos and Peralta-Alva (2014) review the existing literature. But, despite all this notable work, this is an area in dire need of further investigation.

Remark 27 (Preliminary assessments) Before performing a formal error analysis, researchers should undertake several preliminary assessments. First, we need to check that the computed solution satisfies theoretical properties, such as concavity or monotonicity of the decision rules. Second, we need to check the shape and structure of decision rules, impulse response functions, and basic statistics of the model. Third, we need to check how the solution varies as we change the calibration of the model.

These steps often tell us more about the (lack of) accuracy of an approximated solution than any formal method. Obviously, the researcher should also take aggressive steps to verify that her code is correct and that she is, in fact, computing what she is supposed to compute. The use of modern, industry-tested software engineering techniques is crucial in ensuring code quality.

7.1 A χ^2 Accuracy Test

Den Haan and Marcet (1994) noted that, if some of the equilibrium conditions of the model are given by:

$$f(y_t) = \mathbb{E}_t(\phi(y_{t+1}, y_{t+2}, ..))$$

where the vector y_t contains n variables of interest at time t, $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $\phi : \mathbb{R}^n \times \mathbb{R}^\infty \rightarrow \mathbb{R}^m$ are known functions, then:

$$\mathbb{E}_t(u_{t+1} \otimes h(x_t)) = 0 \tag{48}$$

for any vector x_t measurable with respect to t with $u_{t+1} = \phi(y_{t+1}, y_{t+2}, ..) - f(y_t)$ and $h : \mathbb{R}^k \rightarrow \mathbb{R}^q$ being an arbitrary function.

If we simulate a series of length T from the DSGE model using a given solution method, $\left\{y_t^j\right\}_{t=1:T}$, we can find $\left\{u_{t+1}^j, x_t^j\right\}_{t=1:T}$ and compute the sample analog of (48):

$$B_T^j = \frac{1}{T}\sum_{t=1}^{T} u_{t+1}^j \otimes h\left(x_t^j\right). \tag{49}$$

The moment (49) would converge to zero as N increases almost surely if we were using the exact solution to the model. When, instead, we are using an approximation, the statistic $B\left(B_T^j\right)' \left(A_T^j\right)^{-1} B_T^j$ where A_T^j is a consistent estimate of the matrix:

$$\sum_{t=-\infty}^{\infty} \mathbb{E}_t\left[(u_{t+1} \otimes h(x_t))(u_{t+1} \otimes h(x_t))'\right]$$

converges to a χ^2 distribution with qm degrees of freedom under the null that the population moment (48) holds. Values of the test above the critical value can be interpreted as evidence against the accuracy of the solution. Since any solution method is an approximation, as T grows we will eventually reject the null. To control for this problem, Den Haan and Marcet (1990) suggest repeating the test for many simulations and report the percentage of statistics in the upper and lower critical 5% of the distribution. If the solution provides a good approximation, both percentages should be close to 5%.

This χ^2−test helps the researcher to assess how the errors of the approximated solution accumulate over time. Its main disadvantage is that rejections of accuracy may be difficult to interpret.

7.2 Euler Equation Errors

Judd (1992) proposed determining the quality of the solution method by defining normalized Euler equation errors. The idea is to measure how close the Euler equation at the core of nearly DSGE models is to be satisfied when we use the approximated solution.

The best way to understand how to implement this idea is with an example. We can go back to the stochastic neoclassical growth model that we solved in Section 5.6. This model generates an Euler equation:

$$u_{c'}(c_t, l_t) = \beta\mathbb{E}_t\{u_{c'}(c_{t+1}, l_{t+1})R_{t+1}\} \tag{50}$$

where

$$u_c\prime(c_t,l_t)=\frac{\left(c_t^{\tau}(1-l_t)^{1-\tau}\right)^{1-\eta}}{c_t}$$

is the marginal utility of consumption and $R_{t+1}=\left(1+\alpha e^{z_{t+1}}k_t^{\alpha-1}l_{t+1}^{1-\alpha}-\delta\right)$ is the gross return rate of capital. If we take the inverse of the marginal utility of consumption and do some algebra manipulations, we get:

$$1-\frac{u_c'\left(\beta\mathbb{E}_t\left\{u_c'(c_{t+1},l_{t+1})R_{t+1}\right\},l_t\right)^{-1}}{c_t}=0 \tag{51}$$

If we plug into Eq. (51) the exact decision rules for consumption:

$$c_t=c(k_t,z_t),$$

labor

$$l_t=l(k_t,z_t)$$

and capital:

$$k_{t+1}=k(k_t,z_t)$$

we get:

$$1-\frac{u_c'\left(\beta\mathbb{E}_t\left\{u_c'(c(k(k_t,z_t),z_{t+1}),l(k(k_t,z_t),z_{t+1}))R_{t+1}(k_t,z_t,z_{t+1})\right\},l(k_t,z_t)\right)^{-1}}{c(k_t,z_t)}=0 \tag{52}$$

where $R(k_t,z_t,z_{t+1})=\left(1+\alpha e^{z_{t+1}}k(k_t,z_t)^{\alpha-1}l(k(k_t,z_t),z_{t+1})^{1-\alpha}-\delta\right)$. Eq. (52) will hold exactly for any k_t and z_t.

If, instead, we plug into Eq. (52) the approximated decision rules $c^j(k_t,z_t)$, $l^j(k_t,z_t)$, and $k^j(k_t,z_t)$, we will have:

$$EEE(k_t,z_t)=\left\{\frac{1-}{u_c'\left(\beta\mathbb{E}_t\left\{u_c'\left(c^j\left(k^j\left(k_t^j,z_t\right),z_{t+1}\right),l^j(k^j(k_t,z_t),z_{t+1})\right)R_{t+1}^j(k_t,z_t,z_{t+1})\right\},l^j(k_t,z_t)\right)^{-1}}{c^j(k_t,z_t)}\right\} \tag{53}$$

where $R^j(k_t,z_t,z_{t+1})=\left(1+\alpha e^{z_{t+1}}k^j(k_t,z_t)^{\alpha-1}l^i\left(k^i(k_t,z_t),z_{t+1}\right)^{1-\alpha}-\delta\right)$. Eq. (53) defines a function, $EEE(k_t,z_t)$, that we call the Euler equation error.

We highlight three points about Eq. (53). First, the error in the Euler equation depends on the value of the state variables k_t and z_t. Perturbation methods will tend to have a small Euler equation error close to the point where the perturbation is undertaken and a larger Euler equation error farther from it. In contrast, projection methods will deliver a more uniform Euler equation error across Ω. Consequently, researchers have found it useful to summarize the Euler equation error. Proposals include the mean of the Euler equation error (either a simple average or using some estimate of the ergodic distribution of state variables[ad]) or the maximum of the Euler equation error in some region of Ω. Second, due to the algebraic transformation that we took on the Euler equation, $EEE(k_t, z_t)$ is expressed in consumption units, which have a meaningful economic interpretation as the relative optimization error incurred by the use of the approximated policy rule (Judd and Guu, 1997). For instance, if $EEE(k_t, z_t) = 0.01$, then the agent is making a $1 mistake for each $100 spent. In comparison, $EEE(k_t, z_t) = 1e^{-6}$ implies that the agent is making a 1 cent mistake for each 1 million spent. Third, the Euler equation error is also important because we know that, under certain conditions, the approximation error of the decision rule is of the same order of magnitude as the size of the Euler equation error. Correspondingly, the change in welfare is of the square order of the Euler equation error. Furthermore, the constants involved in these error bounds can be related to model primitives (Santos, 2000). Unfortunately, in some DSGE models it can be difficult to use algebraic transformations to achieve an expression for the Euler equation error that is interpretable as consumption units (or other natural economic unit).

Following the convention in the literature, we plot in Fig. 13, the $\log_{10}|EEE(k_t, z_t)|$ of the stochastic neoclassical growth model from Section 5.6. Taking the $\log_{10}$ eases reading: a value of -3 means $1 mistake for each $1000, a value of -4 a $1 mistake for each $10,000, and so on. Fig. 13 shows five lines, one for each value of productivity. As we hinted when we described the Chebyshev-collocation projection method, this accuracy is outstanding.

To compare this performance of Chebyshev-collocation with other solution methods, we reproduce, in Figs. 14 and 15, results from Aruoba et al. (2006). That paper uses the same stochastic neoclassical growth model with only a slightly different calibration (plus a few smaller details about how to handle z_t). Both figures display a transversal cut of the Euler equation errors when $z_t = 0$ and for values of capital between 70% and 130% of its steady-state value (23.14).

In Fig. 14, we plot the results for a first-order perturbation (in levels and in logs), a second-order perturbation, and a fifth-order perturbation. First, perturbations have smaller errors around the steady-state value of capital and deteriorate away from it. Second, there is a considerable improvement when we go from a first- to a second-order approximation. Third, a fifth-order approximation displays a great performance even 30% away from the steady state.

[ad] Using the ergodic distribution has the complication that we may not have access to it, since it is derived from the solution of the model, the object we are searching for. See Aruoba et al. (2006) for suggestions on how to handle this issue.

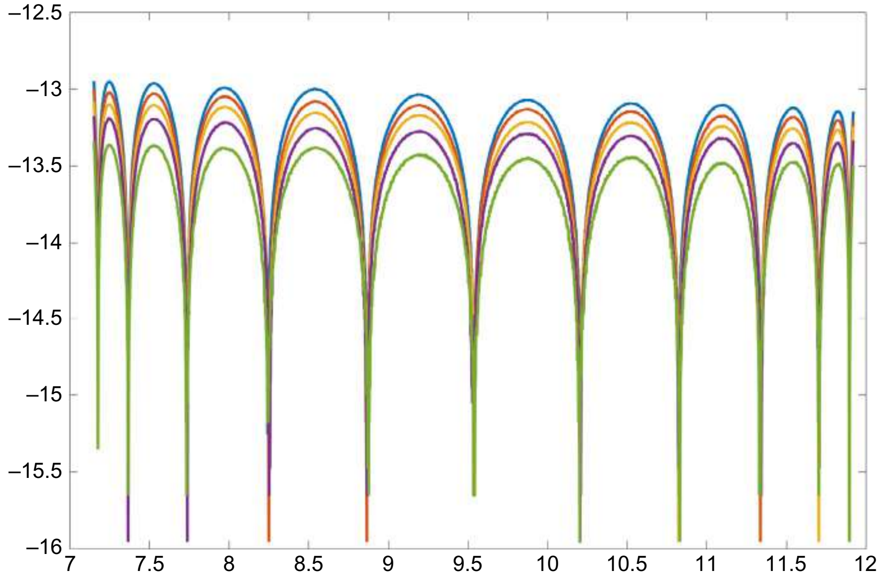


Fig. 13 *Log_{10}* of absolute value of Euler equation error.

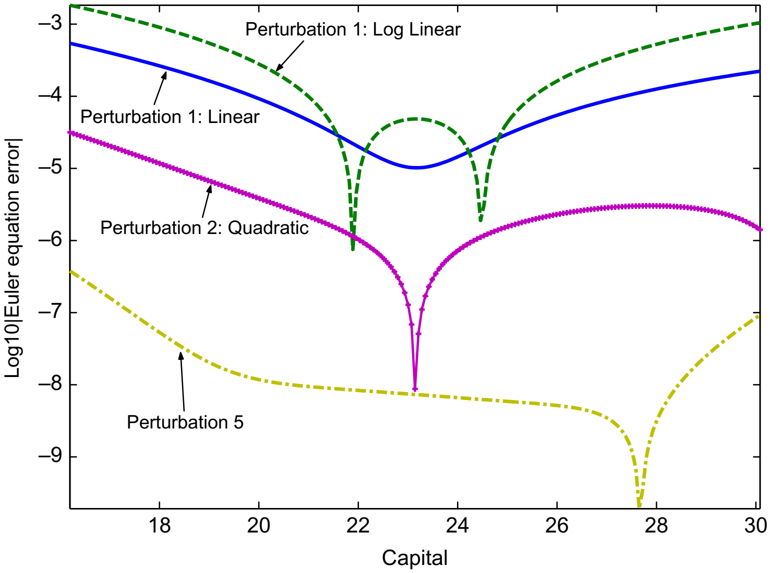


Fig. 14 *Log_{10}* of absolute value of Euler equation error.

In Fig. 15, we plot the results from the first-order perturbation (as a comparison with the previous graph), value function iteration (with a grid of one million points: 25,000 points for capital and 40 for the productivity level), finite elements (with 71 elements), and Chebyshev polynomials (as in Section 5.6, still with 11 polynomials). The main

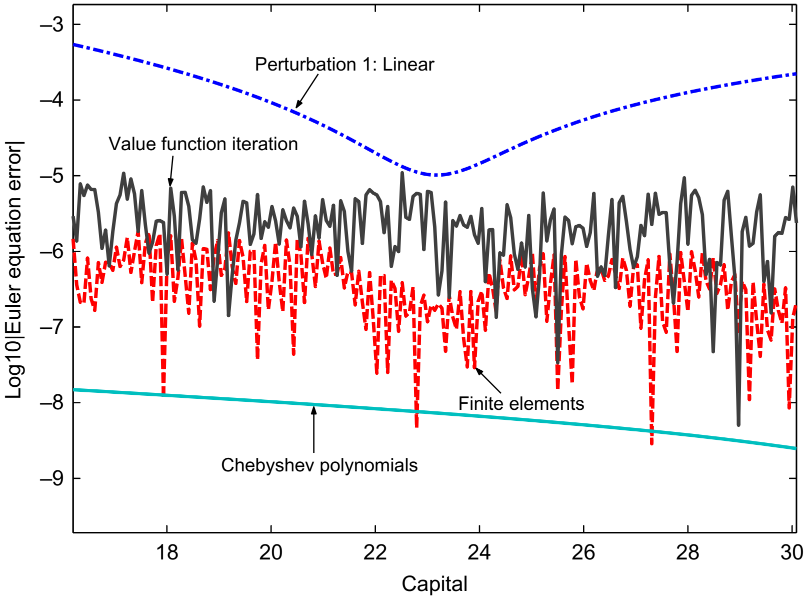


Fig. 15 *Log_{10}* of absolute value of Euler equation error.

lesson from this graph is that the Euler equation errors are much flatter for projection methods and value function iteration (another algorithm that delivers a global solution). The level of each of the three functions is harder to interpret, since it depends on the number of grid points (value function iteration), elements (finite elements), and Chebyshev polynomials. Nevertheless, the performance of Chebyshev is again excellent and its run time much lower than value function iteration and finite elements. This is not a surprise, since the decision rules for the stochastic neoclassical growth model are sufficiently well behaved for a spectral basis to do an extraordinary job.

Computing the Euler equation error has become standard in the literature because it often offers sharp assessments. However, Euler equation errors fail at giving a clear evaluation of how the errors of the approximated solution accumulate over time (see Santos and Peralta-Alva, 2005, for how to think about the impact of Euler equation errors on computed moments from the model). Thus, Euler equation errors should be understood as a complement to, not a substitute for, Den Haan and Marcet (1994)'s χ^2 −test.

7.3 Improving the Error

Once we have gauged the error in the solution to the DSGE, we can decide whether to improve the accuracy of the solution. Everything else equal, more accuracy is better than less accuracy. But, in real-life applications, everything else is rarely equal. More accuracy can come at the cost of more coding time and, in particular, longer running time. For example, in the exercise with the stochastic neoclassical growth model reported in Fig. 15, we could subdivide the finite elements as much as we want and use modern scientific libraries such as the *GNU multiple precision arithmetic library* to achieve any arbitrary level of accuracy, but at the cost of longer running times and more memory requirements. The researcher must look at her needs and resources and, once inferior solution methods are rejected, select those that best fit her goals.

But if the goal is indeed dependent on achieving additional accuracy, there are different possibilities available. If a perturbation is being used, we can increase the order of the perturbation. If a projection method is being used, we can increase the number of elements in the basis. The researcher can also explore changes of variables to make the problem more linear or switch the solution method.

Once the error of the model has been assessed, we are finally ready to move to Part II and see how the DSGE model can account for the observed data.

PART II. ESTIMATING DSGE MODELS

8. CONFRONTING DSGE MODELS WITH DATA

The preceding sections discussed how to compute an approximate solution for a DSGE model conditional on its parameterization. Part II focuses on determining the DSGE model parameters based on the empirical evidence and assessing the model's fit. More specifically, we ask four fundamental questions: (i) How can one estimate the DSGE model parameters from the observed macroeconomic time series? (ii) How well does the estimated DSGE model capture salient features of the data? (iii) What are the quantitative implications of the estimated DSGE models with respect to, for instance, sources of business cycle fluctuations, propagation of exogenous shocks, the effect of changes in macroeconomic policies, and the future path of macroeconomic time series? (iv) How should one construct measures of uncertainty for the parameters and the quantitative implications of the DSGE model? To answer these questions, we begin by analytically solving a stylized New Keynesian DSGE model in Section 8.1 and studying its properties in Section 8.2. DSGE model-implied population moments, autocovariances, spectra, and impulse response functions have sample analogs in the data, which are examined in Section 8.3. Macroeconomic time series exhibit trends that may or may not be captured by the DSGE model, which is discussed in Section 8.4.

Part II of this chapter assumes that the reader has some basic familiarity with econometrics, at the level of a first-year PhD sequence in a US graduate program. With the exception of Canova (2007) and DeJong and Dave (2007) there are no textbooks that focus on the estimation of DSGE models. The literature has progressed quickly since these two books were first written. The subsequent sections contain, in addition to a critical introduction to "standard methods," an overview of the most recent developments in the literature, which include identification conditions for DSGE models, identification-robust frequentist inference, and sequential Monte Carlo techniques for Bayesian analysis. Unlike the recent monograph by Herbst and Schorfheide (2015) which focuses on Bayesian computations, Part II of this chapter also contains extensive discussions of the consequences of misspecification for econometric inference and covers frequentist methods.

8.1 A Stylized DSGE Model

Throughout Part II we consider a stylized New Keynesian DSGE model in its loglinearized form.[ae] This model shares many of the features of its more realistic siblings that have been estimated in the literature. It is a stripped-down version of the model developed in the work by Christiano et al. (2005) and Smets and Wouters (2003). The specific version presented below is taken from Del Negro and Schorfheide (2008) and obtained by imposing several parameter restrictions. It is not suitable to be confronted with actual data, but it can be solved analytically, which is useful for the subsequent exposition. For brevity, we refer to this model as the stylized DSGE model in the remainder of this chapter.

The model economy consists of households, intermediate goods producers, final goods producers, a monetary policy authority, and a fiscal authority. Macroeconomic fluctuations are generated by four exogenous processes: a technology growth shock, z_t, a shock that generates shifts in the preference for leisure, ϕ_t, a price markup shock, λ_t, and a monetary policy shock $\epsilon_{R,t}$. We assume that the level of productivity Z_t in the economy is evolving exogenously according to a random walk with drift:

$$\log Z_t = \log \gamma + \log Z_{t-1} + z_t, \quad z_t = \rho_z z_{t-1} + \sigma_z \epsilon_{z,t}. \tag{54}$$

The productivity process Z_t induces a stochastic trend in output X_t and real wages W_t. To facilitate the model solution, it is useful to detrend output and real wages by the level of technology, defining $x_t = X_t/Z_t$ and $w_t = W_t/Z_t$, respectively. In terms of the detrended variables, the model has the following steady state:

$$\bar{x} = x^*, \quad \bar{w} = \overline{lsh} = \frac{1}{1+\lambda}, \quad \bar{\pi} = \pi^*, \quad \bar{R} = \pi^* \frac{\gamma}{\beta}. \tag{55}$$

Here x^* and π^* are free parameters. The latter can be interpreted as the central bank's target inflation rate, whereas the former can in principle be derived from the weight on leisure in the households' utility function. The steady-state real wage $\bar{w}$ is equal to the steady-state labor share $\overline{lsh}$. The parameter λ can be interpreted as the steady-state markup charged by the monopolistically competitive intermediate goods producers, β is the discount factor of the households, and γ is the growth rate of technology. Under the assumption that the production technology is linear in labor and labor is the only

[ae] See Sections 4.1 and 4.5 for how to think about loglinearizations as a first-order perturbations.

factor of production, the steady state labor share equals the steady state of detrended wages. We also assume that all output is consumed, which means that x can be interpreted as aggregate consumption.

8.1.1 Loglinearized Equilibrium Conditions

In terms of log-deviations from the steady state (denoted by ˆ), ie, $\hat{x} = \log(x_t/\bar{x})$, $\hat{w}_t = \log(w_t/\bar{w})$, $\hat{\pi}_t = \log(\pi_t/\bar{\pi})$, and $\hat{R}_t = \log(R_t/\bar{R})$, the equilibrium conditions of the model can be stated as follows. The consumption Euler equation of the households takes the form

$$\hat{x}_t = \mathbb{E}_{t+1}[\hat{x}_{t+1}] - \left(\hat{R}_t - \mathbb{E}[\hat{\pi}_{t+1}]\right) + \mathbb{E}_t[z_{t+1}]. \tag{56}$$

The expected technology growth rate arises because the Euler equation is written in terms of output in deviations from the stochastic trend induced by Z_t. Assuming the absence of nominal wage rigidities, the intratemporal Euler equation for the households leads to the following labor supply equation:

$$\hat{w}_t = (1+\nu)\hat{x}_t + \phi_t, \tag{57}$$

where $\hat{w}_t$ is the real wage, $1/(1+\nu)$ is the Frisch labor supply elasticity, $\hat{x}_t$ is proportional to hours worked, and ϕ_t is an exogenous labor supply shifter

$$\phi_t = \rho_\phi \phi_{t-1} + \sigma_\phi \epsilon_{\phi,t}. \tag{58}$$

We refer to ϕ_t as preference shock.

The intermediate goods producers hire labor from the households and produce differentiated products, indexed by j, using a linear technology of the form $X_t(j) = Z_t L_t(j)$. After detrending and loglinearization around steady-state aggregate output, the production function becomes

$$\hat{x}_t(j) = \hat{L}_t(j). \tag{59}$$

Nominal price rigidity is introduced via the Calvo mechanism. In each period, firm j is unable to reoptimize its nominal price with probability ζ_p. In this case, the firm simply adjusts its price from the previous period by the steady-state inflation rate. With probability $1 - \zeta_p$, the firm can choose its price to maximize the expected sum of future profits. The intermediate goods are purchased and converted into an aggregate good X_t by a collection of perfectly competitive final goods producers using a constant-elasticity-of-substitution aggregator.

The optimality conditions for the two types of firms can be combined into the so-called New Keynesian Phillips curve, which can be expressed as

$$\hat{\pi}_t = \beta \mathbb{E}_t[\hat{\pi}_{t+1}] + \kappa_p(\hat{w}_t + \lambda_t), \quad \kappa_p = \frac{(1-\zeta_p\beta)(1-\zeta_p)}{\zeta_p}, \tag{60}$$

where β is the households' discount factor and λ_t can be interpreted as a price mark-up shock, which exogenously evolves according to

$$\lambda_t = \rho_\lambda \lambda_{t-1} + \sigma_\lambda \epsilon_{\lambda,t}. \tag{61}$$

It is possible to derive an aggregate resource constraint that relates the total amount of labor L_t hired by the intermediate goods producers to the total aggregate output X_t produced in the economy. Based on this aggregate resource constraint, it is possible to compute the labor share of income, which, in terms of deviations from steady state is given by

$$\widehat{lsh}_t = \hat{w}_t. \tag{62}$$

Finally, the central bank sets the nominal interest rate according to the feedback rule

$$\hat{R}_t = \psi \hat{\pi}_t + \sigma_R \epsilon_{R,t} \quad \psi = 1/\beta. \tag{63}$$

We abstract from interest rate smoothing and the fact that central banks typically also react to some measure of real activity, eg, the gap between actual output and potential output. The shock $\epsilon_{R,t}$ is an unanticipated deviation from the systematic part of the interest rate feedback rule and is called a monetary policy shock. We assume that $\psi = 1/\beta$, which ensures the existence of a unique stable solution to the system of linear rational expectations difference equations and, as will become apparent below, simplifies the solution of the model considerably. The fiscal authority determines the level of debt and lump-sum taxes such that the government budget constraint is satisfied.

8.1.2 Model Solution

To solve the model, note that the economic state variables are ϕ_t, λ_t, z_t, and $\epsilon_{R,t}$. Due to the fairly simple loglinear structure of the model, the aggregate laws of motion $\hat{x}(\cdot)$, $\hat{l}sh(\cdot)$, $\hat{\pi}(\cdot)$, and $\hat{R}(\cdot)$ are linear in the states and can be determined sequentially. We first eliminate the nominal interest rate from the consumption Euler equation using (63):

$$\hat{x}_t = \mathbb{E}_{t+1}[\hat{x}_{t+1}] - \left(\frac{1}{\beta}\hat{\pi}_t + \sigma_R \epsilon_{R,t} - \mathbb{E}[\hat{\pi}_{t+1}] \right) + \mathbb{E}_t[z_{t+1}]. \tag{64}$$

Now notice that the New Keynesian Phillips curve can be rewritten as

$$\frac{1}{\beta}\hat{\pi}_t - \mathbb{E}_t[\hat{\pi}_{t+1}] = \frac{\kappa_p}{\beta}((1+\nu)\hat{x}_t + \phi_t + \lambda_t). \tag{65}$$

Here we replaced wages $\hat{w}_t$ with the right-hand side of (57). Substituting (65) into (64) and rearranging terms leads to the following expectational difference equation for output $\hat{x}_t$

$$\hat{x}_t = \psi_p \mathbb{E}_t[\hat{x}_{t+1}] - \frac{\kappa_p \psi_p}{\beta}(\phi_t + \lambda_t) + \psi_p \mathbb{E}_t[z_{t+1}] - \psi_p \sigma_R \epsilon_{R,t}, \tag{66}$$

where $0 \leq \psi_p \leq 1$ is given by

$$\psi_p = \left(1 + \frac{\kappa_p}{\beta}(1+\nu)\right)^{-1}.$$

We now need to find a law of motion for output (and, equivalently, consumption) of the form

$$\hat{x}_t = \hat{x}(\phi_t, \lambda_t, z_t, \epsilon_{R,t}) = x_\phi \phi_t + x_\lambda \lambda_t + x_z z_t + x_{\epsilon_R}\epsilon_{R,t} \tag{67}$$

that solves the functional equation

$$\begin{aligned} &\mathbb{E}_t \mathcal{H}(\hat{x}(\cdot)) \\ &\quad = \mathbb{E}_t \Big[\hat{x}(\phi_t, \lambda_t, z_t, \epsilon_{R,t}) - \psi_p \hat{x}\big(\rho_\phi \phi_t + \sigma_\phi \epsilon_{\phi,t+1}, \rho_\lambda \lambda_t + \sigma_\lambda \epsilon_{\lambda,t+1}, \rho z_t + \sigma_z \epsilon_{z,t+1}, \epsilon_{R,t+1}\big) \\ &\qquad + \frac{\kappa_p \psi_p}{\beta}(\phi_t + \lambda_t) - \psi_p z_{t+1} + \psi_p \sigma_R \epsilon_{R,t}\Big] = 0. \end{aligned} \tag{68}$$

Here, we used the laws of motion of the exogenous shock processes in (54), (58), and (61). Assuming that the innovations ϵ_t are Martingale difference sequences, it can be verified that the coefficients of the linear decision rule are given by

$$x_\phi = -\frac{\kappa_p \psi_p/\beta}{1-\psi_p\rho_\phi}, \quad x_\lambda = -\frac{\kappa_p \psi_p/\beta}{1-\psi_p\rho_\lambda}, \quad x_z = \frac{\rho_z \psi_p}{1-\psi_p\rho_z} z_t, \quad x_{\epsilon_R} = -\psi_p \sigma_R. \tag{69}$$

After having determined the law of motion for output, we now solve for the labor share, inflation, and nominal interest rates. Using (57) and (62) we immediately deduce that the labor share evolves according to

$$\widehat{lsh}_t = \left[1 + (1+\nu)x_\phi\right]\phi_t + (1+\nu)x_\lambda \lambda_t + (1+\nu)x_z z_t + (1+\nu)x_{\epsilon_R}\epsilon_{R,t}. \tag{70}$$

To obtain the law of motion of inflation, we have to solve the following functional equation derived from the New Keynesian Phillips curve (60):

$$\begin{aligned} &\mathbb{E}_t \mathcal{H}(\hat{\pi}(\cdot)) \\ &\quad = \mathbb{E}_t \big[\hat{\pi}(\phi_t, \lambda_t, z_t, \epsilon_{R,t}) - \beta \hat{\pi}\big(\rho_\phi \phi_t + \sigma_\phi \epsilon_{\phi,t+1}, \rho_\lambda \lambda_t + \sigma_\lambda \epsilon_{\lambda,t+1}, \rho z_t + \sigma_z \epsilon_{z,t+1}, \epsilon_{R,t+1}\big) \\ &\qquad - \kappa_p \widehat{lsh}(\phi_t, \lambda_t, z_t, \epsilon_{R,t}) - \kappa_p \lambda_t\big] = 0, \end{aligned} \tag{71}$$

where $\widehat{lsh}(\cdot)$ is given by (70). The solution takes the form

$$\begin{aligned} \hat{\pi}_t = {} & \frac{\kappa_p}{1-\beta\rho_\phi}\left[1 + (1+\nu)x_\phi\right]\phi_t + \frac{\kappa_p}{1-\beta\rho_\lambda}[1 + (1+\nu)x_\lambda]\lambda_t \\ & + \frac{\kappa_p}{(1-\beta\rho_z)}(1+\nu)x_z z_t + \kappa_p(1+\nu)x_{\epsilon_R}\epsilon_{R,t}. \end{aligned} \tag{72}$$

Finally, combining (72) with the monetary policy rule (63) yields the solution for the nominal interest rate

$$\begin{aligned}\hat{R}_t = & \frac{\kappa_p/\beta}{1-\beta\rho_\phi}\left[1+(1+\nu)x_\phi\right]\phi_t + \frac{\kappa_p/\beta}{1-\beta\rho_\lambda}[1+(1+\nu)x_\lambda]\lambda_t \\ & + \frac{\kappa_p/\beta}{1-\beta\rho_z}(1+\nu)x_z z_t + \left[\kappa_p(1+\nu)x_{\epsilon_R}/\beta + \sigma_R\right]\epsilon_{R,t}.\end{aligned} \tag{73}$$

8.1.3 State-Space Representation

To confront the model with data, one has to account for the presence of the model-implied stochastic trend in aggregate output and to add the steady states to all model variables. Measurement equations for output growth, the labor share, net inflation rates and net interest rates take the form

$$\begin{aligned}\log(X_t/X_{t-1}) &= \hat{x}_t - \hat{x}_{t-1} + z_t + \log\gamma \\ \log(lsh_t) &= \widehat{lsh}_t + \log(lsh) \\ \log\pi_t &= \hat{\pi}_t + \log\pi^* \\ \log R_t &= \hat{R}_t + \log(\pi^*\gamma/\beta).\end{aligned} \tag{74}$$

The DSGE model solution has the form of a generic state-space model. Define the $n_s \times 1$ vector of econometric state variables s_t as

$$s_t = [\phi_t, \lambda_t, z_t, \epsilon_{R,t}, \hat{x}_{t-1}]\prime$$

and the vector of DSGE model parameters[af]

$$\theta = [\beta, \gamma, \lambda, \pi^*, \zeta_p, \nu, \rho_\phi, \rho_\lambda, \rho_z, \sigma_\phi, \sigma_\lambda, \sigma_z, \sigma_R]'. \tag{75}$$

We omitted the steady-state output x^* from the list of parameters because it does not affect the law of motion of output growth. Using this notation, we can express the state transition equation as

$$s_t = \Phi_1(\theta)s_{t-1} + \Phi_\epsilon(\theta)\epsilon_t, \tag{76}$$

where the $n_\epsilon \times 1$ vector ϵ_t is defined as $\epsilon_t = [\epsilon_{\phi,t}, \epsilon_{\lambda,t}, \epsilon_{z,t}, \epsilon_{R,t}]'$. The coefficient matrices $\Phi_1(\theta)$ and $\Phi_\epsilon(\theta)$ are determined by (54), (58), (61), the identity $\epsilon_{R,t} = \epsilon_{R,t}$, and a lagged version of (69) to determine $\hat{x}_{t-1}$. If we define the $n_y \times 1$ vector of observables as

$$y_t = M_y'[\log(X_t/X_{t-1}), \log lsh_t, \log\pi_t, \log R_t]', \tag{77}$$

where M_y' is a matrix that selects rows of the vector $[\log(X_t/X_{t-1}), \log lsh_t, \log\pi_t, \log R_t]'$ then the measurement equation can be written as

$$y_t = \Psi_0(\theta) + \Psi_1(\theta)s_t. \tag{78}$$

[af] From now on, we will use θ to denote the parameters of the DSGE model as opposed to the coefficients of a decision rule conditional on a particular set of DSGE model parameters. Also, to reduce clutter, we no longer distinguish vectors and matrices from scalars by using boldfaced symbols.

The coefficient matrices $\Psi_0(\theta)$ and $\Psi_1(\theta)$ can be obtained from (74), the equilibrium law of motion for the detrended model variables given by (69), (70), (72), and (73). They are summarized in Table 4.

The state-space representation of the DSGE model given by (76) and (78) provides the basis for the subsequent econometric analysis. It characterizes the joint distribution of the observables y_t and the state variables s_t conditional on the DSGE model parameters θ

$$p(Y_{1:T}, S_{1:T}|\theta) = \int \left(\prod_{t=1}^{T} p(y_t|s_t,\theta) p(s_t|s_{t-1},\theta) \right) p(s_0|\theta) ds_0, \tag{79}$$

where $Y_{1:t} = \{y_1,\ldots,y_t\}$ and $S_{1:t} = \{s_1,\ldots,s_t\}$. Because the states are (at least partially) unobserved, we will often work with the marginal distribution of the observables defined as

$$p(Y_{1:T}|\theta) = \int p(Y_{1:T}, S_{1:T}|\theta) dS_{1:T}. \tag{80}$$

As a function of θ the density $p(Y_{1:T}|\theta)$ is called the likelihood function. It plays a central role in econometric inference and its evaluation will be discussed in detail in Section 10.

Remark 28 First, it is important to distinguish economic state variables, namely, ϕ_t, λ_t, z_t, and $\epsilon_{R,t}$, that are relevant for the agents' intertemporal optimization problems, from the econometric state variables s_t, which are used to cast the DSGE model solution into the state-space form given by (76) and (78). The economic state variables of our simple model are all exogenous. As we have seen in Section 4.3, the vector of state variables of a richer DSGE model also may include one or more endogenous variables, eg, the capital stock. Second, output growth in the measurement equation could be replaced by the level of output. This would require adding x^* to the parameter vector θ, eliminating $\hat{x}_{t-1}$ from s_t, adding $\log Z_t/\gamma^t$ to s_t, and accounting for the deterministic trend component $(\log\gamma)t$ in log output in the measurement equation. Third, the measurement Eq. (78) could be augmented by measurement errors. Fourth, if a DSGE model is solved with a higher-order perturbation or projection method, then, depending on how exactly the state vector s_t is defined, the state-transition Eq. (76), the measurement Eq. (78), or both are nonlinear.

8.2 Model Implications

Once we specify a distribution for the innovation vector ϵ_t the probability distribution of the DSGE model variables is fully determined. Recall that the innovation standard deviations were absorbed into the definition of the matrix $\Phi_\epsilon(\theta)$ in (76). For the sake of concreteness, we assume that

$$\epsilon_t \sim iidN(0, I), \tag{81}$$

where I denotes the identity matrix. Based on the probabilistic structure of the DSGE model, we can derive a number of implications from the DSGE model that will later

Table 4 System matrices for DSGE model

State-space representation:

$$y_t = \boldsymbol{\Psi}_0(\theta) + \boldsymbol{\Psi}_1(\theta) s_t$$
$$s_t = \boldsymbol{\Phi}_1(\theta) s_{t-1} + \boldsymbol{\Phi}_\epsilon(\theta)\epsilon_t$$

System matrices:

$$\boldsymbol{\Psi}_0(\theta) = M_y' \begin{bmatrix} \log\gamma \\ \log(lsh) \\ \log\pi^* \\ \log(\pi^*\gamma/\beta) \end{bmatrix}, \quad x_\phi = -\frac{\kappa_p\psi_p/\beta}{1-\psi_p\rho_\phi}, \quad x_\lambda = -\frac{\kappa_p\psi_p/\beta}{1-\psi_p\rho_\lambda}, \quad x_z = \frac{\rho_z\psi_p}{1-\psi_p\rho_z}, \quad x_{\epsilon_R} = -\psi_p\sigma_R$$

$$\boldsymbol{\Psi}_1(\theta) = M_y' \begin{bmatrix} x_\phi & x_\lambda & x_z + 1 & x_{\epsilon_R} & -1 \\ 1 + (1+\nu)x_\phi & (1+\nu)x_\lambda & (1+\nu)x_z & (1+\nu)x_{\epsilon_R} & 0 \\ \frac{\kappa_p}{1-\beta\rho_\phi}(1+(1+\nu)x_\phi) & \frac{\kappa_p}{1-\beta\rho_\lambda}(1+(1+\nu)x_\lambda) & \frac{\kappa_p}{1-\beta\rho_z}(1+\nu)x_z & +\kappa_p(1+\nu)x_{\epsilon_R} & 0 \\ \frac{\kappa_p/\beta}{1-\beta\rho_\phi}(1+(1+\nu)x_\phi) & \frac{\kappa_p/\beta}{1-\beta\rho_\lambda}(1+(1+\nu)x_\lambda) & \frac{\kappa_p/\beta}{1-\beta\rho_z}(1+\nu)x_z & (\kappa_p(1+\nu)x_{\epsilon_R}/\beta + \sigma_R) & 0 \end{bmatrix}$$

$$\boldsymbol{\Phi}_1(\theta) = \begin{bmatrix} \rho_\phi & 0 & 0 & 0 & 0 \\ 0 & \rho_\lambda & 0 & 0 & 0 \\ 0 & 0 & \rho_z & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ x_\phi & x_\lambda & x_z & x_{\epsilon_R} & 0 \end{bmatrix}, \quad \boldsymbol{\Phi}_\epsilon(\theta) = \begin{bmatrix} \sigma_\phi & 0 & 0 & 0 \\ 0 & \sigma_\lambda & 0 & 0 \\ 0 & 0 & \sigma_z & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

M_y' is an $n_y \times 4$ selection matrix that selects rows of $\boldsymbol{\Psi}_0$ and $\boldsymbol{\Psi}_1$.

Table 5 Parameter values for stylized DSGE model

Parameter	Value	Parameter	Value
β	$1/1.01$	γ	$\exp(0.005)$
λ	0.15	π^*	$\exp(0.005)$
ζ_p	0.65	ν	0
ρ_ϕ	0.94	ρ_λ	0.88
ρ_z	0.13		
σ_ϕ	0.01	σ_λ	0.01
σ_z	0.01	σ_R	0.01

be used to construct estimators of the parameter vector θ and evaluate the fit of the model. For now, we fix θ to the values listed in Table 5.

8.2.1 Autocovariances and Forecast Error Variances

DSGE models are widely used for business cycle analysis. In this regard, the model-implied variances, autocorrelations, and cross-correlations are important objects. For linear DSGE models it is straightforward to compute the autocovariance function from the state-space representation given by (76) and (78).[ag] Using the notation

$$\Gamma_{yy}(h) = \mathbb{E}[y_t y_{t-h}], \quad \Gamma_{ss}(h) = \mathbb{E}[s_t s_{t-h}], \quad \text{and} \quad \Gamma_{ys}(h) = \mathbb{E}[y_t s'_{t-h}]$$

and the assumption that $\mathbb{E}[\epsilon_t \epsilon_t']= I$, we can express the autocovariance matrix of s_t as the solution to the following Lyapunov equation:[ah]

$$\Gamma_{ss}(0) = \Phi_1 \Gamma_{ss}(0) \Phi_1' + \Phi_\epsilon \Phi_\epsilon'. \tag{82}$$

Once the covariance matrix of s_t has been determined, it is straightforward to compute the autocovariance matrices for $h \neq 0$ according to

$$\Gamma_{ss}(h) = \Phi_1^h \Gamma_{ss}(0). \tag{83}$$

Finally, using the measurement Eq. (78), we deduce that

$$\Gamma_{yy}(h) = \Psi_1 \Gamma_{ss}(h) \Psi_1', \quad \Gamma_{ys}(h) = \Psi_1 \Gamma_{ss}(h). \tag{84}$$

[ag] For the parameters in Table 5, the largest (in absolute value) eigenvalue of the matrix $\Phi_1(\theta)$ in (76) is less than one, which implies that the VAR(1) law of motion for s_t is covariance stationary.

[ah] Efficient numerical routines to solve Lyapunov equations are readily available in many software packages, eg, the function *dylap* in MATLAB.

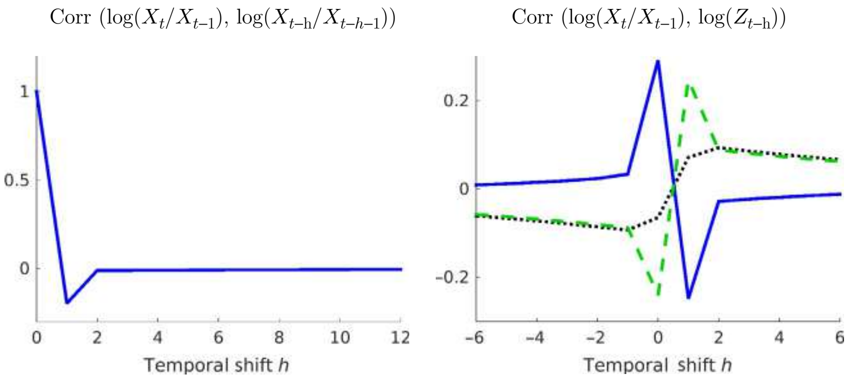


Fig. 16 Autocorrelations. *Notes*: *Right panel*: correlations of output growth with labor share (*solid*), inflation (*dotted*), and interest rates (*dashed*).

Correlations can be easily computed by normalizing the entries of the autocovariance matrices using the respective standard deviations. Fig. 16 shows the model-implied autocorrelation function of output growth and the cross-correlations of output growth with the labor share, inflation, and interest rates as a function of the temporal shift h.

The law of motion for the state vector s_t can also be expressed as the infinite-order vector moving average (MA) process

$$y_t = \Psi_0 + \Psi_1 \sum_{s=0}^{\infty} \Phi_1^s \Phi_\epsilon \epsilon_{t-s}. \tag{85}$$

Based on the moving average representation, it is straightforward to compute the h-step-ahead forecast error, which is given by

$$e_{t|t-h} = y_t - \mathbb{E}_{t-h}[y_t] = \Psi_1 \sum_{s=0}^{h-1} \Phi_1^s \Phi_\epsilon \epsilon_{t-s}. \tag{86}$$

The h-step-ahead forecast error covariance matrix is given by

$$\mathbb{E}[e_{t|t-h} e'_{t|t-h}] = \Psi_1 \left(\sum_{s=0}^{h-1} \Phi_1^s \Phi_\epsilon \Phi_\epsilon' \Phi_1^{s'} \right) \Psi_1' \quad \text{with} \quad \lim_{h\to\infty} \mathbb{E}[e_{t|t-h} e'_{t|t-h}] = \Gamma_{ss}(0). \tag{87}$$

Under the assumption that $\mathbb{E}[\epsilon_t \epsilon_t'] = I$, it is possible to decompose the forecast error covariance matrix as follows. Let $I^{(j)}$ be defined by setting all but the j-th diagonal element of the identity matrix I to zero. Then we can write

$$I = \sum_{j=1}^{n_\epsilon} I^{(j)}. \tag{88}$$

Moreover, we can express the contribution of shock j to the forecast error for y_t as

$$e^{(j)}_{t|t-h} = \Psi_1 \sum_{s=0}^{h-1} \Phi_1^s \Phi_\epsilon I^{(j)} \epsilon_{t-s}. \tag{89}$$

Thus, the contribution of shock j to the forecast error variance of observation $y_{i,t}$ is given by the ratio

$$\text{FEVD}(i,j,h) = \frac{\left[\Psi_1 \left(\sum_{s=0}^{h-1} \Phi_1^s \Phi_\epsilon I^{(j)} \Phi_\epsilon' \Phi_1^{s'}\right) \Psi_1'\right]_{ii}}{\left[\Psi_1 \left(\sum_{s=0}^{h-1} \Phi_1^s \Phi_\epsilon \Phi_\epsilon' \Phi_1^{s'}\right) \Psi_1'\right]_{ii}}, \tag{90}$$

where $[A]_{ij}$ denotes element (i,j) of a matrix A. Fig. 17 shows the contribution of the four shocks to the forecast error variance of output growth, the labor share, inflation, and interest rates in the stylized DSGE model. Given the choice of parameters θ in

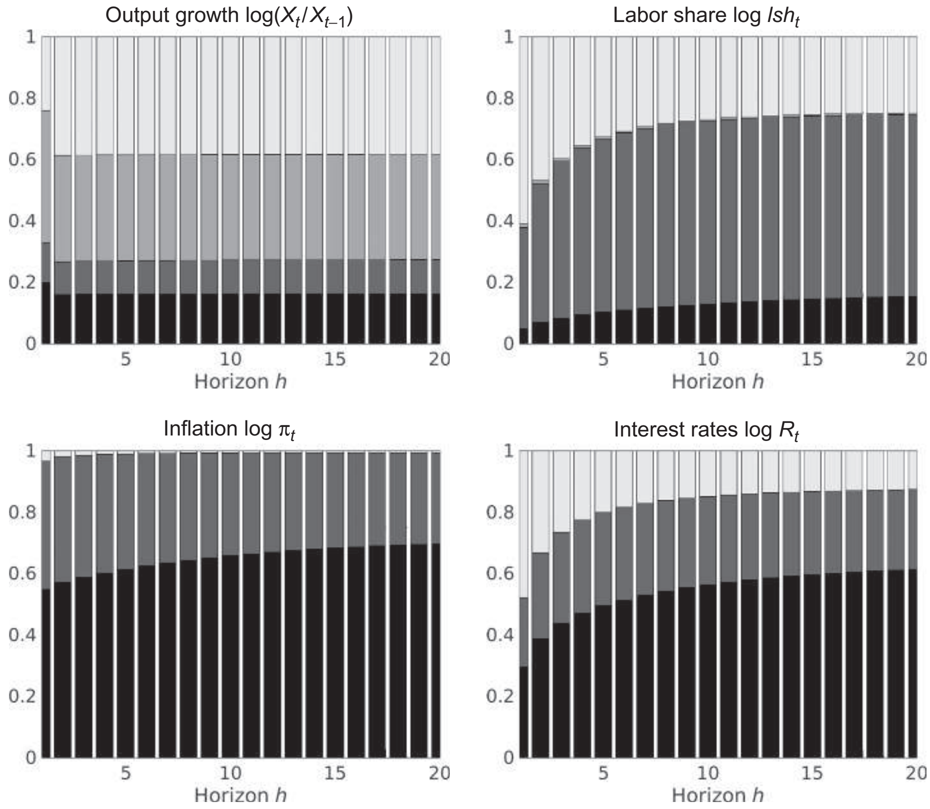


Fig. 17 Forecast error variance decomposition. *Notes*: The stacked bar plots represent the cumulative forecast error variance decomposition. The bars, from darkest to lightest, represent the contributions of ϕ_t, λ_t, z_t, and $\varepsilon_{R,t}$.

Table 5, most of the variation in output growth is due to the technology and the monetary policy shocks. The labor share fluctuations are dominated by the mark-up shock λ_t, in particular in the long run. Inflation and interest rate movements are strongly influenced by the preference shock ϕ_t and the mark-up shock λ_t.

8.2.2 Spectrum

Instead of studying DSGE model implications over different forecasting horizons, one can also consider different frequency bands. There is a long tradition of frequency domain analysis in the time series literature. A classic reference is Priestley (1981). We start with a brief discussion of the linear cyclical model, which will be useful for interpreting some of the formulas presented subsequently. Suppose that y_t is a scalar time series that follows the process

$$y_t = 2\sum_{j=1}^{m} a_j\big(\cos\theta_j \cos(\omega_j t) - \sin\theta_j \sin(\omega_j t)\big), \tag{91}$$

where $\theta_j \sim iidU[-\pi,\pi]$ and $0 \le \omega_j \le \omega_{j+1} \le \pi$. The random variables θ_j cause a phase shift of the cycle and are assumed to be determined in the infinite past. In a nutshell, the model in (91) expresses the variable y_t as the sum of sine and cosine waves that differ in their frequency. The interpretation of the ω_j's depends on the length of the period t. Suppose the model is designed for quarterly data and $\omega_j = (2\pi)/32$. This means that it takes 32 periods to complete the cycle. Business cycles typically comprise cycles that have a duration of 8–32 quarters, which would correspond to $\omega_j \in [0.196, 0.785]$ for quarterly t.

Using Euler's formula, we rewrite the cyclical model in terms of an exponential function:

$$y_t = \sum_{j=-m}^{m} A(\omega_j)e^{i\omega_j t}, \tag{92}$$

where $\omega_{-j} = -\omega_j$, $i = \sqrt{-1}$, and

$$A(\omega_j) = \begin{cases} a_j\big(\cos\theta_{|j|} + i\sin\theta_{|j|}\big) & \text{if } j > 0 \\ a_j\big(\cos\theta_{|j|} - i\sin\theta_{|j|}\big) & \text{if } j < 0 \end{cases} \tag{93}$$

It can be verified that expressions (91) and (92) are identical. The function $A(\omega_j)$ captures the amplitude of cycles with frequency ω_j.

The spectral distribution function of y_t on the interval $\omega \in (-\pi,\pi]$ is defined as

$$F_{yy}(\omega) = \sum_{j=-m}^{m} \mathbb{E}[A(\omega_j)\overline{A(\omega_j)}]\mathbb{I}\{\omega_j \le \omega\}, \tag{94}$$

where $\mathbb{I}\{\omega_j \le \omega\}$ denotes the indicator function that is one if $\omega_j \le \omega$ and $\bar{z} = x - iy$ is the complex conjugate of $z = x + iy$. If $F_{yy}(\omega)$ is differentiable with respect to ω, then we can define the spectral density function as

$$f_{yy}(\omega) = dF_{yy}(\omega)d\omega. \tag{95}$$

If a process has a spectral density function $f_{yy}(\omega)$, then the covariances can be expressed as

$$\Gamma_{yy}(h) = \int_{(-\pi,\pi]} e^{ih\omega} f_{yy}(\omega)d\omega. \tag{96}$$

For the linear cyclical model in (91) the autovariances are given by

$$\Gamma_{yy}(h) = \sum_{j=-m}^{m} \mathbb{E}[A(\omega_j)\overline{A(\omega_j)}]e^{i\omega_j h} = \sum_{j=-m}^{m} a_j^2 e^{i\omega_j h}. \tag{97}$$

The spectral density uniquely determines the entire sequence of autocovariances. Moreover, the converse is also true. The spectral density can be obtained from the autocovariances of y_t as follows:

$$f_{yy}(\omega) = \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} \Gamma_{yy}(h)e^{-i\omega h}. \tag{98}$$

The formulas (96) and (98) imply that the spectral density function and the sequence of autocovariances contain the same information. Their validity is not restricted to the linear cyclical model and they extend to vector-valued y_t's. Recall that for the DSGE model defined by the state-space system (76) and (78) the autocovariance function for the state vector s_t was defined as $\Gamma_{ss}(h) = \Phi_1^h \Gamma_{ss}(0)$. Thus,

$$\begin{aligned} f_{ss}(\omega) &= \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} \Phi_1^h \Gamma_{ss}(0)e^{-i\omega h} \\ &= \frac{1}{2\pi}\left(I - \Phi_1' e^{i\omega}\right)^{-1} \Phi_\epsilon \Phi_\epsilon' \left(I - \Phi_1 e^{-i\omega}\right)^{-1}. \end{aligned} \tag{99}$$

The contribution of shock j to the spectral density is given by

$$f_{ss}^{(j)}(\omega) = \frac{1}{2\pi}\left(I - \Phi_1' e^{i\omega}\right)^{-1} \Phi_\epsilon \mathcal{I}^{(j)} \Phi_\epsilon' \left(I - \Phi_1 e^{-i\omega}\right)^{-1}. \tag{100}$$

The spectral density for the observables y_t (and the contribution of shock j to the spectral density) can be easily obtained as

$$f_{yy}(\omega) = \Psi_1 f_{ss}(\omega)\Psi_1' \quad \text{and} \quad f_{yy}^{(j)}(\omega) = \Psi_1 f_{ss}^{(j)}(\omega)\Psi_1'. \tag{101}$$

Fig. 18 depicts the spectral density functions for output growth, the labor share, inflation, and interest rates for the stylized DSGE model conditional on the parameters in Table 5. Note that $f_{yy}(\omega)$ is a matrix valued function. The four panels correspond to the diagonal elements of this function, providing a summary of the univariate autocovariance properties of the four series. Each panel stacks the contributions of the four shocks to the spectral densities. Because the shocks are independent and evolve according to AR(1) processes, the spectral density peaks at the origin and then decays as the frequency increases.

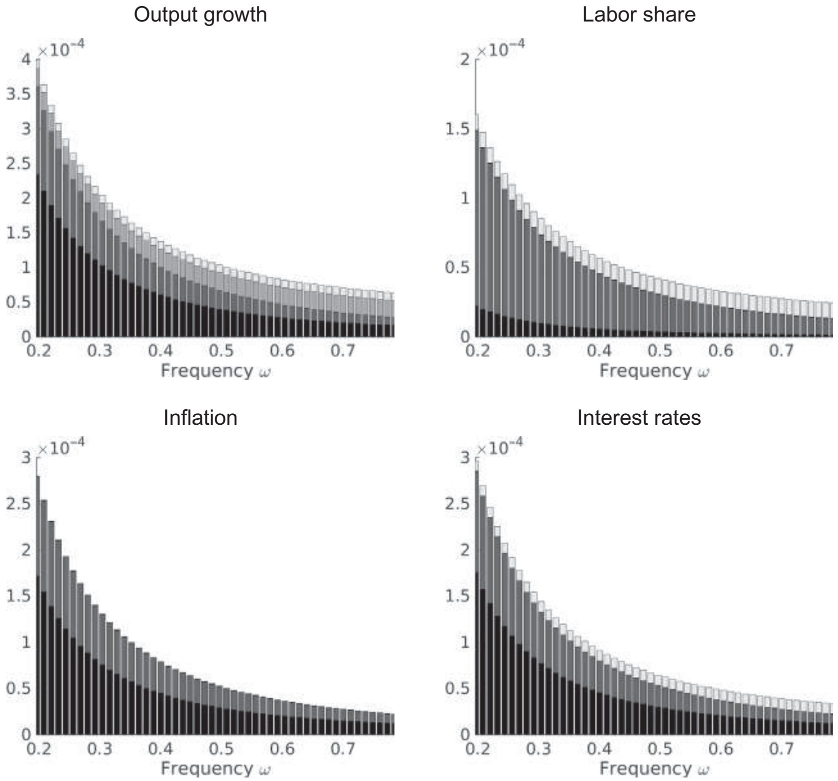


Fig. 18 Spectral decomposition. *Notes*: The stacked bar plots depict cumulative spectral densities. The bars, from darkest to lightest, represent the contributions of ϕ_t, λ_t, z_t, and $\varepsilon_{R,t}$.

8.2.3 Impulse Response Functions

An important tool for studying the dynamic effects of exogenous shocks are impulse response functions (IRFs). Formally, impulse responses in a DSGE model can be defined as the difference between two conditional expectations:

$$\mathrm{IRF}(i,j,h|s_{t-1}) = \mathbb{E}\big[y_{i,t+h} \mid s_{t-1}, \epsilon_{j,t} = 1\big] - \mathbb{E}[y_{i,t+h}| \, s_{t-1}]. \tag{102}$$

Both expectations are conditional on the initial state s_{t-1} and integrate over current and future realizations of the shocks ϵ_t. However, the first term also conditions on $\epsilon_{j,t} = 1$, whereas the second term averages of $\epsilon_{j,t}$. In a linearized DSGE model with a state-space representation of the form (76) and (78), we can use the linearity and the property that $\mathbb{E}[\epsilon_{t+h}|s_{t-1}] = 0$ for $h = 0,1,\ldots$ to deduce that

$$\mathrm{IRF}(.,j,h) = \Psi_1 \frac{\partial}{\partial \epsilon_{j,t}} s_{t+h} = \Psi_1 \Phi_1^h [\Phi_\epsilon]_{.j}, \tag{103}$$

where $[A]_{.j}$ is the j-th column of a matrix A. We dropped s_{t-1} from the conditioning set to simplify the notation.

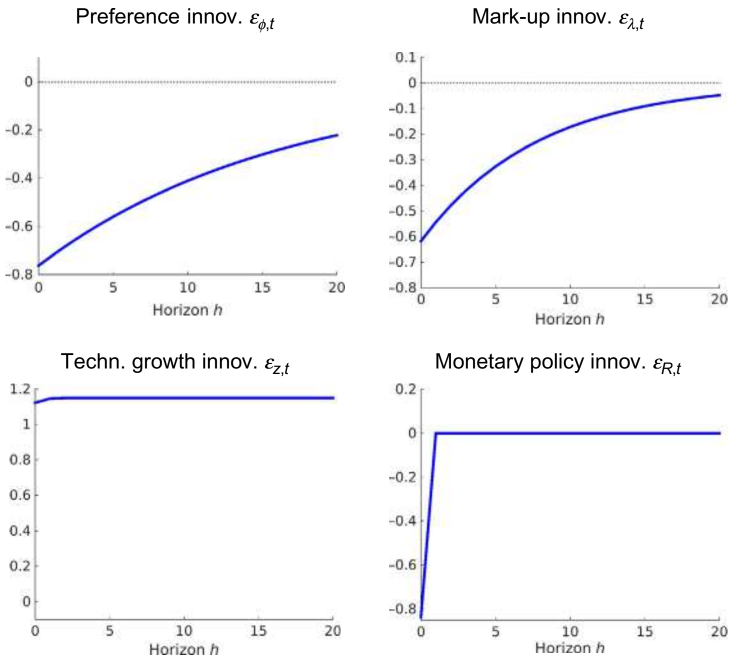


Fig. 19 Impulse responses of log output $100\log(X_{t+h}/X_t)$.

Fig. 19 depicts the impulse response functions for the stylized DSGE model of log output to the four structural shocks, which can be easily obtained from (69) and the laws of motion of the exogenous shock processes. The preference and mark-up shocks lower output upon impact. Subsequently, output reverts back to its steady state. The speed of the reversion is determined by the autoregressive coefficient associated with the exogenous shock process. The technology growth shock raises the log level of output permanently, whereas a monetary policy shock has only a one-period effect on output.

8.2.4 Conditional Moment Restrictions

The intertemporal optimality conditions take the form of conditional moment restrictions. For instance, rearranging the terms in the New Keynesian Phillips (60) curve, we can write

$$\mathbb{E}_{t-1}\left[\hat{\pi}_{t-1}-\beta\hat{\pi}_t-\kappa_p\left(\widehat{lsh}_{t-1}+\lambda_{t-1}\right)\right]=0. \tag{104}$$

The conditional moment condition can be converted into a vector of unconditional moment conditions as follows. Let $\mathcal{F}_t$ denote the sigma algebra generated by the infinite histories of $\{y_\tau, s_\tau, \epsilon_\tau\}_{\tau=-\infty}^{t}$ and let $\tilde{Z}_t$ be a vector of random variables that is measurable with respect to $\mathcal{F}_t$, meaning that its value is determined based on information on current and past (y_t, s_t, ϵ_t). Then for every such vector $\tilde{Z}_{t-1}$,

$$\begin{aligned}&\mathbb{E}\left[\widetilde{Z}_{t-1}\left(\hat{\pi}_{t-1}-\beta\hat{\pi}_t-\kappa_p(\widehat{lsh}_{t-1}+\lambda_{t-1})\right)\right]\\&\quad=\mathbb{E}\left[\widetilde{Z}_{t-1}\mathbb{E}_{t-1}\left[\hat{\pi}_{t-1}-\beta\hat{\pi}_t-\kappa_p(\widehat{lsh}_{t-1}+\lambda_{t-1})\right]\right]=0,\end{aligned} \tag{105}$$

where $\mathbb{E}_{t-1}[\cdot]=\mathbb{E}[\cdot|\mathcal{F}_{t-1}]$.

The moment conditions derived from the New Keynesian Phillips curve involve the latent price mark-up shock λ_t, which will cause difficulties if one tries to use (105) in an estimation objective function. Now consider the consumption Euler equation (56) instead. Recall that the measurement equations imply that

$$\hat{x}_t-\hat{x}_{t-1}+z_t=\log X_t-\log X_{t-1}-\log\gamma \quad\text{and}\quad \hat{R}_t=\log R_t-\log(\pi^*\gamma/\beta).$$

Thus, we can write

$$\mathbb{E}_{t-1}[-\log(X_t/X_{t-1})+\log R_{t-1}-\log\pi_t-\log(1/\beta)]=0. \tag{106}$$

The terms γ and $\log\pi^*$ that appear in the steady-state formulas for the nominal interest rate and inflation cancel and the conditional moment condition only depends on observables and the model parameters, but not on latent variables. Finally, as long as the monetary policy shock satisfies the martingale difference sequences property $\mathbb{E}_{t-1}[\epsilon_{R,t}]=0$, we obtain from the monetary policy rule the condition that

$$\mathbb{E}_{t-1}[\log R_t-\log(\gamma/\beta)-\psi\log\pi_t-(1-\psi)\log\pi^*]=0. \tag{107}$$

Both (106) and (107) can be converted into an unconditional moment condition using an $\mathcal{F}_{t-1}$ measurable random vector $\mathcal{Z}_{t-1}$ as in (105).

8.2.5 Analytical Calculation of Moments vs Simulation Approximations

As previously shown, formulas for autocovariance functions, spectra, and impulse response functions for a linearized DSGE model can be derived analytically from the state-space representation. These analytical expressions can then be numerically evaluated for different vectors of parameter values θ. For DSGE models solved with perturbation methods, there are also analytical formulas available that exploit a conditionally linear structure of some perturbation solutions; see Andreasen et al. (2013). For a general nonlinear DSGE model, the implied moments have to be computed using Monte Carlo simulation. For instance, let $Y^*_{1:T}$ denote a sequence of observations simulated from the state-space representation of the DSGE model by drawing an initial state vector s_0 and innovations ϵ_t from their model-implied distributions, then

$$\frac{1}{T}\sum_{t=1}^{T}y_t^*\xrightarrow{a.s.}\mathbb{E}[y_t], \tag{108}$$

provided that the DSGE model-implied y_t is strictly stationary and ergodic.[ai] The downside of Monte Carlo approximations is that they are associated with a simulation error. We will come back to this problem in Section 11.2, when we use simulation approximations of moments to construct estimators of θ.

8.3 Empirical Analogs

We now examine sample analogs of the population moments derived from the state-space representation of the DSGE model using US data. The time series were downloaded from the FRED database maintained by the Federal Reserve Bank of St. Louis and we report the series labels in parentheses. For real aggregate output, we use quarterly, seasonally adjusted GDP at the annual rate that has been pegged to 2009 dollars (GDPC96). We turn GDP into growth rates by taking logs and then differencing. The labor share is defined as Compensation of Employees (COE) divided by nominal GDP (GDP). Both series are quarterly and seasonally adjusted at the annual rate. We use the log labor share as the observable. Inflation rates are computed from the implicit price deflator (GDPDEF) by taking log differences. Lastly, for the interest rate, we use the Effective Federal Funds Rate (FEDFUNDS), which is monthly, and not seasonally adjusted. Quarterly interest rates are obtained by taking averages of the monthly rates. Throughout this section we focus on the post-Great Moderation and pre-Great Recession period and restrict our sample from 1984:Q1 to 2007:Q4.

8.3.1 *Autocovariances*

The sample analog of the population autocovariance $\Gamma_{yy}(h)$ is defined as

$$\hat{\Gamma}_{yy}(h) = \frac{1}{T}\sum_{t=h}^{T}(y_t - \hat{\mu}_y)(y_{t-h} - \hat{\mu}_y)', \quad \text{where} \quad \hat{\mu}_y = \frac{1}{T}\sum_{t=1}^{T} y_t. \tag{109}$$

Under suitable regularity conditions, eg, covariance stationarity of the vector process y_t, a sufficiently fast decay of the serial correlation in y_t, and some bounds on higher-order moments of y_t, the sample autocovariance $\hat{\Gamma}_{yy}(h)$ converges to the population autocovariance $\Gamma_{yy}(h)$, satisfying a strong law of large numbers (SLLN) and a central limit theorem (CLT).

If the object of interest is a sequence of autocovariance matrices, then it might be more efficient to first estimate an auxiliary model and then convert the parameter estimates of the auxiliary model into estimates of the autocovariance sequence. A natural class of auxiliary models is provided by linear vector autoregressions (VARs). For illustrative purposes consider the following VAR(1):

[ai] A sequence of random variables X_T converges to a limit random variable X almost surely (a.s.) if the set of trajectories for which $X_T \not\to X$ has probability zero.

$$y_t = \Phi_1 y_{t-1} + \Phi_0 + u_t, \quad u_t \sim iid(0, \Sigma). \tag{110}$$

The OLS estimator of Φ_1 can be approximated by

$$\hat{\Phi}_1 = \hat{\Gamma}_{yy}(1)\hat{\Gamma}_{yy}^{-1}(0) + O_p(T^{-1}), \quad \hat{\Sigma} = \hat{\Gamma}_{yy}(0) - \hat{\Gamma}_{yy}(1)\hat{\Gamma}_{yy}^{-1}(0)\hat{\Gamma}'_{yy}(1) + O_p(T^{-1}) \tag{111}$$

The $O_p(T^{-1})$ terms arise because the range of the summations in the definition of the sample autocovariances in (109) and the definition of the OLS estimator are not exactly the same.[aj] Suppose that now we plug the OLS estimator into the autocovariance formulas associated with the VAR(1) (see (82) and (83)), then:

$$\hat{\Gamma}_{yy}^V(0) = \hat{\Gamma}_{yy}(0) + O_p(T^{-1}), \quad \hat{\Gamma}_{yy}^V(h) = \left(\hat{\Gamma}_{yy}(1)\hat{\Gamma}_{yy}^{-1}(0)\right)^h \hat{\Gamma}_{yy}(0) + O_p(T^{-1}). \tag{112}$$

Note that for $h = 0,1$ we obtain $\hat{\Gamma}_{yy}^V(1) = \hat{\Gamma}_{yy}(1) + O_p(T^{-1})$. For $h > 1$ the VAR(1) plug-in estimate of the autocovariance matrix differs from the sample autocovariance matrix. If the actual time series are well approximated by a VAR(1), then the plug-in autocovariance estimate tends to be more efficient than the sample autocovariance estimate $\hat{\Gamma}_{yy}(h)$; see, for instance, Schorfheide (2005b).

In practice, a VAR(1) may be insufficient to capture the dynamics of a time series y_t. In this case the autocovariances can be obtained from a VAR(p)

$$y_t = \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + \Phi_0 + u_t, \quad u_t \sim iid(0, \Sigma). \tag{113}$$

The appropriate lag length p can be determined with a model selection criterion, eg, the Schwarz (1978) criterion, which is often called the Bayesian information criterion (BIC). The notationally easiest way (but not the computationally fastest way) is to rewrite the VAR(p) in companion form. This entails expressing the law of motion for the stacked vector $\tilde{y}_t = [y'_t, y'_{t-1}, \ldots, y'_{t-p+1}]$ as VAR(1):

$$\tilde{y}_t = \tilde{\Phi}_1 \tilde{y}_{t-1} + \tilde{\Phi}_0 + \tilde{u}_t, \quad \tilde{u}_t \sim iid(0, \tilde{\Sigma}), \tag{114}$$

where

$$\tilde{\Phi}_1 = \begin{bmatrix} \Phi_1 & \ldots & \Phi_{p-1} & \Phi_p \\ I_{n\times n} & \ldots & 0_{n\times n} & 0_{n\times n} \\ \vdots & \ddots & \vdots & \vdots \\ 0_{n\times n} & \ldots & I_{n\times n} & 0_{n\times n} \end{bmatrix}, \quad \tilde{\Phi}_0 = \begin{bmatrix} \Phi_0 \\ 0_{n(p-1)\times 1} \end{bmatrix},$$

$$\tilde{\epsilon}_t = \begin{bmatrix} \epsilon_t \\ 0_{n(p-1)\times 1} \end{bmatrix}, \quad \tilde{\Sigma} = \begin{bmatrix} \Sigma & 0_{n\times n(p-1)} \\ 0_{n(p-1)\times n} & 0_{n(p-1)\times n(p-1)} \end{bmatrix}.$$

[aj] We say that a sequence of random variables is $O_p(T^{-1})$ if TX_T is stochastically bounded as $T \to \infty$.

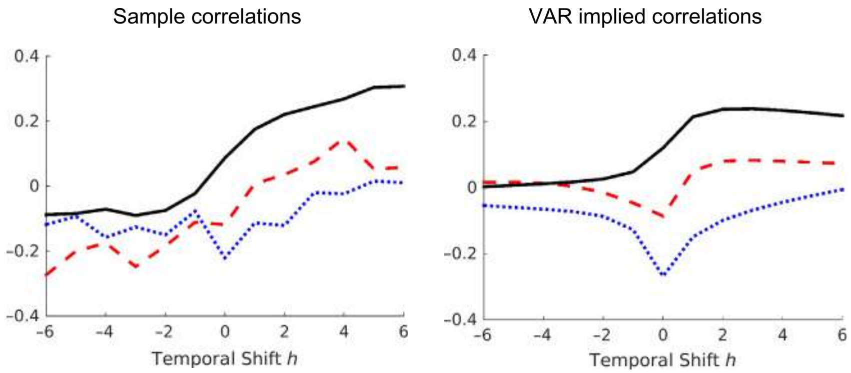


Fig. 20 Empirical cross-correlations $\text{Corr}(\log(X_t/X_{t-1}), \log Z_{t-h})$. *Notes*: Each plot shows the correlation of output growth $\log(X_t/X_{t-1})$ with interest rates (*solid*), inflation (*dashed*), and the labor share (*dotted*), respectively. *Left panel*: correlation functions are computed from sample autocovariance matrices $\hat{\Gamma}_{yy}(h)$. *Right panel*: correlation functions are computed from estimated VAR(1).

The autocovariances for $\widetilde{y}_t$ are then obtained by adjusting the VAR(1) formulas (112) to $\widetilde{y}_t$ and reading off the desired submatrices that correspond to the autocovariance matrices for y_t using the selection matrix $M' = [I_n, 0_{n\times n(p-1)}]$ such that $y_t = M'\widetilde{y}_t$.

We estimate a VAR for output growth, labor share, inflation, and interest rates. The lag length $p = 1$ is determined by the BIC. The left panel of Fig. 20 shows sample cross-correlations (obtained from $\hat{\Gamma}_{yy}(h)$ in (109)) between output growth and leads and lags of the labor share, inflation, and interest rates, respectively. The right panel depicts correlation functions derived from the estimated VAR(1). The two sets of correlation functions are qualitatively similar but quantitatively different. Because the VAR model is more parsimonious, the VAR-implied correlation functions are smoother.

8.3.2 Spectrum

An intuitively plausible estimate of the spectrum is the sample periodogram, defined as

$$\hat{f}_{yy}(\omega) = \frac{1}{2\pi}\sum_{h=-T+1}^{T-1}\hat{\Gamma}_{yy}(h)e^{-i\omega h} = \frac{1}{2\pi}\left(\hat{\Gamma}_{yy}(0) + \sum_{h=1}^{T-1}(\hat{\Gamma}_{yy}(h) + \hat{\Gamma}_{yy}(h)')\cos\omega h\right). \quad (115)$$

While the sample periodogram is an asymptotically unbiased estimator of the population spectral density, it is inconsistent because its variance does not vanish as the sample size $T \longrightarrow \infty$. A consistent estimator can be obtained by smoothing the sample periodogram across adjacent frequencies. Define the fundamental frequencies

$$\omega_j = j\frac{2\pi}{T}, \quad j = 1, \ldots, (T-1)/2$$

and let $K(x)$ denote a kernel function with the property that $\int K(x)dx = 1$. A smoothed periodogram can be defined as

$$\bar{f}_{yy}(\omega) = \frac{\pi}{\lambda(T-1)/2} \sum_{j=1}^{(T-1)/2} K\left(\frac{\omega_j - \omega}{\lambda}\right) \hat{f}_{yy}(\omega_j). \tag{116}$$

An example of a simple kernel function is

$$K\left(\frac{\omega_j - \omega}{\lambda}\right)\hat{f}_{yy}(\omega_j) = \mathbb{I}\left\{-\frac{1}{2} < \frac{\omega_j - \omega}{\lambda} < \frac{1}{2}\right\} = \mathbb{I}\{\omega_j \in B(\omega|\lambda)\},$$

where $B(\omega|\lambda)$ is a frequency band. The smoothed periodogram estimator $\bar{f}_{yy}(\omega)$ is consistent, provided that the bandwidth shrinks to zero, that is, $\lambda \longrightarrow 0$ as $T \longrightarrow \infty$, and the number of ω_j's in the band, given by $\lambda T(2\pi)$, tends to infinity. In the empirical application below we use a Gaussian kernel, meaning that $K(x)$ equals the probability density function of a standard normal random variable.

An estimate of the spectral density can also be obtained indirectly through the estimation of the VAR(p) in (113). Define

$$\Phi = [\Phi_1, \ldots, \Phi_p, \Phi_0]' \quad \text{and} \quad M(z) = [Iz, \ldots, Iz^p],$$

and let $\hat{\Phi}$ be an estimator of Φ. Then a VAR(p) plug-in estimator of the spectral density is given by

$$\hat{f}^{V}_{yy}(\omega) = \frac{1}{2\pi}[I - \hat{\Phi}'M'(e^{-i\omega})]^{-1}\hat{\Sigma}[I - M(e^{-i\omega})\hat{\Phi}]^{-1}. \tag{117}$$

This formula generalizes the VAR(1) spectral density in (99) to a spectral density for a VAR(p).

Estimates of the spectral densities of output growth, log labor share, inflation, and interest rates are reported in Fig. 21. The shaded areas highlight the business cycle frequencies. Because the autocorrelation of output growth is close to zero, the spectral density is fairly flat. The other three series have more spectral mass at the low frequency, which is a reflection of the higher persistence. The labor share has a pronounced hump-shaped spectral density, whereas the other spectral densities of interest and inflation rates are monotonically decreasing in the frequency ω. The smoothness of the periodogram estimates $\bar{f}_{yy}(\omega)$ depends on the choice of the bandwidth. The figure is based on a Gaussian kernel with standard deviation 0.15, which, roughly speaking, averages the sample periodogram over a frequency band of 0.6. While the shapes of the smoothed periodograms and the VAR-based spectral estimates are qualitatively similar, the spectral density is lower according to the estimated VAR.

8.3.3 Impulse Response Functions

The VAR(p) in (113) is a so-called reduced-form VAR because the innovations u_t do not have a specific structural interpretation—they are simply one-step-ahead forecast errors. The impulse responses that we constructed for the DSGE model are responses to

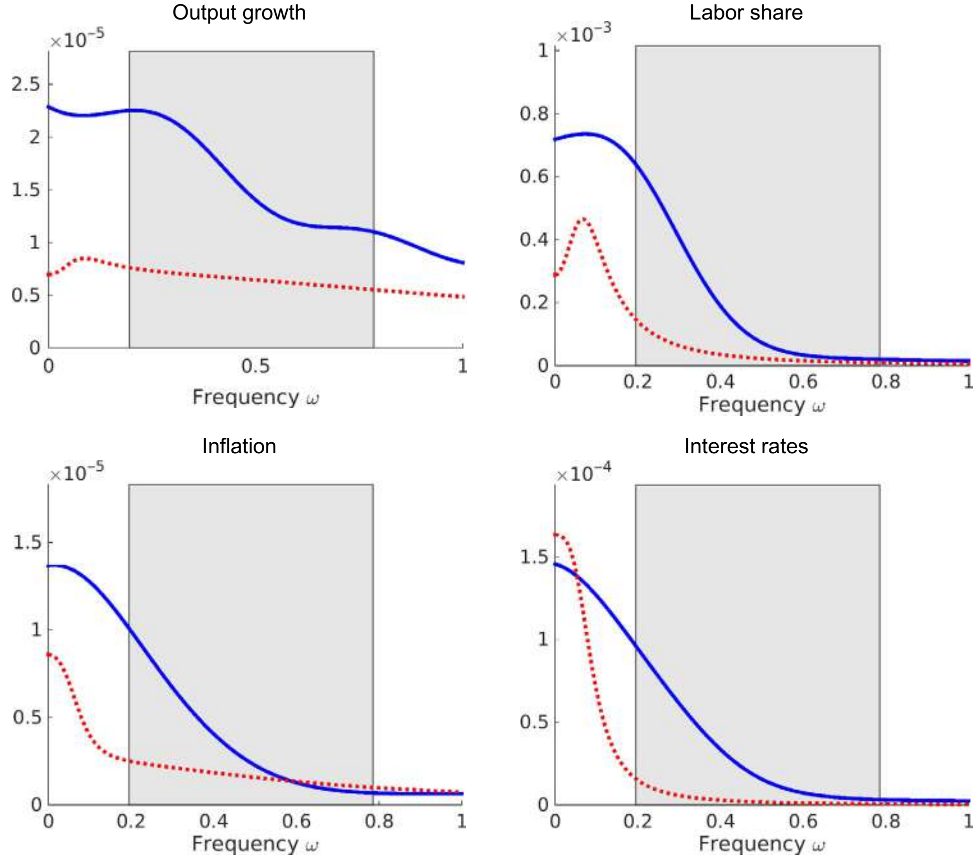


Fig. 21 Empirical spectrum. *Notes*: The *dotted lines* are spectra computed from an estimated VAR(1); the *solid lines* are smoothed periodograms based on a Gaussian kernel with standard deviation 0.15. The *shaded areas* indicate business cycle frequencies (0.196–0.785).

innovations in the structural shock innovations that contribute to the forecast error for several observables simultaneously. In order to connect VAR-based impulse responses to DSGE model-based responses, one has to link the one-step-ahead forecast errors to a vector of structural innovations ϵ_t. We assume that

$$u_t = \Phi_\epsilon \epsilon_t = \Sigma_{tr}\Omega\epsilon_t, \tag{118}$$

where Σ_{tr} is the unique lower-triangular Cholesky factor of Σ with nonnegative diagonal elements, and Ω is an $n \times n$ orthogonal matrix satisfying $\Omega\Omega' = I$. The second equality ensures that the covariance matrix of u_t is preserved in the sense that

$$\Phi_\epsilon \Phi_\epsilon' = \Sigma_{tr}\Omega\Omega'\Sigma_{tr}' = \Sigma. \tag{119}$$

By construction, the covariance matrix of the forecast error is invariant to the choice of Ω, which implies that it is not possible to identify Ω from the data. In turn, much of the literature on structural VARs reduces to arguments about an appropriate set of restrictions for the matrix Ω. Detailed surveys about the restrictions, or identification schemes, that have been used in the literature to identify innovations to technology, monetary policy, government spending, and other exogenous shocks can be found, for instance, in Cochrane (1994), Christiano et al. (1999), Stock and Watson (2001), and Ramey (2016). Conditional on an estimate of the reduced-form coefficient matrices Φ and Σ and an identification scheme for one or more columns of Ω, it is straightforward to express the impulse response as

$$\widehat{IRF}^{V}(.,j,h) = C_h(\hat{\Phi})\hat{\Sigma}_{tr}[\Omega]_{.j}, \tag{120}$$

where the moving average coefficient matrix $C_h(\hat{\Phi})$ can be obtained from the companion form representation of the VAR in (114): $C_h(\Phi) = M'\tilde{\Phi}_1^h M$ with $M' = [I_n, 0_{n\times n(p-1)}]$.

For illustrative purposes, rather than conditioning the computation of impulse response functions on a particular choice of Ω, we follow the recent literature on sign restrictions; see Faust (1998), Canova and De Nicoló (2002), and Uhlig (2005). The key idea of this literature is to restrict the matrices Ω to a set $\mathcal{O}(\Phi, \Sigma)$ such that the implied impulse response functions satisfy certain sign restrictions. This means that the magnitude of the impulse responses are only set-identified. Using our estimated VAR(1) in output growth, log labor share, inflation, and interest rates, we impose the condition that in response to a contractionary monetary policy shock interest rates increase and inflation is negative for four quarters. Without loss of generality, we can assume that the shocks are ordered such that the first column of Ω, denoted by q, captures the effect of the monetary policy shock. Conditional on the reduced-form VAR coefficient estimates $(\hat{\Phi}, \hat{\Sigma})$, we can determine the set of unit-length vectors q such that the implied impulse responses satisfy the sign restrictions. The bands depicted in Fig. 22 delimit the upper and lower bounds of the estimated identified sets for the pointwise impulse responses of output, labor share, inflation, and interest rates to a monetary policy shock. The sign restrictions that are imposed on the monetary policy shock are not sufficiently strong to determine the sign of the output and labor share responses to a monetary policy shock. Note that if a researcher selects a particular q (possibly as a function of the reduced-form parameters Φ and Σ), then the bands in the figure would reduce to a single line, which is exemplified by the solid line in Fig. 22.

8.3.4 Conditional Moment Restrictions

The unconditional moment restrictions derived from the equilibrium conditions of the DSGE model discussed in Section 8.2.4 have sample analogs in which the population expectations are replaced by sample averages. A complication arises if the moment

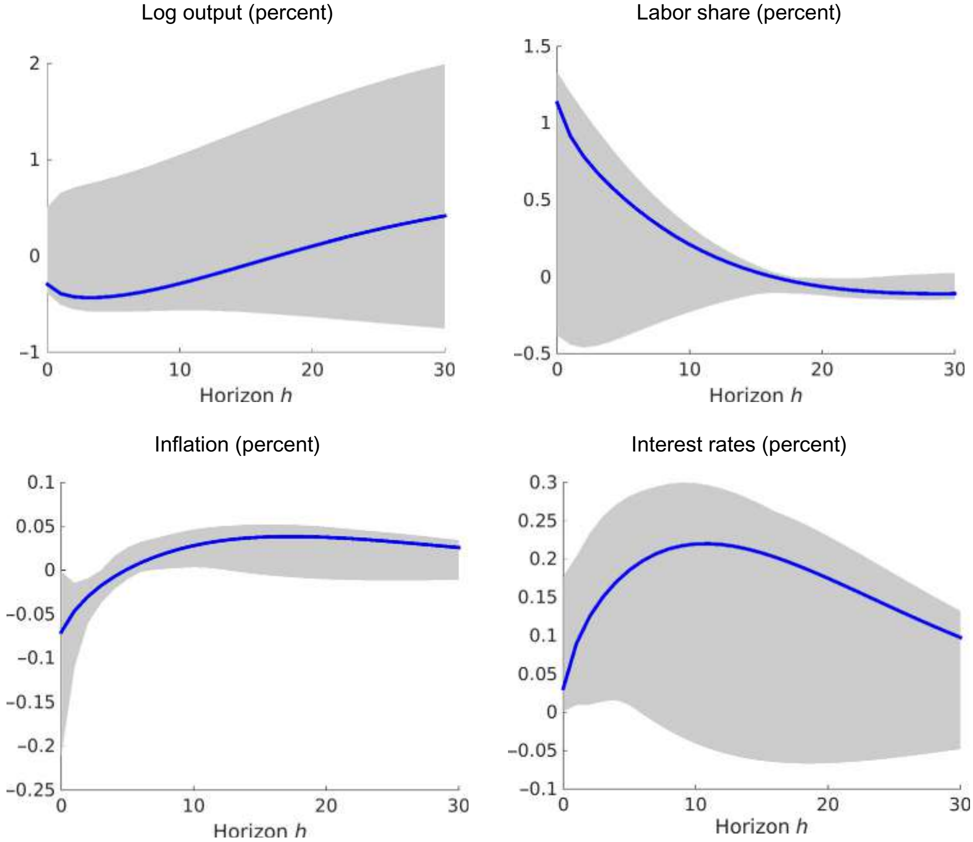


Fig. 22 Impulse responses to a monetary policy shock. *Notes*: Impulse responses to a one-standard-deviation monetary policy shock. Inflation and interest rate responses are not annualized. The bands indicate pointwise estimates of identified sets for the impulse responses based on the assumption that a contractionary monetary policy shock raises interest rates and lowers inflation for four quarters. The *solid line* represents a particular impulse response function contained in the identified set.

conditions contain latent variables, eg, the shock process λ_t in the moment condition (105) derived from the New Keynesian Phillips curve. Sample analogs of population moment conditions can be used to form generalized method of moments (GMM) estimators, which are discussed in Section 11.4.

8.4 Dealing with Trends

Trends are a salient feature of macroeconomic time series. The stylized DSGE model presented in Section 8.1 features a stochastic trend generated by the productivity process $\log Z_t$, which evolves according to a random walk with drift. While the trend in productivity induces a common trend in consumption, output, and real wages, the model

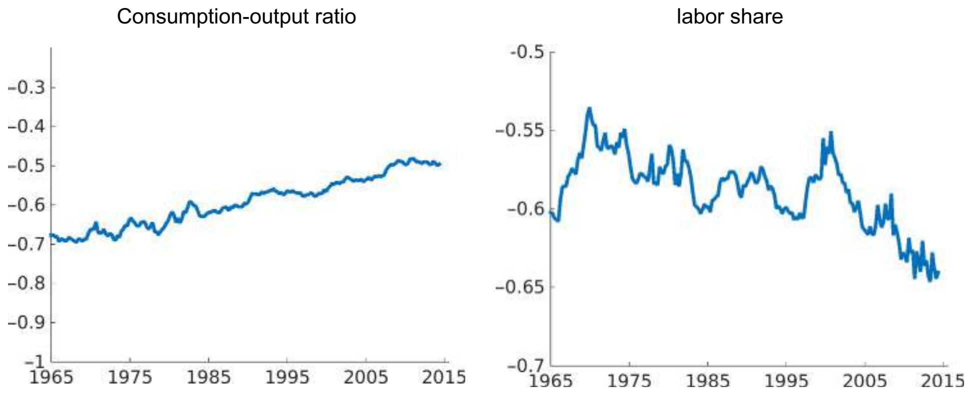


Fig. 23 Consumption-output ratio and Labor share (in logs).

specification implies that the log consumption-output ratio and the log labor share are stationary. Fig. 23 depicts time series of the US log consumption-output ratio and the log labor share for the United States from 1965 to 2014. Here the consumption-output ratio is defined as Personal Consumption Expenditure on Services (PCESV) plus Personal Consumption Expenditure on nondurable goods (PCND) divided by nominal GDP. The consumption-output ratio has a clear upward trend and the labor share has been falling since the late 1990s. Because these trends are not captured by the DSGE model, they lead to a first-order discrepancy between actual US and model-generated data.

Most DSGE models that are used in practice have counterfactual trend implications because they incorporate certain cotrending restrictions, eg, a balanced growth path along which output, consumption, investment, the capital stock, and real wages exhibit a common trend and hours worked and the return on capital are stationary, that are to some extent violated in the data as we have seen in the above example. Researchers have explored various remedies to address the mismatch between model and data, including: (i) detrending each time series separately and fitting the DSGE model to detrended data; (ii) applying an appropriate trend filter to both actual data and model-implied data when confronting the DSGE model with data; (iii) creating a hybrid model, eg, Canova (2014) that consists of a flexible, nonstructural trend component and uses the structural DSGE model to describe fluctuations around the reduced-form trend; and (iv) incorporating more realistic trends directly into the structure of the DSGE model. From a modeling perspective, option (i) is the least desirable and option (iv) is the most desirable choice.

9. STATISTICAL INFERENCE

DSGE models have a high degree of theoretical coherence. This means that the functional forms and parameters of equations that describe the behavior of macroeconomic aggregates are tightly restricted by optimality and equilibrium conditions. In turn, the family of probability distributions $p(Y|\theta)$, $\theta \in \Theta$, generated by a DSGE model tends to be more restrictive than the family of distributions associated with an atheoretical model, such as a reduced-form VAR as in (113). This may place the empirical researcher in a situation in which the data favor the atheoretical model and the atheoretical model generates more accurate forecasts, but a theoretically coherent model is required for the analysis of a particular economic policy. The subsequent discussion of statistical inference will devote special attention to this misspecification problem.

The goal of statistical inference is to infer an unknown parameter vector θ from observations Y; to provide a measure of uncertainty about θ; and to document the fit of the statistical model. The implementation of these tasks becomes more complicated if the statistical model suffers from misspecification. Confronting DSGE models with data can essentially take two forms. If it is reasonable to assume that the probabilistic structure of the DSGE model is well specified, then one can ask how far the observed data $Y^o_{1:T}$ or sample statistics $\mathcal{S}(Y^o_{1:T})$ computed from the observed data fall into the tails of the model-implied distribution derived from $p(Y_{1:T}|\theta)$. The parameter vector θ can be chosen to ensure that the density (likelihood) of $\mathcal{S}(Y^o_{1:T})$ is high under the distribution $p(Y_{1:T}|\theta)$. If, on the other hand, there is a strong belief (possibly supported by empirical evidence) that the probabilistic structure of the DSGE model is not rich enough to capture the salient features of the observed data, it is more sensible to consider a reference model with a well-specified probabilistic structure, use it to estimate some of the population objects introduced in Section 8.2, and compare these estimates to their model counterparts.

In Section 9.1 we ask the question whether the DSGE model parameters can be determined based on observations Y and review the recent literature on identification. We then proceed by reviewing two modes of statistical inference: frequentist and Bayesian.[ak] We pay special attention to the consequences of model misspecification. Frequentist inference, introduced in Section 9.2, takes a preexperimental perspective and focuses on the behavior of estimators and test statistics, which are functions of the observations Y, in repeated sampling under the distribution $\mathbb{P}^Y_\theta$. Frequentist inference is conditioned on a "true" but unknown parameter θ, or on a data-generating process (DGP), which is a hypothetical probability distribution under which the data are assumed to be generated. Frequentist procedures have to be well behaved for all values of $\theta \in \Theta$. Bayesian inference, introduced in Section 9.3, takes a postexperimental perspective by treating the unknown parameter θ as a random variable and updating a prior distribution $p(\theta)$ in view of the data Y using Bayes Theorem to obtain the posterior distribution $p(\theta|Y)$.

Estimation and inference requires that the model be solved many times for different parameter values θ. The subsequent numerical illustrations are based on the stylized

[ak] A comparison between econometric inference approaches and the calibration approach advocated by Kydland and Prescott (1982) can be found in Ríos-Rull et al. (2012).

DSGE model introduced in Section 8.1, for which we have a closed-form solution. However, such closed-form solutions are the exception and typically not available for models used in serious empirical applications. Thus, estimation methods, both frequentist and Bayesian, have to be closely linked to model solution procedures. This ultimately leads to a trade-off: given a fixed amount of computational resources, the more time is spent on solving a model conditional on a particular θ, eg, through the use of a sophisticated projection technique, the less often an estimation objective function can be evaluated. For this reason, much of the empirical work relies on first-order perturbation approximations of DSGE models, which can be obtained very quickly. The estimation of models solved with numerically sophisticated projection methods is relatively rare, because it requires a lot of computational resources. Moreover, as discussed in Part I, perturbation solutions are more easily applicable to models with a high-dimensional state vector and such models, in turn, are less prone to misspecification and are therefore more easily amenable to estimation. However, the recent emergence of low-cost parallel programming environments and cloud computing will make it feasible for a broad group of researchers to solve and estimate elaborate nonlinear DSGE models in the near future.

9.1 Identification

The question of whether a parameter vector θ is identifiable based on a sample Y is of fundamental importance for statistical inference because one of the main objectives is to infer the unknown θ based on the sample Y. Suppose that the DSGE model generates a family of probability distributions $p(Y|\theta)$, $\theta \in \Theta$. Moreover, imagine a stylized setting in which data are in fact generated from the DSGE model conditional on some "true" parameter θ_0. The parameter vector θ_0 is globally identifiable if

$$p(Y|\theta) = p(Y|\theta_0) \;\text{ implies }\; \theta = \theta_0. \tag{121}$$

The statement is somewhat delicate because it depends on the sample Y. From a preexperimental perspective, the sample is unobserved and it is required that (121) holds with probability one under the distribution $p(Y|\theta_0)$. From a postexperimental perspective, the parameter θ may be identifiable for some trajectories Y, but not for others. The following example highlights the subtle difference. Suppose that

$$y_{1,t}|(\theta, y_{2,t}) \sim iidN(\theta y_{2,t}, 1), \quad y_{2,t} = \begin{cases} 0 & \text{w.p. } 1/2 \\ \sim iidN(0,1) & \text{w.p. } 1/2 \end{cases}$$

Thus, with probability (w.p.) 1/2, one observes a trajectory along which θ is not identifiable because $y_{2,t} = 0$ for all t. If, on the other hand, $y_{2,t} \neq 0$, then θ is identifiable.

9.1.1 Local Identification

If condition (121) is only satisfied for values of θ in an open neighborhood of θ_0, then θ_0 is locally identified. Most of the literature has focused on devising procedures to check local identification in linearized DSGE models with Gaussian innovations. In this case the distribution of $Y|\theta$ is a joint normal distribution and can be characterized by a $Tn_y \times 1$ vector of means $\mu(\theta)$ (where n is the dimension of the vector y_t) and a $Tn_y \times Tn_y$ covariance matrix $\Sigma(\theta)$. Defining $m(\theta) = [\mu(\theta)', vech(\Sigma(\theta))']'$, where $vech(\cdot)$ vectorizes the nonredundant elements of a symmetric matrix, we can restate the identification condition as

$$m(\theta) = m(\theta_0) \text{ implies } \theta = \theta_0. \tag{122}$$

Thus, verifying the local identification condition is akin to checking whether the Jacobian

$$\mathcal{J}(\theta) = \frac{\partial}{\partial \theta'} m(\theta) \tag{123}$$

is of full rank. This approach was proposed and applied by Iskrev (2010) to examine the identification of linearized DSGE models. If the joint distribution of Y is not Gaussian, say because the DSGE model innovations ϵ_t are non-Gaussian or because the DSGE model is nonlinear, then it is possible that θ_0 is not identifiable based on the first and second moments $m(\theta)$, but that there are other moments that make it possible to distinguish θ_0 from $\tilde{\theta} \neq \theta_0$.

Local identification conditions are often stated in terms of the so-called information matrix. Using Jensen's inequality, it is straightforward to verify that the Kullback–Leibler discrepancy between $p(Y|\theta_0)$ and $p(Y|\theta)$ is nonnegative:

$$\Delta_{KL}(\theta|\theta_0) = -\int \log\left(\frac{p(Y|\theta)}{p(Y|\theta_0)}\right) p(Y|\theta_0) dY \geq 0. \tag{124}$$

Under a nondegenerate probability distribution for Y, the relationship holds with equality only if $p(Y|\theta) = p(Y|\theta_0)$. Thus, we deduce that the Kullback–Leibler distance is minimized at $\theta = \theta_0$ and that θ_0 is identified if θ_0 is the unique minimizer of $\Delta_{KL}(\theta|\theta_0)$. Let $\ell(\theta|Y) = \log p(Y|\theta)$ denote the log-likelihood function and $\nabla^2_\theta \ell(\theta|Y)$ denote the matrix of second derivatives of the log-likelihood function with respect to θ (Hessian), then (under suitable regularity conditions that allow the exchange of integration and differentiation)

$$\nabla_{\theta^2} \Delta_{KL}(\theta_0|\theta_0) = \int \nabla_{\theta^2} \ell(\theta_0|Y) p(Y|\theta_0) dY. \tag{125}$$

In turn, the model is locally identified at θ_0 if the expected Hessian matrix is nonsingular.

For linearized Gaussian DSGE models that can be written in the form $Y \sim N(\mu(\theta), \Sigma(\theta))$ we obtain

$$\int \nabla^2_\theta \ell(\theta_0|Y) p(Y|\theta_0) dY = \mathcal{J}(\theta)' \Omega \mathcal{J}(\theta), \tag{126}$$

where $\boldsymbol{\Omega}$ is the Hessian matrix associated with the unrestricted parameter vector $m = [\mu', vech(\Sigma)']'$ of a $N(\mu,\Sigma)$. Because $\boldsymbol{\Omega}$ is a symmetric full-rank matrix of dimension dim(m), we deduce that the Hessian is of full rank whenever the Jacobian matrix in (123) is of full rank.

Qu and Tkachenko (2012) focus on the spectral density matrix of the process y_t. Using a frequency domain approximation of the likelihood function and utilizing the information matrix equality, they express the Hessian as the outer product of the Jacobian matrix of derivatives of the spectral density with respect to θ

$$G(\theta_0) = \int_{-\pi}^{\pi} \left(\frac{\partial}{\partial \theta'} vec(f_{yy}(\omega)') \right)' \left(\frac{\partial}{\partial \theta'} vec(f_{yy}(\omega)) \right) d\omega \tag{127}$$

and propose to verify whether $G(\theta_0)$ is of full rank. The identification checks of Iskrev (2010) and Qu and Tkachenko (2012) have to be implemented numerically. For each conjectured θ_0 the user has to compute the rank of the matrices $\mathcal{J}(\theta_0)$ or $G(\theta_0)$, respectively. Because in a typical implementation the computation of the matrices relies on numerical differentiation (and integration), careful attention has be paid to the numerical tolerance level of the procedure that computes the matrix rank. Detailed discussions can be found in the two referenced papers.

Komunjer and Ng (2011) take a different route to assess the local identification of linearized DSGE models. They examine the relationship between the coefficients of the state-space representation of the DSGE model and the parameter vector θ. Recall that the state-space representation takes the form

$$y_t = \Psi_0(\theta) + \Psi_1(\theta), \quad s_t = \Phi_1(\theta) s_{t-1} + \Phi_\epsilon(\theta)\epsilon_t. \tag{128}$$

The notation highlights the dependence of the coefficient matrices on θ. Now stack the coefficients of the Ψ and Φ matrices in the vector ϕ:

$$\phi = \left[vec(\Psi_0)', vec(\Psi_1)', vec(\Phi_1)', vec(\Phi_\epsilon)' \right]'.$$

It is tempting to conjecture that θ is locally identifiable if the Jacobian matrix associated with the mapping from economic parameters θ to the reduced-form state-space parameters ϕ

$$\frac{\partial}{\partial \theta'} \phi(\theta) \tag{129}$$

has full column rank at θ_0. The problem with this conjecture is that the reduced-form parameters ϕ themselves are not identifiable. Let A be a nonsingular $n_s \times n_s$ matrix and $\boldsymbol{\Omega}$ an $n_\epsilon \times n_\epsilon$ orthogonal matrix, then we can define

$$\tilde{s}_t = A s_t, \quad \tilde{\epsilon}_t = \boldsymbol{\Omega} \epsilon_t, \quad \tilde{\Psi}_1 = \Psi_1 A^{-1}, \quad \tilde{\Phi}_1 = \Phi_1 A^{-1}, \quad \tilde{\Phi}_\epsilon = A \Phi_\epsilon \boldsymbol{\Omega}'$$

to obtain an observationally equivalent state-space system

$$y_t = \Psi_0 + \tilde{\Psi}_1 \tilde{s}_t, \quad s_t = \tilde{\Phi}_1 \tilde{s}_{t-1} + \tilde{\Phi}_\epsilon \epsilon_t \tag{130}$$

with $\phi \neq \tilde{\phi}$. Thus, the number of identifiable reduced-form parameters is smaller than the number of elements in the Ψ and Φ matrices. The main contribution in Komunjer and Ng (2011) is to account for the nonidentifiability of the reduced-form state-space parameters when formulating a rank condition along the lines of (129). In many DSGE models a subset of the state transitions are deterministic, which complicates the formal analysis.

Identification becomes generally more tenuous the fewer variables are included in the vector y_t. For instance, in the context of the stylized DSGE model, suppose y_t only includes the labor share. According to (70) the law of motion for the labor share is the sum of three AR(1) processes and an *iid* monetary policy shock. It can be rewritten as an ARMA(3,3) process and therefore has at most 8 identifiable reduced-form parameters. Thus, the upper bound on the number of reduced-form parameters is less than the number of DSGE model parameters, which is 13. In turn, it is not possible to identify the entire θ vector.

9.1.2 Global Identification

Global identification is more difficult to verify than local identification. Consider the following example from Schorfheide (2013):

$$y_t = [1 \quad 1]s_t, \quad s_t = \begin{bmatrix} \theta_1^2 & 0 \\ 1 - \theta_1^2 - \theta_1\theta_2 & (1-\theta_1^2) \end{bmatrix} s_{t-1} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \epsilon_t, \quad \epsilon_t \sim iidN(0,1). \tag{131}$$

Letting L denote the lag operator with the property that $Ly_t = y_{t-1}$, one can write the law of motion of y_t as an restricted ARMA(2,1) process:

$$\left(1-\theta_1^2 L\right)\left(1-(1-\theta_1^2)L\right)y_t = (1-\theta_1\theta_2 L)\epsilon_t. \tag{132}$$

It can be verified that given θ_1 and θ_2, an observationally equivalent process can be obtained by choosing $\tilde{\theta}_1$ and $\tilde{\theta}_2$ such that

$$\tilde{\theta}_1 = \sqrt{1-\theta_1^2}, \quad \tilde{\theta}_2 = \theta_1\theta_2/\tilde{\theta}_1.$$

Here we switched the values of the two roots of the autoregressive lag polynomial. Qu and Tkachenko (2014) propose to check for global identification by searching for solutions to the equation

$$0 = \triangle_{KL}(\theta|\theta_0), \quad \theta \in \Theta. \tag{133}$$

If θ_0 is the unique solution, then the DSGE model is globally identified. The authors evaluate the Kullback–Leibler discrepancy using a frequency domain transformation. The computational challenge is to find all the roots associated with (133). Kociecki and Kolasa (2015) follow a slightly different approach that is attractive because it requires the user to solve the DSGE model only at θ_0, but not at all the other values of $\theta \in \Theta$.

9.2 Frequentist Inference

The fundamental problem of statistical inference is to infer the parameter vector θ, in our case the DSGE model parameters, based on a random sample Y. Frequentist inference adopts a preexperimental perspective and examines the sampling distribution of estimators and test statistics, which are transformations of the random sample Y, conditional on a hypothetical DGP. We will distinguish between two cases. First, we consider the stylized case in which the DSGE model is correctly specified. Formally, this means that Y is sampled from $p(Y|\theta_0)$, where the density $p(Y|\theta_0)$ is derived from the DSGE model and θ is the "true" but unknown parameter vector.[al] Second, we consider the case of misspecification, meaning the DSGE model is too stylized to capture some of the key features of the data Y. As a consequence, the sampling distribution of Y has to be characterized by a reference model, for instance, a VAR or a linear process. In terms of notation, we will distinguish between the DSGE model, denoted by M_1, and the reference model M_0. To avoid confusion about which model generates the sampling distribution of Y, we add the model indicator to the conditioning set and write, eg, $p(Y|\theta,M_1)$ or $p(Y|M_0)$. We also use the notation $\|a\|_W = a'Wa$.

9.2.1 "Correct" Specification of DSGE Model

Under the assumption of correct specification, the DSGE model itself is the DGP and $p(Y|\theta_0,M_1)$ describes the sampling distribution of Y under which the behavior of estimators and test statistics is being analyzed. In this case it is desirable to let the model-implied probability distribution $p(Y|\theta_0,M_1)$ determine the choice of the objective function for estimators and test statistics to obtain a statistical procedure that is efficient (meaning that the estimator is close to θ_0 with high probability in repeated sampling). In this regard, the maximum likelihood (ML) estimator

$$\hat{\theta}_{ml} = \operatorname{argmax}_{\theta\in\Theta}\ \log p(Y|\theta,M_1) \tag{134}$$

plays a central role in frequentist inference, because it is efficient under fairly general regularity conditions. One of these conditions is that θ_0 is identifiable.

Alternative estimators can be obtained by constructing an objective function $Q_T(\theta|Y)$ that measures the discrepancy between sample statistics $\hat{m}_T(Y)$ (see Section 8.3) and model-implied population statistics $\mathbb{E}[\hat{m}_T(Y)|\theta,M_1]$ (see Section 8.2). Examples of the vector $\hat{m}_T(Y)$ are, for instance, vectorized sample autocovariances such as

[al] In reality, of course, the observed Y is never generated from a probabilistic mechanism. Instead it reflects measured macroeconomic activity. Thus, by "correct specification of a DSGE model" we mean that we believe that its probabilistic structure is rich enough to assign high probability to the salient features of macroeconomic time series.

$$\hat{m}_T(Y) = \left[vech(\hat{\Gamma}_{yy}(0))', vec(\hat{\Gamma}_{yy}(1))' \right] = \frac{1}{T}\sum_{t=1}^{T} m(y_{t-1:t})$$

or the OLS estimator of the coefficients of a VAR(1) (here without intercept)

$$\hat{m}_T(Y) = vec\left(\left(\frac{1}{T}\sum_{t=1}^{T} y_{t-1}y'_{t-1} \right)^{-1} \frac{1}{T}\sum_{t=1}^{T} y_{t-1}y'_t \right).$$

We write the estimation objective function as

$$Q_T(\theta|Y) = \|\hat{m}_T(Y) - \mathbb{E}[\hat{m}_T(Y)|\theta, M_1]\|_{W_T}, \tag{135}$$

where W_T is a symmetric positive-definite weight matrix. Under the assumption of a correctly specified DSGE model, the optimal choice of the weight matrix W_T is the inverse of the DSGE model-implied covariance matrix of $\hat{m}_T(Y)$. Thus, more weight is assigned to sample moments that accurately approximate the underlying population moment. The minimum distance (MD) estimator of θ is defined as

$$\hat{\theta}_{md} = \text{argmax}_{\theta \in \Theta}\ Q_T(\theta|Y). \tag{136}$$

Econometric inference is based on the sampling distribution of the estimator $\hat{\theta}_{md}$ and confidence sets and test statistics derived from $\hat{\theta}_{md}$ and $Q_T(\theta|Y)$ under the distribution $p(Y|\theta_0, M_1)$.

9.2.2 Misspecification and Incompleteness of DSGE Models

Model misspecification can be interpreted as a violation of the cross-coefficient restrictions embodied in the mapping from the DSGE model parameters θ into the system matrices Ψ_0, Ψ_1, Φ_1, and Φ_ϵ of the state-space representation in (76) and (78). An example of an incomplete model is a version of the stylized DSGE model in which we do not fully specify the law of motion for the exogenous shock processes and restrict our attention to certain moment conditions, such as the consumption Euler equation. In some cases, incompleteness and misspecification are two sides of the same coin. Consider a version of the stylized DSGE model with only one structural shock, namely, the monetary policy shock. This version does not contain sufficiently many shocks to explain the observed variability in output growth, the labor share, inflation, and the interest rate. More specifically, the one-shock DSGE model implies, for instance, that the linear combination

$$\frac{1}{\kappa_p(1+\nu)x_{\epsilon_R}/\beta + \sigma_R}\hat{R}_t - \frac{1}{\kappa_p(1+\nu)x_{\epsilon_R}}\hat{\pi}_t = 0$$

is perfectly predictable; see (72) and (73). This prediction is clearly counterfactual. We could regard the model as misspecified, in the sense that its predictions are at odds with the data; or as incomplete, in the sense that adding more structural shocks could reduce the gap between model and reality.

Regardless of whether the DSGE model is incomplete or misspecificed, it does not produce a sampling distribution for the data Y that can be used to determine the frequentist behavior of estimators and test statistics. In order to conduct a frequentist analysis, we require a reference model M_0 that determines the distribution of the data $p(Y|M_0)$ and can be treated as a DGP. The reference model could be a fully specified parametric model such as a VAR, $p(Y|\phi, M_0)$, where ϕ is a finite-dimensional parameter vector. Alternatively, the reference model could be a general stochastic process for $\{y_t\}$ that satisfies a set of regularity conditions necessary to establish large sample approximations of the sampling distributions of estimators and test statistics.

If the DSGE model is incompletely specified, it is still possible to uphold the notion of a "true" parameter vector θ_0, in the sense that one could imagine the DGP to be the incompletely specified DSGE model augmented by a set of equations (potentially with additional parameters). If the DSGE model is misspecified, then the concept of a "true" parameter value has to be replaced by the notion of a pseudo-true (or pseudo-optimal) parameter value. The definition of a pseudo-true parameter value requires a notion of discrepancy between the DGP $p(Y|M_0)$ and the DSGE model $p(Y|\theta, M_1)$. Different discrepancies lead to different pseudo-optimal values. Likelihood-based inference is associated with the Kullback–Leibler discrepancy and would lead to

$$\theta_0(KL) = \text{argmin}_{\theta\in\Theta} \; -\int \log\left(\frac{p(Y|\theta, M_1)}{p(Y|M_0)}\right) p(Y|M_0)dY. \tag{137}$$

Moment-based inference based on the sample objective function $Q_T(\theta|Y)$ is associated with a pseudo-optimal value

$$\theta_0(Q, W) = \text{argmin}_{\theta\in\Theta}\; Q(\theta|M_0), \tag{138}$$

where

$$Q(\theta|M_0) = \|\mathbb{E}[\hat{m}_T(Y)|M_0] - \mathbb{E}[\hat{m}(Y)|\theta, M_1]\|_W.$$

Ultimately, the sampling properties of estimators and test statistics have to be derived from the reference model M_0.

9.3 Bayesian Inference

Under the Bayesian paradigm, the calculus of probability is used not only to deal with uncertainty about shocks ϵ_t, states s_t, and observations y_t, but also to deal with uncertainty about the parameter vector θ. The initial state of knowledge (or ignorance) is summarized by a prior distribution with density $p(\theta)$. This prior is combined with the conditional distribution of the data given θ, ie, the likelihood function, to characterize the joint distribution of parameters and data. Bayes Theorem is applied to obtain the conditional distribution of the parameters given the observed data Y. This distribution is called the posterior distribution:

$$p(\theta|Y,M_1)=\frac{p(Y|\theta,M_1)p(\theta|M_1)}{p(Y|M_1)},\quad p(Y|M_1)=\int p(Y|\theta,M_1)p(\theta|M_1)d\theta. \tag{139}$$

The posterior distribution contains all the information about θ conditional on sample information Y. In a Bayesian setting a model comprises the likelihood function $p(Y|\theta, M_1)$ *and* the prior $p(\theta|M_1)$.

The posterior distribution of transformations of the DSGE model parameters θ, say, $h(\theta)$, eg, autocovariances and impulse response functions, can be derived from $p(\theta|Y, M_1)$. For instance,

$$\mathbb{P}_Y\{h(\theta)\leq\bar{h}\}=\int_{\theta\ |\ h(\theta)\leq\bar{h}}p(\theta|Y,M_1)d\theta. \tag{140}$$

Solutions to inference problems can generally be obtained by specifying a suitable loss function, stating the inference problem as a decision problem, and minimizing posterior expected loss. For instance, to obtain a point estimator for $h(\theta)$, let $L(h(\theta),\delta)$ describe the loss associated with reporting δ if $h(\theta)$ is correct. The optimal decision δ_* is obtained by minimizing the posterior expected loss:

$$\delta_*=\text{argmin}_{\delta\in\mathcal{D}}\int L(h(\theta),\delta)p(\theta|Y,M_1)d\theta. \tag{141}$$

If the loss function is quadratic, then the optimal point estimator is the posterior mean of $h(\theta)$.

The most difficult aspect of Bayesian inference is the characterization of the posterior moments of $h(\theta)$. Unfortunately, it is not possible to derive these moments analytically for DSGE models. Thus, researchers have to rely on numerical methods. The Bayesian literature has developed a sophisticated set of algorithms to generate draws θ^i from the posterior distribution, such that averages of these draws converge to posterior expectations:

$$\mathbb{E}[h(\theta)|Y,M_1]=\int h(\theta)p(\theta|Y,M_1)d\theta\approx\frac{1}{N}\sum_{i=1}^{N}h(\theta^i). \tag{142}$$

Several of these computational techniques are discussed in more detail in Section 12.

9.3.1 "Correct" Specification of DSGE Models

The use of Bayes Theorem to learn about the DSGE model parameters implicitly assumes that the researcher regards the probabilistic structure of the DSGE model as well specified in the sense that there are parameters θ in the support of the prior distribution conditional on which the salient features of the data Y are assigned a high probability. Of course, in practice there is always concern that an alternative DSGE model may deliver a better description of the data. The Bayesian framework is well suited to account for model uncertainty.

Suppose the researcher contemplates two model specifications M_1 and M_2, assuming that one of them is correct. It is natural to place prior probabilities on the two models, which we denote by $\pi_{j,0}$. Ratios of model probabilities are called model odds. The posterior odds of M_1 vs M_2 are given by

$$\frac{\pi_{1,T}}{\pi_{2,T}} = \frac{\pi_{1,0}}{\pi_{2,0}} \frac{p(Y|M_1)}{p(Y|M_2)}, \tag{143}$$

where the first factor on the right-hand side captures the prior odds and the second factor, called Bayes factor, is the ratio of marginal data densities. Note that $p(Y|M_i)$ appears in the denominator of Bayes Theorem (139). Posterior model odds and probabilities have been widely used in the DSGE model literature to compare model specification or to take averages across DSGE models. Prominent applications include Rabanal and Rubio-Ramírez (2005) and Smets and Wouters (2007).

9.3.2 Misspecification of DSGE Models

As in the frequentist case, model misspecification complicates inference. Several approaches have been developed in the literature to adapt Bayesian analysis to the potential misspecification of DSGE models. In general, the model space needs to be augmented by a more densely parameterized reference model, M_0, that provides a more realistic probabilistic representation of the data.

Schorfheide (2000) considers a setting in which a researcher is interested in the relative ability of two (or more) DSGE models, say, M_1 and M_2, to explain certain population characteristics φ, eg, autocovariances or impulse responses.[am] However, the DSGE models may be potentially misspecified and the researcher considers a reference model M_0. As long as it is possible to form a posterior distribution for φ based on the reference model, the overall posterior can be described by

$$p(\varphi|Y) = \sum_{j=0,1,2} \pi_{j,T} p(\varphi|Y, M_j). \tag{144}$$

If one of the DSGE models is well specified, this model receives high posterior probability and dominates the mixture. If both DSGE models are at odds with the data, the posterior probability of the reference model will be close to one. Given a loss function over predictions of φ, one can compute DSGE model-specific predictions:

$$\hat{\varphi}_{(j)} = \text{argmin}_{\tilde{\varphi}} \int L(\tilde{\varphi}, \varphi) p(\varphi|Y, M_j) d\varphi, \quad j = 1, 2. \tag{145}$$

[am] Frequentist versions of this approach have been developed in Hnatkosvaka et al. (2012) and Marmer and Otsu (2012).

Finally, the two DSGE models can be ranked based on the posterior risk

$$\int L(\hat{\varphi}_{(j)}, \varphi) p(\varphi | Y) d\varphi. \tag{146}$$

Geweke (2010) assumes that the researcher regards the DSGE models not as models of the data Y, but as models of some population moments φ. A reference model M_0, eg, a VAR, provides the model for Y, but also permits the computation of implied population moments. He shows that under these assumptions, one can define the posterior odds of DSGE models as

$$\frac{\pi_{1,T}}{\pi_{2,T}} = \frac{\pi_{1,0}}{\pi_{2,0}} \frac{\int p(\varphi | M_1) p(\varphi | Y, M_0) d\varphi}{\int p(\varphi | M_2) p(\varphi | Y, M_0) d\varphi}. \tag{147}$$

Roughly, if we were able to observe φ, then $p(\varphi | M_j)$ is the marginal likelihood. However, φ is unobservable and therefore replaced by a posterior predictive distribution obtained from a reference model M_0. The odds in favor of model M_1 are high if there is a lot of overlap between the preditive distribution for the population moments φ under the DSGE model, and the posterior distribution of φ obtained when estimating the reference model M_0.

Building on work by Ingram and Whiteman (1994); Del Negro and Schorfheide (2004) do not treat the DSGE model as a model of the data Y, but instead use it to construct a prior distribution for a VAR. Consider the companion form VAR in (114). Use the DSGE model to generate a prior distribution for $(\tilde{\Phi}_1, \tilde{\Phi}_0, \tilde{\Sigma})$ and combine this prior with the VAR likelihood function

$$p(Y, \tilde{\Phi}_0, \tilde{\Phi}_1, \tilde{\Sigma}, \theta | \lambda) = p(Y | \tilde{\Phi}_0, \tilde{\Phi}_1, \tilde{\Sigma}) p(\tilde{\Phi}_0, \tilde{\Phi}_1, \tilde{\Sigma} | \theta, \lambda) p(\theta). \tag{148}$$

The resulting hierarchical model is called a DSGE-VAR. The prior $p(\tilde{\Phi}_0, \tilde{\Phi}_1, \tilde{\Sigma} | \theta, \lambda)$ is centered on restriction functions

$$\tilde{\Phi}_0^*(\theta), \quad \tilde{\Phi}_1^*(\theta), \quad \tilde{\Sigma}^*(\theta),$$

but allows for deviations from these restriction functions to account for model misspecification. The parameter λ is a hyperparameter that controls the magnitude of the deviations (prior variance) from the restriction function. This framework can be used for forecasting, to assess the fit of DSGE models, eg, Del Negro et al. (2007), and to conduct policy analysis, eg, Del Negro and Schorfheide (2009).

In a setting in which the reference model M_0 plays a dominating role, Fernández-Villaverde and Rubio-Ramírez (2004) show that choosing the DSGE model that attains the highest posterior probability (among, say, competing DSGE models M_1 and M_2) leads asymptotically to the specification that is closest to M_0 in a Kullback–Leibler sense. Rather than using posterior probabilities to select among or average across two DSGE models, one can form a prediction pool, which is essentially a linear combination of two predictive densities:

$$\lambda p(y_t|Y_{1:t-1},M_1) + (1-\lambda)p(y_t|Y_{1:t-1},M_2).$$

The weight $\lambda \in [0,1]$ can be determined based on

$$\prod_{t=1}^{T} [\lambda p(y_t|Y_{1:t-1},M_1) + (1-\lambda)p(y_t|Y_{1:t-1},M_2)].$$

This objective function could either be maximized with respect to λ or it can be treated as a likelihood function for λ and embedded in a Bayesian inference procedure. This idea is developed in Geweke and Amisano (2011) and Geweke and Amisano (2012). Dynamic versions with λ depending on time t are provided by Waggoner and Zha (2012) and Del Negro et al. (2014).

10. THE LIKELIHOOD FUNCTION

The likelihood function plays a central role in both frequentist and Bayesian inference. The likelihood function treats the joint density of the observables conditional on the parameters, $p(Y_{1:T}|\theta)$, as a function of θ. The state-space representation of the DSGE model leads to a joint distribution $p(Y_{1:T},S_{1:T}|\theta)$; see (79). In order to obtain the likelihood function, one needs to integrate out the (hidden) states $S_{1:T}$. This can be done recursively, using an algorithm that is a called a *filter*.

This section focuses on the numerical evaluation of the likelihood function conditional on a particular parameterization θ through the use of linear and nonlinear filters. We assume that the DSGE model has the following, possibly nonlinear, state-space representation:

$$\begin{aligned} y_t &= \Psi(s_t, t;\theta) + u_t, \quad u_t \sim F_u(\cdot;\theta) \\ s_t &= \Phi(s_{t-1}, \epsilon_t;\theta), \quad \epsilon_t \sim F_\epsilon(\cdot;\theta). \end{aligned} \tag{149}$$

The state-space system is restricted in two dimensions. First, the errors in the measurement equation enter in an additively separable manner. This implies that the conditional density $p(y_t|s_t,\theta)$ is given by $p_u(y_t - \Psi(s_t,t;\theta)|\theta)$, where $p_u(\cdot|\theta)$ is the pdf associated with the measurement error distribution $F_u(\cdot;\theta)$. In the absence of measurement errors, the distribution $y_t|(s_t,\theta)$ is a pointmass at $\Psi(s_t,t;\theta)$. Second, the state-transition equation has a first-order Markov structure.[an] Owing to the first-order Markov structure of the state-transition equation, neither the states $s_{t-2},s_{t-3},\ldots$ nor the observations $y_{t-1},y_{t-2},\ldots$ provide any additional information about s_t conditional on s_{t-1}. Thus,

[an] Additional lags of the state vector could be easily incorporated using a companion form representation of the state vector as in (114).

$$p(s_t|s_{t-1},\theta)=p(s_t|s_{t-1},S_{1:t-2},\theta)=p(s_t|s_{t-1},S_{1:t-2},Y_{1:t-1},\theta). \tag{150}$$

For the linearized DSGE model of Section 8.1 with normally distributed measurement errors $u_t \sim N(0,\Sigma_u)$ the conditional distributions are given by $s_t|(s_{t-1},\theta) \sim N\left(\Phi_1 s_{t-1}, \Phi_\epsilon \Phi'_\epsilon\right)$ and $y_t|(s_t,\theta) \sim N(\Psi_0 + \Psi_1 s_t, \Sigma_u)$.

10.1 A Generic Filter

We now describe a generic filter that can be used to recursively compute the conditional distributions $p(s_t|Y_{1:t},\theta)$ and $p(y_t|Y_{1:t-1},\theta)$, starting from an initialization $p(s_0|\theta)$. The distributions $p(s_t|Y_{1:t},\theta)$ are a by-product of the algorithm and summarize the information about the state s_t conditional on the current and past observations $Y_{1:t}$, which may be of independent interest. The sequence of predictive distributions $p(y_t|Y_{1:t-1},\theta)$, $t=1,\ldots,T$, can be used to obtain the likelihood function, which can be factorized as follows

$$p(Y_{1:T}|\theta)=\prod_{t=1}^{T} p(y_t|Y_{1:t-1},\theta). \tag{151}$$

The filter is summarized in Algorithm 5. In the description of the filter we drop the parameter θ from the conditioning set to simplify the notation.

Algorithm 5 (Generic Filter). Let $p(s_0) = p(s_0|Y_{1:0})$ be the initial distribution of the state. For $t = 1$ to T:

1. Forecasting t given $t-1$:

(a) Transition equation:

$$p(s_t|Y_{1:t-1}) = \int p(s_t|s_{t-1}, Y_{1:t-1})p(s_{t-1}|Y_{1:t-1})ds_{t-1}$$

(b) Measurement equation:

$$p(y_t|Y_{1:t-1}) = \int p(y_t|s_t, Y_{1:t-1})p(s_t|Y_{1:t-1})ds_t$$

2. Updating with Bayes Theorem. Once y_t becomes available:

$$p(s_t|Y_{1:t}) = p(s_t|y_t, Y_{1:t-1}) = \frac{p(y_t|s_t, Y_{1:t-1})p(s_t|Y_{1:t-1})}{p(y_t|Y_{1:t-1})}.$$

10.2 Likelihood Function for a Linearized DSGE Model

For illustrative purposes, consider the prototypical DSGE model. Owing to the simple structure of the model, we can use (69), (70), (72), and (73) to solve for the latent shocks ϕ_t, λ_t, z_t, and $\epsilon_{R,t}$ as a function of $\hat{x}_t$, $\widehat{lsh}_t$, $\hat{\pi}_t$, and $\hat{R}_t$. Thus, we can deduce from (78) and the definition of s_t that conditional on $\hat{x}_0$, the states s_t can be uniquely inferred from the observables y_t in a recursive manner, meaning that the conditional distributions

$p(s_t|Y_{1:t}, \hat{x}_0)$ are degenerate. Thus, the only uncertainty about the state stems from the initial condition.

Suppose that we drop the labor share and the interest rates from the definition of y_t. In this case it is no longer possible to uniquely determine s_t as a function of y_t and $\hat{x}_0$, because we only have two equations, (69) and (72), and four unknowns. The filter in Algorithm 5 now essentially solves an underdetermined system of equations, taking into account the probability distribution of the four hidden processes. For our linearized DSGE model with Gaussian innovations, all the distributions that appear in Algorithm 5 are Gaussian. In this case the Kalman filter can be used to compute the means and covariance matrices of these distributions recursively. To complete the model specification, we make the following distributional assumptions about the initial state s_0:

$$s_0 \sim N\left(\bar{s}_{0|0}, P_{0|0}\right).$$

In stationary models it is common to set $\bar{s}_{0|0}$ and $P_{0|0}$ equal to the unconditional first and second moments of the invariant distribution associated with the law of motion of s_t in (76). The four conditional distributions in the description of Algorithm 5 for a linear Gaussian state-space model are summarized in Table 6. Detailed derivations can be found in textbook treatments of the Kalman filter and smoother, eg, Hamilton (1994) or Durbin and Koopman (2001).

To illustrate the Kalman filter algorithm, we simulate $T = 50$ observations from the stylized DSGE model conditional on the parameters in Table 5. The two left panels of Fig. 24 depict the filtered shock processes ϕ_t and z_t based on observations of only output growth, which are defined as $\mathbb{E}[s_t|Y_{1:t}]$. The bands delimit 90% credible intervals which are centered around the filtered estimates and based on the standard deviations $\sqrt{\mathbb{V}[s_t|Y_{1:t}]}$. The information in the output growth series is not sufficient to generate a precise estimate of the preference shock process ϕ_t, which, according to the forecast

Table 6 Conditional distributions for the Kalman filter

	Distribution	Mean and variance
$s_{t-1}\|Y_{1:t-1}$	$N\left(\bar{s}_{t-1\|t-1}, P_{t-1\|t-1}\right)$	Given from Iteration $t - 1$
$s_t\|Y_{1:t-1}$	$N\left(\bar{s}_{t\|t-1}, P_{t\|t-1}\right)$	$\bar{s}_{t\|t-1} = \Phi_1 \bar{s}_{t-1\|t-1}$ $P_{t\|t-1} = \Phi_1 P_{t-1\|t-1} \Phi_1' + \Phi_\epsilon \Sigma_\epsilon \Phi_\epsilon'$
$y_t\|Y_{1:t-1}$	$N\left(\bar{y}_{t\|t-1}, F_{t\|t-1}\right)$	$\bar{y}_{t\|t-1} = \Psi_0 + \Psi_1 \bar{s}_{t\|t-1}$ $F_{t\|t-1} = \Psi_1 P_{t\|t-1} \Psi_1' + \Sigma_u$
$s_t\|Y_{1:t}$	$N\left(\bar{s}_{t\|t}, P_{t\|t}\right)$	$\bar{s}_{t\|t} = \bar{s}_{t\|t-1} + P_{t\|t-1} \Psi_1' F_{t\|t-1}^{-1} (y_t - \bar{y}_{t\|t-1})$ $P_{t\|t} = P_{t\|t-1} - P_{t\|t-1} \Psi_1' F_{t\|t-1}^{-1} \Psi_1 P_{t\|t-1}$
$s_t\|(S_{t+1:T}, Y_{1:T})$	$N\left(\bar{s}_{t\|t+1}, P_{t\|t+1}\right)$	$\bar{s}_{t\|t+1} = \bar{s}_{t\|t} + P_{t\|t} \Phi_1' P_{t+1\|t}^{-1} (s_{t+1} - \Phi_1 \bar{s}_{t\|t})$ $P_{t\|t+1} = P_{t\|t} - P_{t\|t} \Phi_1' P_{t+1\|t}^{-1} \Phi_1 P_{t\|t}$

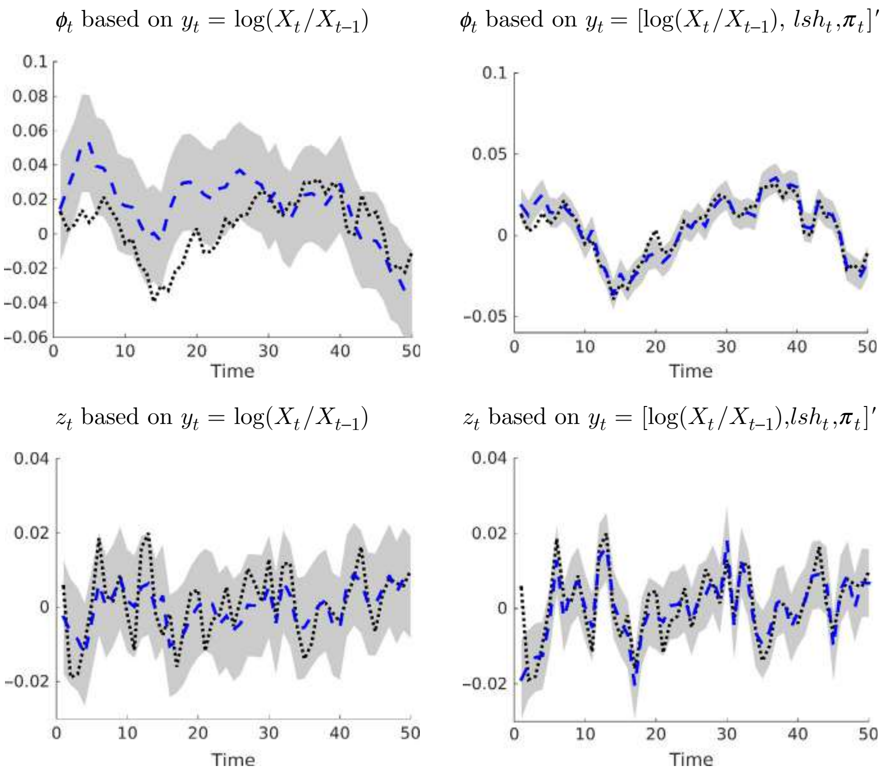


Fig. 24 Filtered states. *Notes*: The filtered states are based on a simulated sample of $T = 50$ observations. Each panel shows the true state s_t (*dotted*), the filtered state $\mathbb{E}[s_t|Y_{1:t}]$ (*dashed*), and 90% credible bands based on $p(s_t|Y_{1:t})$ (*grey area*).

error variance decomposition in Fig. 17, only explains a small fraction of the variation in output growth. The two right panels of Fig. 24 show what happens to the inference about the hidden states if inflation and labor share are added to the set of observables. Conditional on the three series, it is possible to obtain fairly sharp estimates of both the preference shock ϕ_t and the technology growth shock z_t.

Instead of using the Kalman filter, in a linearized DSGE model with Gaussian innovations it is possible to characterize the joint distribution of the observables directly. Let Y be a $T \times n_y$ matrix composed of rows y_t'. Then the joint distribution of Y is given by

$$vec(Y)|\theta \sim N\left(I \otimes \Phi_0(\theta),\ \begin{bmatrix} \Gamma_{yy}(0|\theta) & \Gamma_{yy}(1|\theta) & \ldots & \Gamma_{yy}(T-1|\theta) \\ \Gamma'_{yy}(1|\theta) & \Gamma_{yy}(0|\theta) & \ldots & \Gamma_{yy}(T-2|\theta) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma'_{yy}(T-1|\theta) & \Gamma'_{yy}(T-2|\theta) & \ldots & \Gamma_{yy}(0|\theta) \end{bmatrix} \right). \tag{152}$$

The evaluation of the likelihood function requires the calculation of the autocovariance sequence and the inversion of an $n_y T \times n_y T$ matrix. For large T the joint density can be approximated by the so-called Whittle likelihood function

$$p_W(Y|\theta) \propto \left(\prod_{j=0}^{T-1} \left| 2\pi f_{yy}^{-1}(\omega_j|\theta) \right| \right)^{1/2} \exp\left\{ -\frac{1}{2} \sum_{j=0}^{T-1} tr\left[f_{yy}^{-1}(\omega_j|\theta) \hat{f}_{yy}(\omega_j) \right] \right\} \tag{153}$$

where $f_{yy}(\omega|\theta)$ is the DSGE model-implied spectral density, $\hat{f}_{yy}(\omega)$ is the sample periodogram, and the ω_j's are the fundamental frequencies. The attractive feature of this likelihood function is that the researcher can introduce weights for the different frequencies, and, for instance, only consider business cycle frequencies in the construction of the likelihood function. For the estimation of DSGE models, the Whittle likelihood has been used, for instance, by Christiano and Vigfusson (2003), Qu and Tkachenko (2012), and Sala (2015).

10.3 Likelihood Function for Nonlinear DSGE Models

If the DSGE model is solved using a nonlinear approximation technique, then either the state-transition equation, or the measurement equation, or both become nonlinear. As a consequence, analytical representations of the densities $p(s_{t-1}|Y_{1:t-1})$, $p(s_t|Y_{1:t-1})$, and $p(y_t|Y_{1:t-1})$ that appear in Algorithm 5 are no longer available. While there exists a large literature on nonlinear filtering (see for instance Crisan and Rozovsky, 2011) we focus on the class of particle filters. Particle filters belong to the class of sequential Monte Carlo algorithms. The basic idea is to approximate the distribution $s_t|Y_{1:t}$ through a swarm of particles $\{s_t^j, W_t^j\}_{j=1}^M$ such that

$$\begin{aligned} \bar{h}_{t,M} &= \frac{1}{M}\sum_{j=1}^{M} h(s_t^j) W_t^j \xrightarrow{a.s.} \mathbb{E}[h(s_t)|Y_{1:t}], \\ \sqrt{M}(\bar{h}_{t,M} &- \mathbb{E}[h(s_t)|Y_{1:t}]) \Longrightarrow N(0, \Omega_t[h]), \end{aligned} \tag{154}$$

where $\Longrightarrow$ denotes convergence in distribution.[ao] Here the s_t^j's are particle values and the W_t^j's are the particle weights. The conditional expectation of $h(s_t)$ is approximated by a weighted average of the (transformed) particles $h(s_t^j)$. Under suitable regularity conditions, the Monte Carlo approximation satisfies an SLLN and a CLT. The covariance matrix $\Omega_t[h]$ characterizes the accuracy of the Monte Carlo approximation. Setting $h(s_t) = p(y_{t+1}|s_t)$ yields the particle filter approximation of the likelihood increment $p(y_{t+1}|Y_{1:t}) = \mathbb{E}[p(y_{t+1}|s_t)|Y_{1:t}]$. Each iteration of the filter manipulates the particle values and weights to recursively track the sequence of conditional distributions $s_t|Y_{1:t}$. The paper by Fernández-Villaverde and Rubio-Ramírez (2007) was the first to

[ao] A sequence of random variables X_T converges in distribution to a random variable X if for every measurable and bounded function $f(\cdot)$ that is continuous almost everywhere $\mathbb{E}[f(X_T)] \to \mathbb{E}[f(X)]$.

approximate the likelihood function of a nonlinear DSGE model using a particle filter and many authors have followed this approach.

Particle filters are widely used in engineering and statistics. Surveys and tutorials are provided, for instance, in Arulampalam et al. (2002), Cappé et al. (2007), Doucet and Johansen (2011), and Creal (2012). The basic bootstrap particle filter algorithm is remarkably straightforward, but may perform quite poorly in practice. Thus, much of the literature focuses on refinements of the bootstrap filter that increases the efficiency of the algorithm; see, for instance, Doucet et al. (2001). Textbook treatments of the statistical theory underlying particle filters can be found in Liu (2001), Cappé et al. (2005), and Del Moral (2013).

10.3.1 Generic Particle Filter

The subsequent exposition draws from Herbst and Schorfheide (2015), who provide a detailed presentation of particle filtering techniques in the context of DSGE model applications as well as a more extensive literature survey. In the basic version of the particle filter, the time t particles are generated based on the time $t-1$ particles by simulating the state-transition equation forward. The particle weights are then updated based on the likelihood of the observation y_t under the s_t^j particle, $p(y_t|s_t^j)$. The more accurate the prediction of y_t based on s_t^j, the larger the density $p(y_t|s_t^j)$, and the larger the relative weight that will be placed on particle j. However, the naive forward simulation ignores information contained in the current observation y_t and may lead to a very uneven distribution of particle weights, in particular, if the measurement error variance is small or if the model has difficulties explaining the period t observation in the sense that for most particles s_t^j the actual observation y_t lies far in the tails of the model-implied distribution of $y_t|s_t^j$. The particle filter can be generalized by allowing s_t^j in the forecasting step to be drawn from a generic importance sampling density $g_t(\cdot|s_{t-1}^j)$, which leads to the following algorithm:[ap]

Algorithm 6 (Generic Particle Filter).

1. **Initialization.** Draw the initial particles from the distribution $s_0^j \overset{iid}{\sim} p(s_0)$ and set $W_0^j = 1$, $j = 1,\ldots,M$.
2. **Recursion.** For $t = 1,\ldots,T$:
   (a) **Forecasting** s_t**.** Draw $\tilde{s}_t^j$ from density $g_t(\tilde{s}_t|s_{t-1}^j)$ and define the importance weights

$$\omega_t^j = \frac{p(\tilde{s}_t^j|s_{t-1}^j)}{g_t(\tilde{s}_t^j|s_{t-1}^j)}. \tag{155}$$

An approximation of $\mathrm{E}[h(s_t)|Y_{1:t-1}]$ is given by

[ap] To simplify the notation, we omit θ from the conditioning set.

$$\hat{h}_{t,M} = \frac{1}{M}\sum_{j=1}^{M} h(\tilde{s}_t^j)\omega_t^j W_{t-1}^j. \tag{156}$$

(b) Forecasting y_t. Define the incremental weights

$$\tilde{w}_t^j = p(y_t|\tilde{s}_t^j)\omega_t^j. \tag{157}$$

The predictive density $p(y_t|Y_{1:t-1})$ can be approximated by

$$\hat{p}(y_t|Y_{1:t-1}) = \frac{1}{M}\sum_{j=1}^{M} \tilde{w}_t^j W_{t-1}^j. \tag{158}$$

(c) Updating. Define the normalized weights

$$\tilde{W}_t^j = \frac{\tilde{w}_t^j W_{t-1}^j}{\frac{1}{M}\sum_{j=1}^{M} \tilde{w}_t^j W_{t-1}^j}. \tag{159}$$

An approximation of $\mathbb{E}[h(s_t)|Y_{1:t},\theta]$ is given by

$$\tilde{h}_{t,M} = \frac{1}{M}\sum_{j=1}^{M} h(\tilde{s}_t^j)\tilde{W}_t^j. \tag{160}$$

(d) Selection. Resample the particles via multinomial resampling. Let $\{s_t^j\}_{j=1}^M$ denote M *iid* draws from a multinomial distribution characterized by support points and weights $\{\tilde{s}_t^j, \tilde{W}_t^j\}$ and set $W_t^j = 1$ for $j =,1\ldots, M$. An approximation of $\mathbb{E}[h(s_t)|Y_{1:t},\theta]$ is given by

$$\bar{h}_{t,M} = \frac{1}{M}\sum_{j=1}^{M} h(s_t^j)W_t^j. \tag{161}$$

3. **Likelihood Approximation.** The approximation of the log likelihood function is given by

$$\log\hat{p}(Y_{1:T}|\theta) = \sum_{t=1}^{T}\log\left(\frac{1}{M}\sum_{j=1}^{M}\tilde{w}_t^j W_{t-1}^j\right). \tag{162}$$

Conditional on the stage $t-1$ weights W_{t-1}^j the accuracy of the approximation of the likelihood increment $p(y_t|Y_{1:t-1})$ depends on the variability of the incremental weights $\tilde{\omega}_t^j$ in (157). The larger the variance of the incremental weights, the less accurate the particle filter approximation of the likelihood function. In this regard, the most important choice for the implementation of the particle filter is the choice of the proposal distribution $g_t(\tilde{s}_t^j|s_{t-1}^j)$, which is discussed in more detail below.

The selection step is included in the filter to avoid a degeneracy of particle weights. While it adds additional noise to the Monte Carlo approximation, it simultaneously equalizes the particle weights, which increases the accuracy of subsequent approximations. In the absence of the selection step, the distribution of particle weights would become more uneven from iteration to iteration. The selection step does not have to be executed in every iteration. For instance, in practice, users often apply a threshold rule according to which the selection step is executed whenever the following measure falls below a threshold, eg, 25% or 50% of the nominal number of particles:

$$\widehat{ESS}_t = M \Big/ \left(\frac{1}{M} \sum_{j=1}^{M} (\tilde{W}_t^j)^2 \right). \tag{163}$$

The effective sample size $\widehat{ESS}_t$ (in terms of number of particles) captures the variance of the particle weights. It is equal to M if $\tilde{W}_t^j = 1$ for all j and equal to 1 if one of the particles has weight M and all others have weight 0. The resampling can be executed with a variety of algorithms. We mention multinomial resampling in the description of Algorithm 6. Multinomial resampling is easy to implement and satisfies a CLT. However, there are more efficient algorithms (meaning they are associated with a smaller Monte Carlo variance), such as stratified or systematic resampling. A detailed textbook treatment can be found in Liu (2001) and Cappé et al. (2005).

10.3.2 Bootstrap Particle Filter

The bootstrap particle filter draws $\tilde{s}_t^j$ from the state-transition equation and sets

$$g_t(\tilde{s}_t^j | s_{t-1}^j) = p(\tilde{s}_t^j | s_{t-1}^j). \tag{164}$$

This implies that $\omega_t^j = 1$ and the incremental weight is given by the likelihood $p(y_t | \tilde{s}_t^j)$, which unfortunately may be highly variable. Fig. 25 provides an illustration of the bootstrap particle filter with $M = 100$ particles using the same experimental design as for the Kalman filter in Section 10.2. The observables are output growth, labor share, and inflation and the observation equation is augmented with measurement errors. The measurement error variance amounts to 10% of the total variance of the simulated data. Because the stylized DSGE is loglinearized, the Kalman filter provides exact inference and any discrepancy between the Kalman and particle filter output reflects the approximation error of the particle filter. In this application the particle filter approximations are quite accurate even with a small number of particles. The particle filtered states z_t and $\epsilon_{R,t}$ appear to be more volatile than the exactly filtered states from the Kalman filter.

Fig. 26 illustrates the accuracy of the likelihood approximation. The left panel compares log-likelihood increments $\log p(y_t | Y_{1:t-1}, \theta)$ obtained from the Kalman filter and a single run of the particle filter. The left panel shows the distribution of the approximation errors of the log-likelihood function: $\log \hat{p}(Y_{1:T} | \theta) - \log p(Y_{1:T} | \theta)$. It has been shown,

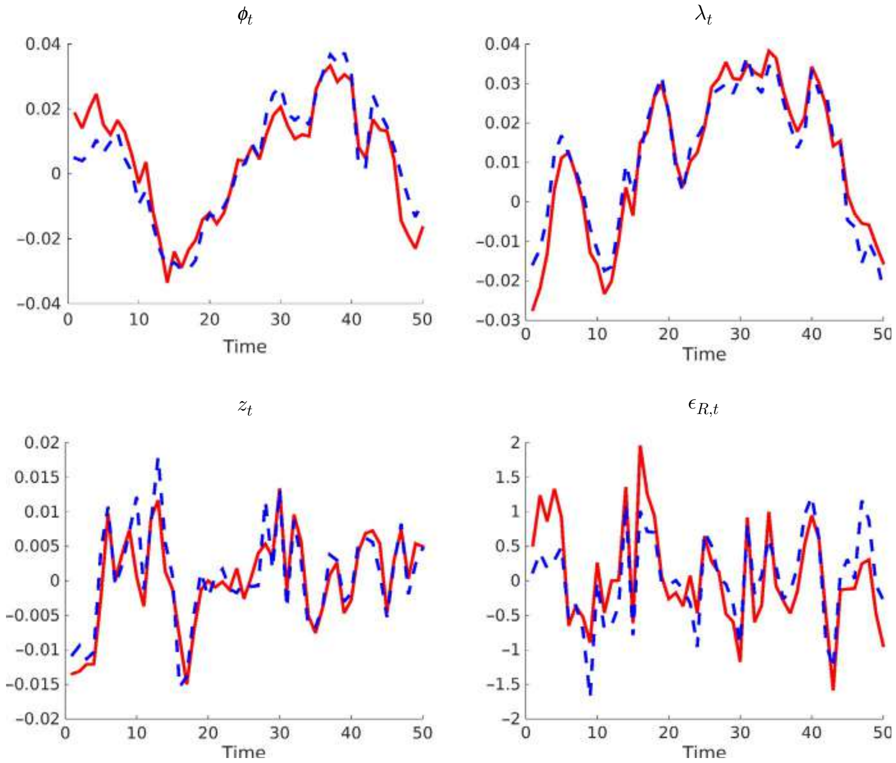


Fig. 25 Particle-filtered states. *Notes*: We simulate a sample of $T = 50$ observations y_t and states s_t from the stylized DSGE model. The four panels compare filtered states from the Kalman filter (*solid*) and a single run of the particle filter (*dashed*) with $M = 100$ particles. The observables used for filtering are output growth, labor share, and inflation. The measurement error variances are 10% of the total variance of the data.

eg, by Del Moral (2004) and Pitt et al. (2012), that the particle filter approximation of the likelihood function is unbiased, which implies that the approximation of the *log*-likelihood function has a downward bias, which is evident in the figure. Under suitable regularity conditions the particle filter approximations satisfy a CLT. The figure clearly indicates that the distribution of the approximation errors becomes more concentrated as the number of particles is increased from $M = 100$ to $M = 500$.

The accuracy of the bootstrap particle filter crucially depends on the quality of the fit of the DSGE model and the magnitude of the variance of the measurement errors u_t. Recall that for the bootstrap particle filter, the incremental weights $\tilde{w}_t^j = p(y_t|\tilde{s}_t^j)$. If the model fits poorly, then the one-step-ahead predictions conditional on the particles $\tilde{s}_t^j$ are inaccurate and the density of the actual observation y_t falls far in the tails of the predictive distribution. Because the density tends to decay quickly in the tails, the

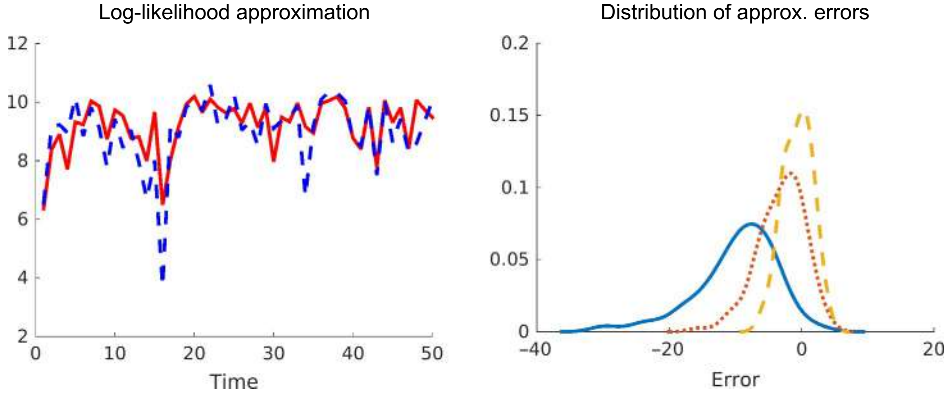


Fig. 26 Particle-filtered log-likelihood. *Notes*: We simulate a sample of $T = 50$ observations y_t and states s_t from the stylized DSGE model. The *left panel* compares log-likelihood increments from the Kalman filter (*solid*) and a single run of the particle filter (*dashed*) with $M = 100$ particles. The *right panel* shows a density plot for approximation errors of $\log \hat{p}(Y_{1:T}|\theta) - \log p(Y_{1:T}|\theta)$ based on $N_{run} = 100$ repetitions of the particle filter for $M = 100$ (*solid*), $M = 200$ (*dotted*), and $M = 500$ (*dashed*) particles. The measurement error variances are 10% of the total variance of the data.

incremental weights will have a high variability, which means that Monte Carlo approximations based on these incremental weights will be inaccurate.

The measurement error defines a metric between the observation y_t and the conditional mean prediction $\Psi(s_t,t;\theta)$. Consider the extreme case in which the measurement error is set to zero. This means that any particle that does not predict y_t exactly would get weight zero. In a model in which the error distribution is continuous, the probability of drawing a $\tilde{s}_t^j$ that receives a nonzero weight is zero, which means that the algorithm would fail in the first iteration. By continuity, the smaller the measurement error variance, the smaller the number of particles that would receive a nontrivial weight, and the larger the variance of the approximation error of particle filter approximations. In practice, it is often useful to start the filtering with a rather large measurement error variance, eg, 10% or 20% of the variance of the observables, and then observing the accuracy of the filter as the measurement error variance is reduced.

10.3.3 (Approximately) Conditionally Optimal Particle Filter

The conditionally optimal particle filter sets

$$g_t(\tilde{s}_t|s_{t-1}^j) = p(\tilde{s}_t|y_t, s_{t-1}^j), \tag{165}$$

that is, $\tilde{s}_t$ is sampled from the posterior distribution of the period t state given (y_t, s_{t-1}^j). In this case

$$\tilde{w}_t^j = \int p(y_t|s_t)p(s_t|s_{t-1}^j)ds_t = p(y_t|s_{t-1}^j). \tag{166}$$

Unfortunately, in a typical nonlinear DSGE model applications it is not possible to sample directly from $p(\tilde{s}_t|y_t, s^j_{t-1})$. In this case the researcher could try to approximate the conditionally optimal proposal density, which leads to an *approximately conditionally optimal particle filter*. For instance, if the DSGE model's nonlinearity arises from a higher-order perturbation solution and the nonlinearities are not too strong, then an approximately conditionally optimal importance distribution could be obtained by applying the one-step Kalman filter updating described in Table 6 to the first-order approximation of the DSGE model. More generally, as suggested in Guo et al. (2005), one could use the updating steps of a conventional nonlinear filter, such as an extended Kalman filter, unscented Kalman filter, or a Gaussian quadrature filter, to construct an efficient proposal distribution. Approximate filters for nonlinear DSGE models have been developed by Andreasen (2013) and Kollmann (2015).

Whenever one uses a proposal distribution that differs from $p(\tilde{s}^j_t|s^j_{t-1})$ it becomes necessary to evaluate the density $p(\tilde{s}^j_t|s^j_{t-1})$. In DSGE model applications, one typically does not have a closed-form representation for this density. It is implicitly determined by the distribution of ϵ_t and the state transition $\Phi(s_{t-1},\epsilon_t)$. The problem of having to evaluate the DSGE model-implied density of $\tilde{s}^j_t$ can be avoided by sampling an innovation from a proposal density $g^\epsilon(\tilde{\epsilon}_t|s^j_{t-1})$ and defining $\tilde{s}^j_t = \Phi(s^j_{t-1}, \tilde{\epsilon}_t)$. In this case the particle weights can be updated by the density ratio

$$\omega^j_t = \frac{p^\epsilon(\tilde{\epsilon}^j_t)}{g_t(\tilde{\epsilon}^j_t|s^j_{t-1})}, \tag{167}$$

where $p^\epsilon(\cdot)$ is the model-implied pdf of the innovation ϵ_t.

Sometimes, DSGE models have a specific structure that may simplify the particle-filter-based likelihood approximation. In models that are linear conditional on a subset of state variables, eg, volatility states or Markov-switching regimes, it is possible to use the Kalman filter to represent the uncertainty about a subset of states. In models in which the number of shocks ϵ_t equals the number of observables y_t, it might be possible (in the absence of measurement errors) conditional on an initial state vector s_0 to directly solve for ϵ_t based on y_t and s_{t-1}, which means that it may be possible to evaluate the likelihood function $p(Y_{1:T}|\theta,s_0)$ recursively. A more detailed discussion of these and other issues related to particle filtering for DSGE models is provided in Herbst and Schorfheide (2015).

11. FREQUENTIST ESTIMATION TECHNIQUES

We will now consider four frequentist inference techniques in more detail: likelihood-based estimation (Section 11.1), simulated method of moments estimation (Section 11.2), impulse response function matching (Section 11.3), and GMM estimation (Section 11.4). All of these econometric techniques, with the exception of the impulse response function matching approach, are widely used in other areas of economics and are associated with extensive literatures that we will not do justice to in this section. We will sketch the main idea behind each of the econometric procedures and then focus on adjustments that have been proposed to tailor the techniques to DSGE model applications. Each estimation method is associated with a model evaluation procedure that essentially assesses the extent to which the estimation objective has been achieved.

11.1 Likelihood-Based Estimation

Under the assumption that the econometric model is well specified, likelihood-based inference techniques enjoy many optimality properties. Because DSGE models deliver a joint distribution for the observables, maximum likelihood estimation of θ is very appealing. The maximum likelihood estimator $\hat{\theta}_{ml}$ was defined in (134). Altug (1989) and McGrattan (1994) are early examples of papers that estimated variants of a neoclassical stochastic growth model by maximum likelihood, whereas Leeper and Sims (1995) estimated a DSGE model meant to be usable for monetary policy analysis.

Even in a loglinearized DSGE model, the DSGE model parameters θ enter the coefficients of the state-space representation in a nonlinear manner, which can be seen in Table 4. Thus, a numerical technique is required to maximize the likelihood function. A textbook treatment of numerical optimization routines can be found, for instance, in Judd (1998) and Nocedal and Wright (2006). Some algorithms, eg, Quasi-Newton methods, rely on the evaluation of the gradient of the objective function (which requires differentiability), and other methods, such as simulated annealing, do not. This distinction is important if the likelihood function is evaluated with a particle filter. Without further adjustments, particle filter approximations of the likelihood function are nondifferentiable in θ even if the exact likelihood function is. This issue and possible solutions are discussed, for instance, in Malik and Pitt (2011) and Kantas et al. (2014).

11.1.1 Textbook Analysis of the ML Estimator

Under the assumption that θ is well identified and the log-likelihood function is sufficiently smooth with respect to θ, confidence intervals and test statistics for the DSGE model parameters can be based on a large sample approximation of the sampling distribution of the ML estimator. A formal analysis in the context of state-space models is provided, for instance, in the textbook by Cappé et al. (2005). We sketch the main steps of the approximation, assuming that the DSGE model is correctly specified and the data are generated by $p(Y|\theta_0,M_1)$. Of course, this analysis could be generalized to a setting in which the DSGE model is misspecified and the data are generated by a reference model $p(Y|M_0)$. In this case the resulting estimator is called quasi-maximum-likelihood estimator and the formula for the asymptotic covariance matrix presented below would have to be adjusted. A detailed treatment of quasi-likelihood inference is provided in White (1994).

Recall from Section 10 that the log-likelihood function can be decomposed as follows:

$$\ell_T(\theta|Y) = \sum_{t=1}^{T} \log p(y_t|Y_{1:t-1},\theta) = \sum_{t=1}^{T} \log \int p(y_t|s_t,\theta)p(s_t|Y_{1:t-1})ds_t. \tag{168}$$

Owing to the time-dependent conditioning information $Y_{1:t-1}$ the summands are not stationary. However, under the assumption that the sequence $\{s_t,y_t\}$ is stationary if initialized in the infinite past, one can approximate the log-likelihood function by

$$\ell^s_T(\theta|Y) = \sum_{t=1}^{T} \log \int p(y_t|s_t,\theta)p(s_t|Y_{-\infty:t-1})ds_t, \tag{169}$$

and show that the discrepancy $\left|\ell_T(\theta|Y) - \ell^s_T(\theta|Y)\right|$ becomes negligible as $T \to \infty$. The ML estimator is consistent if $T^{-1}\ell^s_T(\theta|Y) \xrightarrow{a.s.} \ell^s(\theta)$ uniformly almost surely (a.s.), where $\ell^s(\theta)$ is deterministic and maximized at the "true" θ_0. The consistency can be stated as

$$\hat{\theta}_{ml} \xrightarrow{a.s.} \theta_0. \tag{170}$$

Frequentist asymptotics rely on a second-order approximation of the log-likelihood function. Define the score (vector of first derivatives) $\nabla_\theta \ell^s_T(\theta|Y)$ and the matrix of second derivatives (Hessian, multiplied by minus one) $-\nabla^2_\theta \ell^s_T(\theta|Y)$ and let

$$\begin{aligned}\ell^s_T(\theta|Y) = {} & \ell^s_T(\theta_0|Y) + T^{-1/2}\nabla_\theta \ell^s_T(\theta_0|Y)\sqrt{T}(\theta-\theta_0) \\ & + \frac{1}{2}\sqrt{T}(\theta-\theta_0)'\left[\nabla^2_\theta \ell^s_T(\theta_0|Y)\right]\sqrt{T}(\theta-\theta_0) + \text{small}.\end{aligned}$$

If the maximum is attained in the interior of the parameter space Θ, the first-order conditions can be approximated by

$$\sqrt{T}(\hat{\theta}_{ml} - \theta_0) = \left[-\nabla^2_\theta \ell^s_T(\theta_0|Y)\right]^{-1} T^{-1/2}\nabla_\theta \ell^s_T(\theta_0|Y) + \text{small}. \tag{171}$$

Under suitable regularity conditions, the score process satisfies a CLT:

$$T^{-1/2}\nabla_\theta \ell_T(\theta|Y) \Longrightarrow N(0,\mathcal{I}(\theta_0)), \tag{172}$$

where $\mathcal{I}(\theta_0)$ is the Fisher information matrix.[aq] As long as the likelihood function is correctly specified, the term $\| -\nabla^2_\theta \ell_T(\theta|Y) - \mathcal{I}(\theta_0) \|$ converges to zero uniformly in a neighborhood around θ_0, which is a manifestation of the so-called information matrix equality. This leads to the following result

$$\sqrt{T}(\hat{\theta}_{ml} - \theta_0) \Longrightarrow N\left(0,\mathcal{I}^{-1}(\theta_0)\right). \tag{173}$$

[aq] The formal definition of the information matrix for this model is delicate and therefore omitted.

Thus, standard error estimates for t-tests and confidence intervals for elements of the parameter vector θ can be obtained from the diagonal elements of the inverse Hessian $[-\nabla^2_\theta \ell_T(\theta|Y)]^{-1}$ of the log-likelihood function evaluated at the ML estimator.[ar] Moreover, the maximized likelihood function can be used to construct textbook Wald, Lagrange-multiplier, and likelihood ratio statistics. Model selection could be based on a penalized likelihood function such as the Schwarz (1978) information criterion.

11.1.2 Illustration

To illustrate the behavior of the ML estimator we repeatedly generate data from the stylized DSGE model, treating the values listed in Table 5 as "true" parameters. We fix all parameters except for the Calvo parameter ζ_p at their "true" values and use the ML approach to estimate ζ_p. The likelihood function is based on output growth, labor share, inflation, and interest rate data. The left panel of Fig. 27 depicts the likelihood function for a single simulated data set Y. The right panel shows the sampling distribution of $\hat{\zeta}_{p,ml}$, which is approximated by repeatedly generating data and evaluating the ML estimator. The sampling distribution peaks near the "true" parameter value and becomes more concentrated as the sample size is increased from $T = 80$ to $T = 200$.

In practice, the ML estimator is rarely as well behaved as in this illustration, because the maximization is carried out over a high-dimensional parameter space and the log-likelihood function may be highly nonelliptical. In the remainder of this subsection, we focus on two obstacles that arise in the context of the ML estimation of DSGE models.

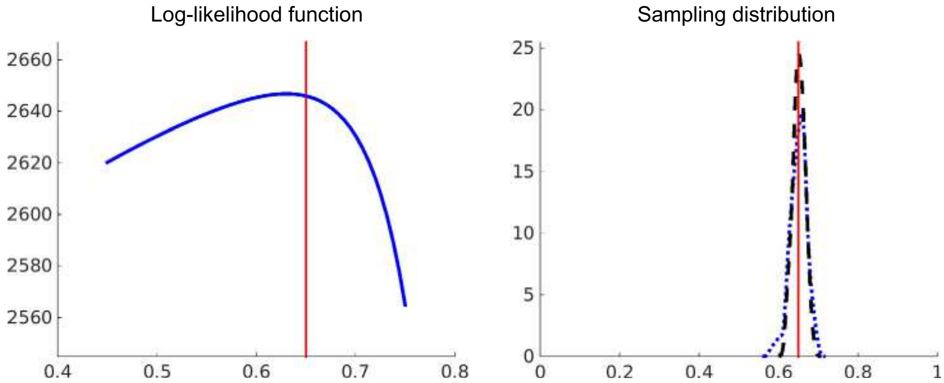


Fig. 27 Log-likelihood function and sampling distribution of $\hat{\zeta}_{p,ml}$. *Notes*: *Left panel*: log-likelihood function $\ell_T(\zeta_p|Y)$ for a single data set of size $T = 200$. *Right panel*: We simulate samples of size $T = 80$ (*dotted*) and $T = 200$ (*dashed*) and compute the ML estimator for the Calvo parameter ζ_p. All other parameters are fixed at their "true" value. The plot depicts densities of the sampling distribution of $\hat{\zeta}_p$. The vertical lines in the two panels indicate the "true" value of ζ_p.

[ar] Owing to the Information Matrix Equality, the standard error estimates can also be obtained from the outer product of the score: $\sum_{t=1}^T (\nabla_\theta \log p(y_t|Y_{1:t-1},\theta))(\nabla_\theta \log p(y_t|Y_{1:t-1},\theta))'$.

The first obstacle is the potential stochastic singularity of the DSGE model-implied conditional distribution of y_t given its past. The second obstacle is caused by a potential lack of identification of the DSGE model parameters.

11.1.3 Stochastic Singularity

Imagine removing all shocks except for the technology shock from the stylized DSGE model, while maintaining that y_t comprises output growth, the labor share, inflation, and the interest rate. In this case, we have one exogenous shock and four observables, which implies, among other things, that the DSGE model places probability one on the event that

$$\beta \log R_t - \log \pi_t = \beta \log (\pi^* \gamma / \beta) - \log \pi^*.$$

Because in the actual data $\beta \log R_t - \log \pi_t$ is time varying, the likelihood function is equal to zero and not usable for inference. The literature has adopted two types of approaches to address the singularity, which we refer to as the "measurement error" approach and the "more structural shocks" approach.

Under the measurement error approach (78) is augmented by a measurement error process u_t, which in general may be serially correlated. The term "measurement error" is a bit of a misnomer. It tries to blame the discrepancy between the model and the data on the accuracy of the latter rather than the quality of the former. In a typical DSGE model application, the blame should probably be shared by both. A key feature of the "measurement error" approach is that the agents in the model do not account for the presence of u_t when making their decisions. The "measurement error" approach has been particularly popular in the real business cycle literature—it was used, for instance, in Altug (1989). The real business cycle literature tried to explain business cycle fluctuations based on a small number of structural shocks, in particular, technology shocks.

The "more structural shocks" approach augments the DSGE model with additional structural shocks until the number of shocks is equal to or exceeds the desired number of observables stacked in the vector y_t. For instance, if we add the three remaining shock processes ϕ_t, λ_t, $\epsilon_{R,t}$ back into the prototypical DSGE model, then a stochastic singularity is no longer an obstacle for the evaluation of the likelihood function. Of course, at a deeper level, the stochastic singularity problem never vanishes, as we could also increase the dimension of the vector y_t. Because the policy functions in the solution of the DSGE model express the control variables as functions of the state variables, the set of potential observables y_t in any DSGE model exceeds the number of shocks (which are exogenous state variables from the perspective of the underlying agents' optimization problems). Most of the literature that estimates loglinearized DSGE models uses empirical specifications in which the number of exogenous shocks is at least as large as the number of observables. Examples are Schorfheide (2000), Rabanal and Rubio-Ramírez (2005), and Smets and Wouters (2007).

The converse of the "more structural shocks" approach would be a "fewer observables" approach, ie, one restricts the number of observables used in the construction of the likelihood function to the number of exogenous shocks included in the model. This raises the question of which observables to include in the likelihood function, which is discussed in Guerrón-Quintana (2010) and Canova et al. (2014). Qu (2015) proposes to use a composite likelihood to estimate singular DSGE models. A composite likelihood function is obtained by partitioning the vector of observables y_t into subsets, eg, $y_t' = [y_{1,t}', y_{2,t}', y_{3,t}']$ for which the likelihood function is nonsingular, eg, "composite likelihood" and then use the product of marginals $p(Y_{1,1:T}|\theta)p(Y_{2,1:T}|\theta)$ $p(Y_{3,1:T}|\theta)$ as the estimation objective function.

11.1.4 Dealing with Lack of Identification

In many applications it is quite difficult to maximize the likelihood function. This difficulty is in part caused by the presence of local extrema and/or weak curvature in some directions of the parameter space and may be a manifestation of identification problems. One potential remedy that has been widely used in practice is to fix a subset of the parameters at plausible values, where "plausible" means consistent with some empirical observations that are not part of the estimation sample Y. Conditional on the fixed parameters, the likelihood function for the remaining parameters may have a more elliptical shape and therefore may be easier to maximize. Of course, such an approach ignores the uncertainty with respect to those parameters that are being fixed. Moreover, if they are fixed at the "wrong" parameter values, inference about the remaining parameters will be distorted.

Building on the broader literature on identification-robust econometric inference, the recent literature has developed inference methods that remain valid even if some parameters of the DSGE model are only weakly or not at all identified. Guerrón-Quintana et al. (2013) propose a method that relies on likelihood-based estimates of the system matrices of the state-space representation $\hat{\Psi}_0$, $\hat{\Psi}_1$, $\hat{\Phi}_1$ and $\hat{\Phi}_\epsilon$. In view of the identification problems associated with the Ψ and Φ matrices discussed in Section 9.1, their approach requires a reparameterization of the state-space matrices in terms of an identifiable reduced-form parameter vector $\phi = f(\theta)$ that, according to the DSGE model, is a function of θ. In the context of our stylized DSGE model, such a reparameterization could be obtained based on the information in Table 4.

Let M_1^ϕ denote the state-space representation of the DSGE model in terms of ϕ and let $\hat{\phi}$ be the ML estimator of ϕ. The hypothesis $H_0 : \theta = \theta_0$ can be translated into the hypothesis $\phi = f(\theta_0)$ and the corresponding likelihood ratio statistic takes the form

$$LR(Y|\theta_0) = 2\left[\log p(Y|\hat{\phi}, M_1^\phi) - \log p(Y|f(\theta_0), M_1^\phi)\right] \Longrightarrow \chi^2_{\dim(\phi)}. \qquad (174)$$

The degrees of freedom of the χ^2 limit distribution depend on the dimension of ϕ (instead of θ), which means that it is important to reduce the dimension of ϕ as much as possible by using a minimal state-variable representation of the DSGE model solution and to remove elements from the Ψ and Φ matrices that are zero for all values of θ. The likelihood ratio statistic can be inverted to generate a $1-\alpha$ joint confidence set for the vector θ:

$$CS^{\theta}(Y)=\{\theta \mid LR(Y|\theta)\leq \chi^2_{crit}\}, \tag{175}$$

where χ^2_{crit} is the $1-\alpha$ quantile of the $\chi^2_{\dim(\phi)}$ distribution. Subvector inference can be implemented by projecting the joint confidence set on the desired subspace. The inversion of test statistics is computationally tedious because the test statistic has to be evaluated for a wide range of θ values. However, it does not require the maximization of the likelihood function. Guerrón-Quintana et al. (2013) show how the computation of the confidence interval can be implemented based on the output from a Bayesian estimation of the DSGE model.

Andrews and Mikusheva (2015) propose an identification-robust Lagrange multiplier test. The test statistic is based on the score process and its quadratic variation

$$s_{T,t}(\theta)=\nabla_{\theta}\ell(\theta|Y_{1:t})-\nabla_{\theta}\ell(\theta|Y_{1:t-1}),\ \ J_T(\theta)=\sum_{t=1}^{T}s_{T,t}(\theta)s'_{T,t}(\theta)$$

and is defined as

$$LM(\theta|Y)=\nabla'_{\theta}\ell_T(\theta_0|Y)[J_T(\theta_0)]^{-1}\nabla_{\theta}\ell_T(\theta_0|Y)\Longrightarrow\chi^2_{\dim(\theta_0)}. \tag{176}$$

Note that the degrees of freedom of the χ^2 limit distribution now depend on the dimension of the parameter vector θ instead of the vector of identifiable reduced-form coefficients. A confidence set for θ can be obtained by replacing the LR statistic in (175) with the LM statistic. Andrews and Mikusheva (2015) also consider subvector inference based on a profile likelihood function that concentrates out a subvector of well-identified DSGE model parameters. A frequency domain version of the LM test based on the Whittle likelihood function is provided by Qu (2014). Both Andrews and Mikusheva (2015) and Qu (2014) provide detailed Monte Carlo studies to assess the performance of the proposed identification-robust tests.

11.2 (Simulated) Minimum Distance Estimation

Minimum distance (MD) estimation is based on the idea of minimizing the discrepancy between sample moments of the data, which we denoted by $\hat{m}_T(Y)$, and model-implied moments, which we denoted by $\mathbb{E}[\hat{m}_T(Y)|\theta,M_1]$. The MD estimator $\hat{\theta}_{md}$ was defined in (135) and (136). Examples of the sample statistics $\hat{m}_T(Y)$ are the sample autocovariances $\hat{\Gamma}_{yy}(h)$, a smoothed periodogram $\bar{f}_{yy}(\omega)$ as in Diebold et al. (1998), or estimates of the parameters of an approximating model, eg, the VAR(p) in (113) as in Smith (1993). If $\hat{m}_T(Y)$ consists of parameter estimates of a reference model, then the moment-based estimation is also called indirect inference; see Gourieroux et al. (1993). In some cases it is possible to calculate the model-implied moments analytically. For instance, suppose that $\hat{m}_T(Y) = \frac{1}{T}\sum y_t y'_{t-1}$, then we can derive

$$\mathbb{E}[\hat{m}_T(Y)|\theta, M_1] = \frac{1}{T}\sum \mathbb{E}[y_t y'_{t-1}|\theta, M_1] = \mathbb{E}[y_2 y'_1|\theta, M_1] \tag{177}$$

from the state-space representation of a linearized DSGE model. Explict formulae for moments of pruned models solved with perturbation methods are provided by Andreasen et al. (2013) (recall Section 4.4). Alternatively, suppose that $\hat{m}_T(Y)$ corresponds to the OLS estimates of a VAR(1). In this case, even for a linear DSGE model, it is not feasible to compute

$$\mathbb{E}[\hat{m}_T(Y)] = \mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T} y_{t-1}y'_{t-1}\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} y_{t-1}y'_t \middle| \theta, M_1\right]. \tag{178}$$

The model-implied expectation of the OLS estimator has to be approximated, for instance, by a population regression:

$$\hat{\mathbb{E}}[\hat{m}_T(Y)] = \left(\mathbb{E}[y_{t-1}y'_{t-1}|\theta, M_1]\right)^{-1}\mathbb{E}[y_{t-1}y'_t|\theta, M_1], \tag{179}$$

or the model-implied moment function has to be replaced by a simulation approximation, which will be discussed in more detail below.

11.2.1 Textbook Analysis

We proceed by sketching the asymptotic approximation of the frequentist sampling distribution of the MD estimator. Define the discrepancy

$$G_T(\theta|Y) = \hat{m}_T(Y) - \hat{\mathbb{E}}[\hat{m}_T(Y)|\theta, M_1], \tag{180}$$

such that the criterion function of the MD estimator in (135) can be written as

$$Q_T(\theta|Y) = \|G_T(\theta|Y)\|_{W_T}. \tag{181}$$

Suppose that there is a unique θ_0 with the property that[as]

$$\hat{m}_T(Y) - \mathbb{E}[\hat{m}_T(Y)|\theta_0, M_1] \xrightarrow{a.s.} 0 \tag{182}$$

and that the sample criterion function $Q_T(\theta|Y)$ converges uniformly almost surely to a limit criterion function $Q(\theta)$, then the MD estimator is consistent in the sense that $\hat{\theta}_{md} \xrightarrow{a.s.} \theta_0$.

The analysis of the MD estimator closely mirrors the analysis of the ML estimator, because both types of estimators are defined as the extremum of an objective function.

[as] In some DSGE models a subset of the series included in y_t is nonstationary. Thus, moments are only well defined after a stationarity-inducing transformation has been applied. This problem is analyzed in Gorodnichenko and Ng (2010).

The sampling distribution of $\hat{\theta}_{md}$ can be derived from a second-order approximation of the criterion function $Q_T(\theta|Y)$ around θ_0:

$$\begin{aligned} TQ_T(\theta|Y) &= \sqrt{T}\nabla_\theta Q_T(\theta_0|Y)\sqrt{T}(\theta-\theta_0)' \\ &\quad + \frac{1}{2}\sqrt{T}(\theta-\theta_0)'\left[\frac{1}{T}\nabla_\theta^2 Q_T(\theta_0|Y)\right]\sqrt{T}(\theta-\theta_0) + \text{small}. \end{aligned} \tag{183}$$

If the minimum of $Q_T(\theta|Y)$ is obtained in the interior, then

$$\sqrt{T}(\hat{\theta}_{md}-\theta_0) = \left[-\frac{1}{T}\nabla_\theta^2 Q_T(\theta_0|Y)\right]^{-1}\sqrt{T}\nabla_\theta Q_T(\theta_0|Y) + \text{small}. \tag{184}$$

Using (180), the "score" process can be expressed as

$$\sqrt{T}\nabla_\theta Q_T(\theta_0|Y) = (\nabla_\theta G_T(\theta_0|Y))W_T\sqrt{T}G_T(\theta_0|Y) \tag{185}$$

and its distribution depends on the distribution of

$$\begin{aligned} \sqrt{T}G_T(\theta_0|Y) &= \sqrt{T}(\hat{m}_T(Y) - \mathbb{E}[\hat{m}_T(Y)|\theta_0, M_1]) \\ &\quad + \sqrt{T}\left(\hat{\mathbb{E}}[\hat{m}_T(Y)|\theta_0, M_1] - \mathbb{E}[\hat{m}_T(Y)|\theta_0, M_1]\right) \\ &= I + II, \end{aligned} \tag{186}$$

say. Term I captures the variability of the deviations of the sample moment $\hat{m}_T(Y)$ from its expected value $\mathbb{E}[\hat{m}_T(Y)|\theta_0, M_1]$ and term II captures the error due to approximating $\mathbb{E}[\hat{m}_T(Y)|\theta_0, M_1]$ by $\hat{\mathbb{E}}[\hat{m}_T(Y)|\theta_0, M_1]$. Under suitable regularity conditions

$$\sqrt{T}G_T(\theta_0|Y) \Longrightarrow N(0, \boldsymbol{\Omega}), \tag{187}$$

and

$$\sqrt{T}(\hat{\theta}_{md}-\theta_0) \Longrightarrow N\left(0, (DWD')^{-1}DW\boldsymbol{\Omega}WD'(DWD')^{-1}\right), \tag{188}$$

where W is the limit of the sequence of weight matrices W_T and the matrix D is defined as the probability limit of $\nabla_\theta G_T(\theta_0|Y)$. To construct tests and confidence sets based on the limit distribution, the matrices D and $\boldsymbol{\Omega}$ have to be replaced by consistent estimates. We will discuss the structure of $\boldsymbol{\Omega}$ in more detail below.

If the number of moment conditions exceeds the number of parameters, then the model specification can be tested based on the overidentifying moment conditions. If $W_T = [\hat{\boldsymbol{\Omega}}_T]^{-1}$, where $\hat{\boldsymbol{\Omega}}_T$ is a consistent estimator of $\boldsymbol{\Omega}$, then

$$TQ_T(\hat{\theta}_{md}|Y) \Longrightarrow \chi^2_{df}, \tag{189}$$

where the degrees of freedom df equal the number of overidentifying moment conditions. The sample objective function can also be used to construct hypothesis tests for θ. Suppose that the null hypothesis is $\theta = \theta_0$. A quasi-likelihood ratio test is based on $T(Q_T(\theta_0|Y) - Q_T(\hat{\theta}_{md}|Y)$; a quasi-Lagrange-multiplier test is based on a properly standardized quadratic form of $\sqrt{T}\nabla_\theta Q_T(\theta_0|Y)$; and a Wald test is based on a properly standardized quadratic form of $\sqrt{T}(\hat{\theta}_{md}-\theta_0)$. Any of these test statistics can be inverted to construct a confidence set. Moreover, if the parameters suffer from identification problems, then the approach of Andrews and Mikusheva (2015) can be used to conduct identification-robust inference based on the quasi-Lagrange-multiplier test.

11.2.2 Approximating Model-Implied Moments

In many instances the model-implied moments $\mathbb{E}[m_T(Y)|\theta,M_1]$ are approximated by an estimate $\hat{\mathbb{E}}[m_T(Y)|\theta,M_1]$. This approximation affects the distribution of $\hat{\theta}_{md}$ through term *II* in (186). Consider the earlier example in (178) and (179) in which $\hat{m}_T(Y)$ corresponds to the OLS estimates of a VAR(1). Because the OLS estimator has a bias that vanishes at rate $1/T$, we can deduce that term *II* converges to zero and does not affect the asymptotic covariance matrix $\mathbf{\Omega}$.

The more interesting case is the one in which $\hat{\mathbb{E}}[m_T(Y)|\theta,M_1]$ is based on the simulation of the DSGE model. The asymptotic theory for simulation-based extremum estimators has been developed in Pakes and Pollard (1989). Lee and Ingram (1991) and Smith Jr. (1993) are the first papers that use simulated method of moments to estimate DSGE models. For concreteness, suppose that $m_T(Y)$ corresponds to the first-order (uncentered) sample autocovariances. We previously showed that, provided the y_t's are stationary, $\mathbb{E}[m_T(Y)|\theta,M_1]$ is given by the DSGE model population autocovariance matrix $\mathbb{E}[y_2y_1'|\theta,M_1]$, which can be approximated by simulating a sample of length λT of artificial observations Y^* from the DSGE model M_1 conditional on θ. Based on these simulated observations one can compute the sample autocovariances $\hat{m}_{\lambda T}(Y^*(\theta,M_1))$. In this case term *II* is given by

$$II=\frac{1}{\sqrt{\lambda}}\sqrt{\lambda T}\left(\frac{1}{\lambda T}\sum_{t=1}^{\lambda T}y_t^*y_{t-1}^*-\mathbb{E}[y_2y_1'|\theta_0,M_1]\right) \tag{190}$$

and satisfies a CLT. Because the simulated data are independent of the actual data, terms *I* and *II* in (186) are independent and we can write

$$\mathbf{\Omega}=\mathbb{V}_\infty[I]+\mathbb{V}_\infty[II], \tag{191}$$

where

$$\mathbb{V}_\infty[II]=\frac{1}{\lambda}\left(\lim_{T\to\infty}T\mathbb{V}[\hat{m}_T(Y^*(\theta_0,M_1))]\right) \tag{192}$$

and can be derived from the DSGE model. The larger λ, the more accurate the simulation approximation and the contribution of $\mathbb{V}_\infty[II]$ to the overall covariance matrix $\mathbf{\Omega}$.

We generated the simulation approximation by simulating one long sample of observations from the DSGE model. Alternatively, we could have simulated λ samples Y^i, $i=1,\lambda$ of size T. It turns out that for the approximation, say, of $\mathbb{E}[y_2y_1'|\theta,M_1]$, it does not matter because $\hat{m}_T(Y^*(\theta,M_1))$ is an unbiased estimator of $\mathbb{E}[y_2y_1'|\theta,M_1]$. However, if

$\hat{m}_T(Y)$ is defined as the OLS estimator of a VAR(1), then the small-sample bias of the OLS estimator generates an $O(T^{-1})$ wedge between

$$\left(\sum_{t=1}^{\lambda T} y_{t-1}^{*} y_{t-1}^{*\prime}\right)^{-1} \sum_{t=1}^{\lambda T} y_{t-1}^{*} y_{t-1}^{*\prime} \quad \text{and} \quad \mathbb{E}\left[\left(\sum_{t=1}^{T} y_{t-1} y_{t-1}^{\prime}\right)^{-1} \sum_{t=1}^{T} y_{t-1} y_{t-1}^{\prime} \middle| \theta, M_1\right].$$

For large values of λ, this wedge can be reduced by using

$$\hat{E}[m_T(Y)|\theta, M_1] = \frac{1}{\lambda} \sum_{i=1}^{\lambda} \left(\sum_{t=1}^{T} y_{t-1}^{i} y_{t-1}^{i\prime}\right)^{-1} \sum_{t=1}^{T} y_{t-1}^{i} y_{t-1}^{i\prime}$$

instead. Averaging OLS estimators from model-generated data reproduces the $O(T^{-1})$ bias of the OLS estimator captured by $\mathbb{E}[\hat{m}_T(Y)|\theta_0, M_1]$ and can lead to a final sample bias reduction in term II, which improves the small sample performance of $\hat{\theta}_{md}$.[at]

When implementing the simulation approximation of the moments, it is important to fix the random seed when generating the sample Y^* such that for each parameter value of θ the same sequence of random variables is used in computing $Y^*(\theta, M_1)$. This ensures that the sample objective function $Q_T(\theta|Y)$ remains sufficiently smooth with respect to θ to render the second-order approximation of the objective function valid.

11.2.3 Misspecification

Under the assumption that the DSGE model is correctly specified, the MD estimator has a well-defined almost-sure limit θ_0 and the asymptotic variance $\mathbb{V}_\infty[I]$ of term I in (186) is given by the model-implied variance

$$\mathbb{V}_\infty[I] = \left(\lim_{T\to\infty} T\mathbb{V}[\hat{m}_T(Y^*(\theta_0, M_1))]\right), \tag{193}$$

which up to the factor of $1/\lambda$ is identical to the contribution $\mathbb{V}_\infty[II]$ of the simulation approximation of the moments to the overall asymptotic variance $\boldsymbol{\Omega}$; see (192). Under the assumption of correct specification, it is optimal to choose the weight matrix W based on the accuracy with which the elements of the moment vector $\hat{m}_T(Y)$ measure the population analog $\mathbb{E}[\hat{m}_T(Y)|\theta_0, M_1]$. If the number of moment conditions exceeds the number of parameters, it is optimal (in the sense of minimizing the sampling variance of $\hat{\theta}_{md}$) to place more weight on matching moments that are accurately measured in the data, by setting $W = \boldsymbol{\Omega}^{-1}$. In finite sample, one can construct W_T from a consistent estimator of $\boldsymbol{\Omega}^{-1}$.

[at] See Gourieroux et al. (2010) for a formal analysis in the context of a dynamic panel data model.

If the DSGE model is regarded as misspecified, then the sampling distribution of the MD estimator has to be derived under the distribution of a reference model $p(Y|M_0)$. In this case we can define

$$\theta_0(Q) = \lim_{T\to\infty} \operatorname{argmin}_\theta \|\mathbb{E}[\hat{m}_T(Y)|M_0] - \mathbb{E}[\hat{m}|\theta, M_1]\|_W \tag{194}$$

and, under suitable regularity, the estimator $\hat{\theta}_{md}$ will converge to the pseudo-optimal value θ_0. Note that θ_0 is a function of the moments $\hat{m}_T(Y)$ that are being matched and the weight matrix W (indicated by the Q argument). Both $\hat{m}$ and W are chosen by the researcher based on the particular application. The vector $\hat{m}$ should correspond to a set of moments that are deemed to be informative about the desired parameterization of the DSGE model and reflect the ultimate purpose of the estimated DSGE model. The weight matrix W should reflect beliefs about the informativeness of certain sample moments with respect to the desired parameterization of the DSGE model.

To provide an example, consider the case of a DSGE model with stochastic singularity that attributes all business cycle fluctuations to technology shocks. To the extent that the observed data are not consistent with this singularity, the model is misspecified. A moment-based estimation of the model will ultimately lead to inflated estimates of the standard deviation of the technology shock innovation, because this shock alone has to generate the observed variability in, say, output growth, the labor share, and other variables. The extent to which the estimated shock variance is upwardly biased depends on exactly which moments the estimator is trying to match. If one of the priorities of the estimation exercise is to match the unconditional variance of output growth, then the weight matrix W should assign a large weight to this moment, even if it is imprecisely measured by its sample analog in the data.

The asymptotic variance $\mathbb{V}_\infty[I]$ of term I in (186) is now determined by the variance of the sample moments implied by the reference model M_0:

$$\mathbb{V}_\infty[I] = \left(\lim_{T\to\infty} T\mathbb{V}[\hat{m}_T(Y)|M_0]\right). \tag{195}$$

Suppose that $\hat{m}_T(Y) = \frac{1}{T}\sum_{t=1}^T y_t y_{t-1}'$, which under suitable regularity conditions converges to the population autocovariance matrix $\mathbb{E}[y_1 y_0'|M_0]$ under the reference model M_0. If the reference model is a linear process, then the asymptotic theory developed in Phillips and Solo (1992) can be used to determine the limit covariance matrix $\mathbb{V}_\infty[I]$. An estimate of $\mathbb{V}_\infty[I]$ can be obtained with a heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimator that accounts for the serial correlation in the matrix-valued sequence $\{y_t y_{t-1}'\}_{t=1}^T$. An extension of indirect inference in which $\hat{m}_T(Y)$ comprises estimates of an approximating model to the case of misspecified DSGE models is provided in Dridi et al. (2007).

11.2.4 Illustration

Detailed studies of the small-sample properties of MD estimators for DSGE models can be found in Ruge-Murcia (2007) and Ruge-Murcia (2012). To illustrate the behavior of the MD estimator we repeatedly generate data from the stylized DSGE model, treating the values listed in Table 5 as "true" parameters. We fix all parameters except for the Calvo parameter ζ_p at their "true" values and use two versions of the MD procedure to estimate ζ_p. The vector of moment conditions $\hat{m}_T(Y)$ is defined as follows. Let $y_t = [\log(X_t/X_{t-1}), \pi_t]'$ and consider a VAR(2) in output growth and inflation:

$$y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \Phi_0 + u_t. \tag{196}$$

Let $\hat{m}_T(Y) = \hat{\Phi}$ be the OLS estimate of $[\Phi_1, \Phi_2, \Phi_0]'$.

The results in the left panel of Fig. 28 are obtained by a simulation approximation of the model-implied expected value of $\hat{m}_T(Y)$. We simulate $N = 100$ trajectories of length $T + T_0$ and discarding the first T_0 observations. Let $Y_{1:T}^{(i)}(\theta)$ be the i-th simulated trajectory and define

$$\mathbb{E}[\hat{m}_T(Y)|\theta, M_1] \approx \frac{1}{N}\sum_{i=1}^{N} \hat{m}_T(Y^{(i)}(\theta)), \tag{197}$$

which can be used to evaluate the objective function (181). For the illustration we use the optimal weight matrix $W_T = \hat{\Sigma}^{-1} \otimes X'X$, where X is the matrix of regressors for the VAR(2) and $\hat{\Sigma}$ an estimate of the covariance matrix of the VAR innovations. Because we are estimating a single parameter, we compute the estimator $\hat{\theta}_{md}$ by grid search. It

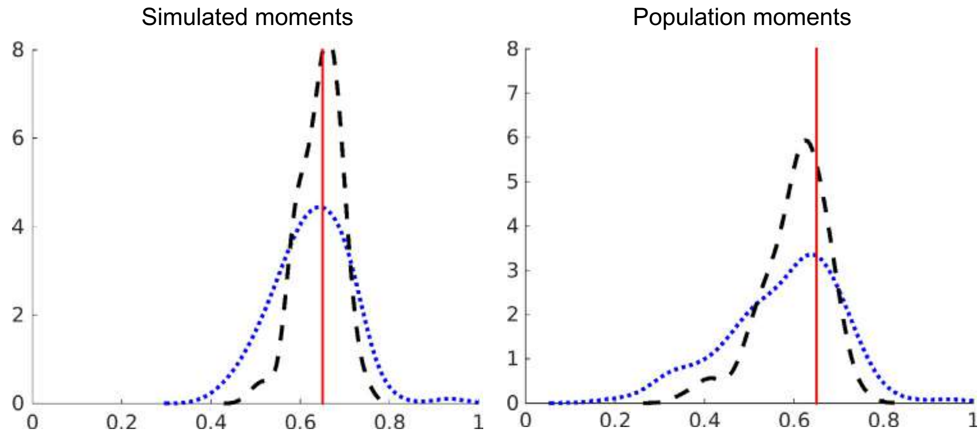


Fig. 28 Sampling distribution of $\hat{\zeta}_{p,md}$. *Notes*: We simulate samples of size $T = 80$ (*dotted*) and $T = 200$ (*dashed*) and compute two versions of an MD estimator for the Calvo parameter ζ_p. All other parameters are fixed at their "true" value. The plots depict densities of the sampling distribution of $\hat{\zeta}_{p,md}$. The vertical line indicates the "true" value of ζ_p.

is important to use the same sequence of random numbers for each value of $\theta \in \mathcal{T}$ to compute the simulation approximation $\mathbb{E}[\hat{m}_T(Y)|\theta, M_1]$. The results in the right panel of Fig. 28 are based on the VAR(2) approximation of the DSGE model based on a population regression. Let $x_t' = [y_{t-1}', y_{t-2}', 1]$ and let

$$\mathbb{E}[\hat{m}_T(Y)|\theta, M_1] \approx \left(\mathbb{E}[x_t x_t'|\theta, M_1]\right)^{-1} \mathbb{E}[x_t y_t'|\theta, M_1]. \tag{198}$$

Fig. 28 depicts density estimates of the sampling distribution of $\hat{\zeta}_{p,md}$. The vertical line indicates the "true" parameter value of ζ_p. As the sample size increases from $T = 80$ to $T = 200$, the sampling distribution concentrates around the "true" value and starts to look more like a normal distribution, as the asymptotic theory presented in this section suggests. The distribution of the estimator based on the simulated objective function is more symmetric around the "true" value and also less variable. However, even based on a sample size of 200 observations, there is considerable uncertainty about the Calvo parameter and hence the slope of the New Keynesian Phillips curve. A comparison with Fig. 27 indicates that the MD estimator considered in this illustration is less efficient than the ML estimator.

11.2.5 Laplace Type Estimators

In DSGE model applications the estimation objective function $Q_T(\theta|Y)$ is often difficult to optimize. Chernozhukov and Hong (2003) proposed computing a mean of a quasi-posterior density instead of computing an extremum estimator. The resulting estimator is called a Laplace-type (LT) estimator and defined as follows (provided the integral in the denominator is well defined):

$$\hat{\theta}_{LT} = \frac{\exp\left\{-\frac{1}{2} Q_T(\theta|Y)\right\}}{\int \exp\left\{-\frac{1}{2} Q_T(\theta|Y)\right\} d\theta}. \tag{199}$$

This estimator can be evaluated using the Metropolis–Hastings algorithm discussed in Section 12.2 or the sequential Monte Carlo algorithm presented in Section 12.3 below. The posterior computations may be more accurate than the computation of an extremum. Moreover, suppose that the objective function is multimodal. In repeated sampling, the extremum of the objective function may shift from one mode to the other, making the estimator appear to be unstable. On the other hand, owing to the averaging, the LT estimator may be more stable. Chernozhukov and Hong (2003) establish the consistency and asymptotic normality of LT estimators, which is not surprising because the sample objective function concentrates around its extremum as $T \longrightarrow \infty$ and the discrepancy between the extremum and the quasi-posterior mean vanishes. DSGE model applications of LT estimators are provided in Kormilitsina and Nekipelov (2012, 2016).

LT estimators can be constructed not only from MD estimators but also from IRF matching estimators and GMM estimators discussed below.

11.3 Impulse Response Function Matching

As discussed previously, sometimes DSGE models are misspecified because researchers have deliberately omitted structural shocks that contribute to business cycle fluctuations. An example of such a model is the one developed by Christiano et al. (2005). The authors focus their analysis on the propagation of a single shock, namely, a monetary policy shock. If it is clear that if the DSGE model does not contain enough structural shocks to explain the variability in the observed data, then it is sensible to try to purge the effects of the unspecified shocks from the data, before matching the DSGE model to the observations. This can be done by "filtering" the data through the lens of a VAR that identifies the impulse responses to those shocks that are included in the DSGE model. The model parameters can then be estimated by minimizing the discrepancy between model-implied and empirical impulse response functions. A mismatch between the two sets of impulse responses provides valuable information about the misspecification of the propagation mechanism and can be used to develop better-fitting DSGE models. Influential papers that estimate DSGE models by matching impulse response functions include Rotemberg and Woodford (1997), Christiano et al. (2005), and Altig et al. (2011). The casual description suggests that impulse response function matching estimators are a special case of the previously discussed MD estimators (the DSGE model M_1 is misspecified and a structural VAR serves as reference model M_0 under which the sampling distribution of the estimator is derived). Unfortunately, several complications arise, which we will discuss in the remainder of this section. Throughout, we assume that the DSGE model has been linearized. An extension to the case of nonlinear DSGE models is discussed in Ruge-Murcia (2014).

11.3.1 Invertibility and Finite-Order VAR Approximations

The empirical impulse responses are based on a finite-order VAR, such as the one in (113). However, even linearized DSGE models typically cannot be written as a finite-order VAR. Instead, they take the form of a state-space model, which typically has a VARMA representation. In general we can distinguish the following three cases: (i) the solution of the DSGE model can be expressed as a VAR(p). For the stylized DSGE model, this is the case if y_t is composed of four observables: output growth, the labor share, inflation, and interest rates. (ii) The moving average polynomial of the VARMA representation of the DSGE model is invertible. In this case the DSGE model can be expressed as an infinite-order VAR driven by the structural shock innovations ϵ_t. (iii) The moving average polynomial of the VARMA representation of the DSGE model is not invertible. In this case the innovation of the VAR(∞) approximation do not correspond to the structural innovations ϵ_t. Only in case (i) can one expect a direct match between the empirical IRFs and the DSGE model IRFs. Cases (ii) and (iii) complicate econometric inference. The extent to which impulse-response-function-based estimation and model evaluation may be misleading has been fiercely debated in Christiano et al. (2007) and Chari et al. (2008).

Fernández-Villaverde et al. (2007) provide formal criteria to determine whether a DSGE model falls under case (i), (ii), or (iii). Rather than presenting a general analysis of this problem, we focus on a simple example. Consider the following two MA processes that represent the DSGE models in this example:

$$\begin{array}{lcl} M_1 & : & y_t = \epsilon_t + \theta\epsilon_{t-1} = (1+\theta L)\epsilon_t \\ M_2 & : & y_t = \theta\epsilon_t + \epsilon_{t-1} = (\theta + L)\epsilon_t, \end{array} \tag{200}$$

where $0 < \theta < 1$, L denotes the lag operator, and $\epsilon_t \sim iidN(0,1)$. Models M_1 and M_2 are observationally equivalent, because they are associated with the same autocovariance sequence. The root of the MA polynomial of model M_1 is outside of the unit circle, which implies that the MA polynomial is invertible and one can express y_t as an AR(∞) process:

$$AR(\infty) \text{ for } M_1: \; y_t = -\sum_{j=1}^{\infty}(-\theta)^j y_{t-j} + \epsilon_t. \tag{201}$$

It is straightforward to verify that the AR(∞) approximation reproduces the impulse response function of M_1:

$$\frac{\partial y_t}{\partial \epsilon_t} = 1, \quad \frac{\partial y_{t+1}}{\partial \epsilon_t} = \theta, \quad \frac{\partial y_{t+h}}{\partial \epsilon_t} = 0 \text{ for } h > 1.$$

Thus, the estimation of an autoregressive model with many lags can reproduce the monotone impulse response function of model M_1.

The root of the MA polynomial of M_2 lies inside the unit circle. While M_2 could also be expressed as an AR(∞), it would be a representation in terms of a serially uncorrelated one-step-ahead forecast error u_t that is a function of the infinite history of the ϵ_t's: $u_t = (1+\theta L)^{-1}(\theta + L)$. As a consequence, the AR(∞) is unable to reproduce the hump-shaped IRF of model M_2. More generally, if the DSGE model is associated with a noninvertible moving average polynomial, its impulse responses cannot be approximated by a VAR(∞) and a direct comparison of VAR and DSGE IRFs may be misleading.

11.3.2 Practical Considerations

The objective function for the IRF matching estimator takes the same form as the criterion function of the method of moments estimator in (180) and (181), where $\hat{m}_T(Y)$ is the VAR IRF. For $\hat{\mathbb{E}}[\hat{m}_T(Y)|\theta, M_1]$ researchers typically just use the DSGE model impulse response, say, $IRF(\cdot|\theta,M_1)$. In view of the problems caused by noninvertible moving-average polynomials and finite-order VAR approximations of infinite-order VAR representations, a more prudent approach would be to replace

$IRF(\cdot|\theta,M_1)$ by average impulse response functions that are obtained by repeatedly simulating data from the DSGE model (given θ) and estimating a structural VAR, as in the indirect inference approach described in Section 11.2. Such a modification would address the concerns about IRF matching estimators raised by Chari et al. (2008).

The sampling distribution of the IRF matching estimator depends on the sampling distribution of the empirical VAR impulse responses $\hat{m}_T(Y)$ under the VAR M_0. An approximation of the distribution of $\hat{m}_T(Y)$ could be obtained by first-order asymptotics and the delta method as in Lütkepohl (1990) and Mittnik and Zadrozny (1993) for stationary VARs; or as in Phillips (1998), Rossi and Pesavento (2006), and Pesavento and Rossi (2007) for VARs with persistent components. Alternatively, one could use the bootstrap approximation proposed by Kilian (1998, 1999). If the number of impulse responses stacked in the vector $\hat{m}_T(Y)$ exceeds the number of reduced-form VAR coefficient estimates, then the sampling distribution of the IRFs becomes asymptotically singular. Guerrón-Quintana et al. (2014) use nonstandard asymptotics to derive the distribution of IRFs for the case in which there are more responses than reduced-form parameters.

Because for high-dimensional vectors $\hat{m}_T(Y)$ the joint covariance matrix may be close to singular, researchers typically choose a diagonal weight matrix W_T, where the diagonal elements correspond to the inverse of the sampling variance for the estimated response of variable i to shock j at horizon h. As discussed in Section 11.2, to the extent that the DSGE model is misspecified, the choice of weight matrix affects the probability limit of the IRF matching estimator and should reflect the researcher's loss function.

In fact, impulse response function matching is appealing only if the researcher is concerned about model misspecification. This misspecification might take two forms: First, the propagation mechanism of the DSGE model is potentially misspecified and the goal is to find pseudo-optimal parameter values that minimize the discrepancy between empirical and model-implied impulse responses. Second, the propagation mechanisms for the shocks of interest are believed to be correctly specified, but the model lacks sufficiently many stochastic shocks to capture the observed variation in the data. In the second case, it is in principle possible to recover the subset of "true" DSGE model parameters θ_0 that affect the propagation of the structural shock for which the IRF is computed. The consistent estimation would require that the DSGE model allow for a VAR(∞) representation in terms of the structural shock innovations ϵ_t; that the number of lags included in the empirical VAR increase with sample size T; and that the VAR identification scheme correctly identify the shock of interest if the data are generated from a version of the DSGE model that is augmented by additional structural shocks.

11.3.3 Illustration

To illustrate the properties of the IRF matching estimator, we simulate data from the stylized DSGE model using the parameter values given in Table 5. We assume that the econometrician considers an incomplete version of the DSGE model that only includes the monetary policy shock and omits the remaining shocks. Moreover, we assume that the econometrician only has to estimate the degree of price stickiness captured by the Calvo parameter ζ_p. All other parameters are fixed at their "true" values during the estimation.

The empirical impulse response functions stacked in the vector $\hat{m}_T(Y)$ are obtained by estimating a VAR(p) for interest rates, output growth, and inflation:

$$y_t = [R_t - \pi_t/\beta, \log(X_t/X_{t-1}), \pi_t]'. \tag{202}$$

The first equation of this VAR represents the monetary policy rule of the DSGE model. The interest rate is expressed in deviations from the central bank's systematic reaction to inflation. Thus, conditional on β, the monetary policy shock is identified as the orthogonalized one-step-ahead forecast error in the first equation of the VAR. Upon impact, the response of y_t to the monetary policy shock is given by the first column of the lower-triangular Cholesky factor of the covariance matrix Σ of the reduced-form innovations u_t.

Because y_t excludes the labor share, the state-space representation of the DSGE model cannot be expressed as a finite-order VAR. However, we can construct a VAR approximation of the DSGE model as follows. Let $x_t = [y'_{t-1},\ldots,y'_{t-p},1']'$ and define the functions[au]

$$\begin{aligned} \Phi*(\theta) &= \left(\mathbb{E}[x_t x'_t|\theta, M_1]\right)^{-1}\left(\mathbb{E}[x_t y'_t|\theta, M_1]\right), \\ \Sigma^*(\theta) &= \mathbb{E}[y_t y'_t|\theta, M_1] - \mathbb{E}[y_t x'_t|\theta, M_1]\left(\mathbb{E}[x_t x'_t|\theta, M_1]\right)^{-1}\mathbb{E}[x_t y'_t|\theta, M_1]. \end{aligned} \tag{203}$$

Note that $\Phi*(\theta)$ and $\Sigma*(\theta)$ are functions of the population autocovariances of the DSGE model. For a linearized DSGE model, these autocovariances can be expressed analytically as a function of the coefficient matrices of the model's state-space representation.

The above definition of $\Phi*(\theta)$ and $\Sigma^*(\theta)$ requires that $\mathbb{E}[x_t x'_t|\theta, M_1]$ is nonsingular. This condition is satisfied as long as $n_y \leq n_\epsilon$. However, the appeal of IRF matching estimators is that they can be used in settings in which only a few important shocks are incorporated into the model and $n_y > n_\epsilon$. In this case, $\Phi*(\theta)$ and $\Sigma^*(\theta)$ have to be modified, for instance, by computing the moment matrices based on $\tilde{y}_t = y_t + u_t$, where u_t is a "measurement error," or by replacing $\left(\mathbb{E}[x_t x'_t|\theta, M_1]\right)^{-1}$ with $\left(\mathbb{E}[x_t x'_t|\theta, M_1] + \lambda I\right)^{-1}$, where λ is a scalar and I is the identity matrix. In the subsequent illustration, we keep all the structural shocks in the DSGE model active, ie, $n_y \leq n_\epsilon$, such that the restriction functions can indeed be computed based on (203).

[au] For the evaluation of the moment matrices $\mathbb{E}[\cdot|\theta, M_1]$ see Section 8.2.1.

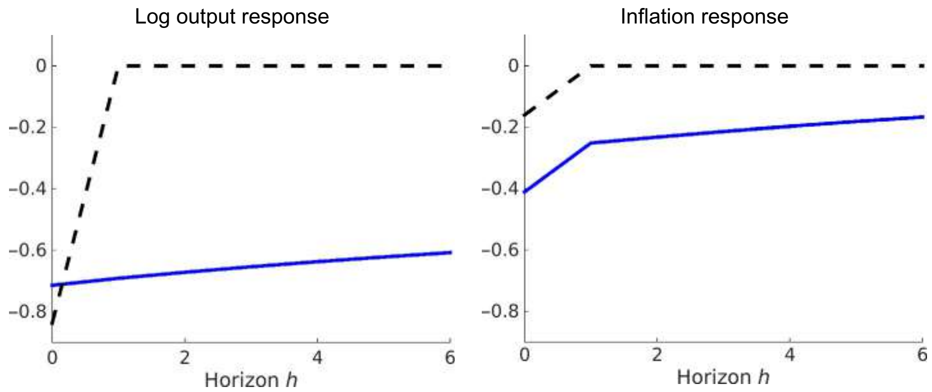


Fig. 29 DSGE model and VAR impulse responses to a monetary policy shock. *Notes*: The figure depicts impulse responses to a monetary policy shock computed from the state-space representation of the DSGE model (*dashed*) and the VAR(1) approximation of the DSGE model (*solid*).

Fig. 29 compares the impulse responses from the state-space representation and the VAR approximation of the DSGE model. It turns out that there is a substantial discrepancy. Because the monetary policy shock is *iid* and the stylized DSGE model does not have an endogenous propagation mechanism, both output and inflation revert back to the steady state after one period. The VAR response, on the other hand, is more persistent and the relative movement of output and inflation is distorted. Augmenting a VAR(1) with additional lags has no noticeable effect on the impulse response.

The IRF matching estimator minimizes the discrepancy between the empirical and the DSGE model-implied impulse responses by varying ζ_p. Fig. 30 illustrates the effect of ζ_p on the response of output and inflation. The larger ζ_p, the stronger the nominal rigidity, and the larger the effect of a monetary policy shock on output. Fig. 31 shows the sampling distribution of the IRF matching estimator for the sample sizes $T = 80$ and $T = 200$. We match IRFs over 10 horizons and use an identity weight matrix. If $\hat{\mathbb{E}}[\hat{m}_T(Y)|\theta, M_1]$ is defined as the IRF implied by the state-space representation, then the resulting estimator of ζ_p has a fairly strong downward bias. This is not surprising in view of the mismatch depicted in Figs. 29 and 30. If the state-space IRF is replaced by the IRF obtained from the VAR approximation of the DSGE model, then the sampling distribution is roughly centered at the "true" parameter value, though it is considerably more dispersed, also compared to the MD estimator in Fig. 28. This is consistent with the fact that the IRF matching estimator does not utilize variation in output and inflation generated by the other shocks.

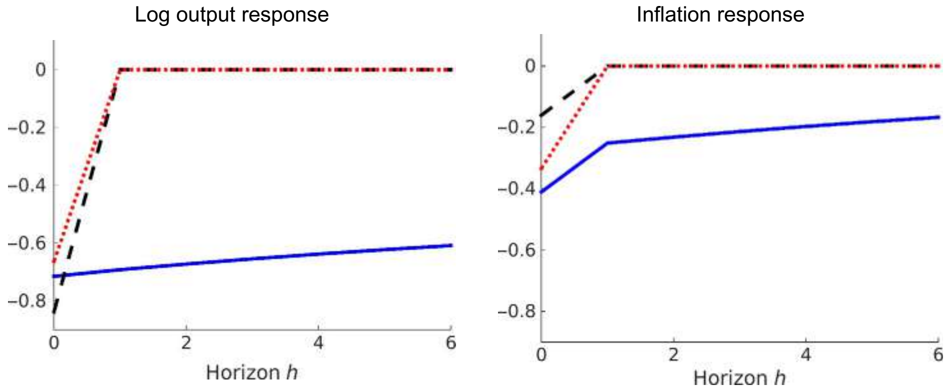


Fig. 30 Sensitivity of IRF to ζ_p. *Notes*: The *solid lines* indicate IRFs computed from the VAR approximation of the DSGE model. The other two lines depict DSGE model-implied IRFs based on $\zeta_p = 0.65$ (*dashed*) and $\zeta_p = 0.5$ (*dotted*).

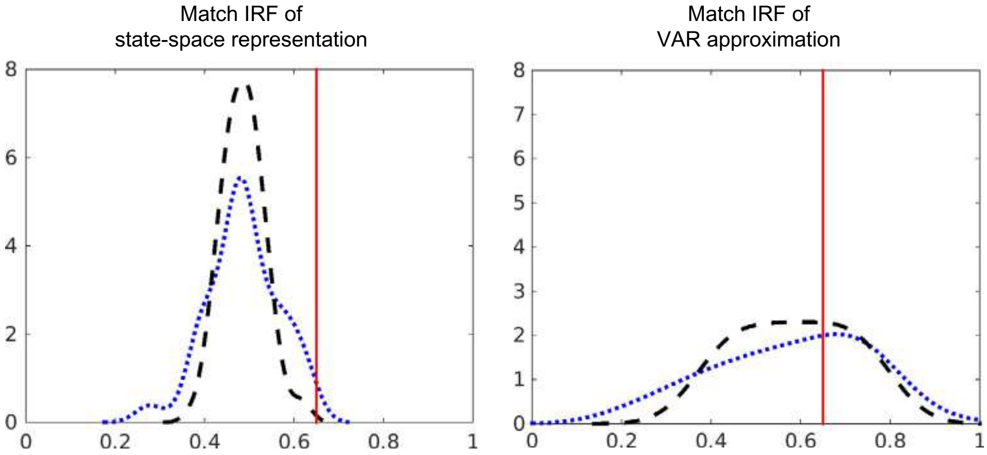


Fig. 31 Sampling distribution of $\hat{\zeta}_{p,irf}$. *Notes*: We simulate samples of size $T = 80$ and $T = 200$ and compute IRF matching estimators for the Calvo parameter ζ_p based on two choices of $\hat{\mathbb{E}}[\hat{m}_T(Y)|\theta,M_1]$. For the *left panel* we use the IRFs from the state-space representation of the DSGE model; for the *right panel* we use the IRF from the VAR approximation of the DSGE model. All other parameters are fixed at their "true" value. The plot depicts densities of the sampling distribution of $\hat{\zeta}_p$ for $T = 80$ (*dotted*) and $T = 200$ (*dashed*). The vertical line indicates the "true" value of ζ_p.

11.4 GMM Estimation

We showed in Section 8.2.4 that one can derive moment conditions of the form

$$\mathbb{E}[g(y_{t-p:t}|\theta, M_1)] = 0 \tag{204}$$

for $\theta = \theta_0$ from the DSGE model equilibrium. For instance, based on (106) and (107) we could define

$$g(y_{t-p:t}|\theta, M_1) = \begin{bmatrix} (-\log(X_t/X_{t-1}) + \log R_{t-1} - \log \pi_t - \log(1/\beta))Z_{t-1} \\ (\log R_t - \log(\gamma/\beta) - \psi \log \pi_t - (1-\psi)\log \pi^*)Z_{t-1} \end{bmatrix}. \tag{205}$$

The identifiability of θ requires that the moments be different from zero whenever $\theta \neq \theta_0$. A GMM estimator is obtained by replacing population expectations by sample averages. Let

$$G_T(\theta|Y) = \frac{1}{T}\sum_{t=1}^{T} g(y_{t-p:t}|\theta, M_1). \tag{206}$$

The GMM objective function is given by

$$Q_T(\theta|Y) = G_T(\theta|Y)' W_T G_T(\theta|Y) \tag{207}$$

and looks identical to the objective function studied in Section 11.2. In turn, the analysis of the sampling distribution of $\hat{\theta}_{md}$ carries over to the GMM estimator.

The theoretical foundations of GMM estimation were developed by Hansen (1982), who derived the first-order asymptotics for the estimator assuming that the data are stationary and ergodic. Christiano and Eichenbaum (1992) and Burnside et al. (1993) use GMM to estimate the parameters of real business cycle DSGE models. These papers use sufficiently many moment conditions to be able to estimate all the parameters of their respective DSGE models. GMM estimation can also be applied to a subset of the equilibrium conditions, eg, the consumption Euler equation or the New Keynesian Phillips curve to estimate the parameters related to these equilibrium conditions.

Unlike all the other estimators considered in this paper, the GMM estimators do not require the researchers to solve the DSGE model. To the extent that solving the model is computationally costly, this can considerably speed up the estimation process. Moreover, one can select moment conditions that do not require assumptions about the law of motion of exogenous driving processes, which robustifies the GMM estimator against misspecification of the exogenous propagation mechanism. However, it is difficult to exploit moment conditions in which some of the latent variables appear explicitly. For instance, consider the Phillips curve relationship of the stylized DSGE model, which suggests setting

$$g(y_{t-p:t}|\theta, M_1) = \left(\hat{\pi}_{t-1} - \beta\hat{\pi}_t - \kappa_p(\widehat{lsh}_{t-1})\right)Z_{t-1}. \tag{208}$$

Note that λ_{t-1} is omitted from the definition of $g(y_{t-p:t}|\theta,M_1)$ because it is unobserved. However, as soon as Z_t is correlated with the latent variable λ_t the expected value of $g(y_{t-p:t}|\theta,M_1)$ is nonzero even for $\theta = \theta_0$:

$$\mathbb{E}[g(y_{t-p:t}|\theta_0, M_1)] = -\kappa_0\mathbb{E}[\lambda_{t-1}Z_{t-1}] \neq 0. \tag{209}$$

To the extent that λ_t is serially correlated, using higher-order lags of y_t as instruments does not solve the problem.[av] Recent work by Gallant et al. (2013) and Shin (2014) considers extensions of GMM estimation to moment conditions with latent variables.

The recent literature on GMM estimation of DSGE models has focused on identification-robust inference in view of the weak identification of Phillips curve and monetary policy rule parameters. Generic identification problems in the context of monetary policy rule estimation are highlighted in Cochrane (2011) and methods to conduct identification-robust inference are developed in Mavroeidis (2010). Identification-robust inference for Phillips curve parameters is discussed in Mavroeidis (2005), Kleibergen and Mavroeidis (2009), and Mavroeidis et al. (2014). Dufour et al. (2013) consider identification-robust moment-based estimation of all of the equilibrium relationships of a DSGE model.

12. BAYESIAN ESTIMATION TECHNIQUES

Bayesian inference is widely used in empirical work with DSGE models. The first papers to estimate small-scale DSGE models using Bayesian methods were DeJong et al. (2000), Schorfheide (2000), Otrok (2001), Fernández-Villaverde and Rubio-Ramírez (2004), and Rabanal and Rubio-Ramírez (2005). Subsequent papers estimated open-economy DSGE models, eg, Lubik and Schorfheide (2006), and larger DSGE models tailored to the analysis of monetary policy, eg, Smets and Wouters (2003) and Smets and Wouters (2007). Because Bayesian analysis treats shock, parameter, and model uncertainty symmetrically by specifying a joint distribution that is updated in view of the observations Y, it provides a conceptually appealing framework for decision making under uncertainty. Levin et al. (2006) consider monetary policy analysis under uncertainty based on an estimated DSGE model and the handbook chapter by Del Negro and Schorfheide (2013) focuses on forecasting with DSGE models.

Conceptually, Bayesian inference is straightforward. A prior distribution is updated in view of the sample information contained in the likelihood function. This leads to a posterior distribution that summarizes the state of knowledge about the unknown parameter vector θ. The main practical difficulty is the calculation of posterior moments and quantiles of transformations $h(\cdot)$ of the parameter vector θ. The remainder of this section is organized as follows. We provide a brief discussion of the elicitation of prior distributions in Section 12.1. Sections 12.2 and 12.3 discuss two important algorithms to generate parameter draws from posterior distributions: Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC). Bayesian model diagnostics are reviewed in

[av] Under the assumption that λ_t follows an AR(1) process, one could quasi-difference the Phillips curve, which would replace the term $\lambda_{t-1}Z_{t-1}$ with $\epsilon_{\lambda,t-1}Z_{t-1}$. If Z_{t-1} is composed of lagged observables dated $t-2$ and earlier, then the validity of the moment condition is restored.

Section 12.4. Finally, we discuss the recently emerging literature on limited-information Bayesian inference in Section 12.5. Sections 12.1 and 12.3 are based on Herbst and Schorfheide (2015), who provide a much more detailed exposition. Section 12.4 draws from Del Negro and Schorfheide (2011).

12.1 Prior Distributions

There is some disagreement in the Bayesian literature about the role of prior information in econometric inference. Some authors advocate "flat" prior distributions that do not distort the shape of the likelihood function, which raises two issues: first, most prior distributions are not invariant under parameter transformations. Suppose a scalar parameter $\theta \sim U[-M,M]$. If the model is reparameterized in terms of $1/\theta$, the implied prior is no longer flat. Second, if the prior density is taken to be constant on the real line, say, $p(\theta) = c$, then the prior is no longer proper, meaning the total prior probability mass is infinite. In turn, it is no longer guaranteed that the posterior distribution is proper.

In many applications prior distributions are used to conduct inference in situations in which the number of unknown parameters is large relative to the number of sample observations. An example is a high-dimensional VAR. If the number of variables in the VAR is n and the number of lags is p, then each equation has at least np unknown parameters. For instance, a 4-variable VAR with $p = 4$ lags has 16 parameters. If this model is estimated based on quarterly post-Great Moderation and pre-Great Recession data, the data-to-parameter ratio is approximately 6, which leads to very noisy parameter estimates. A prior distribution essentially augments the estimation sample Y by artificial observations Y^* such that the model is estimated based on the combined sample (Y,Y^*).

Prior distributions can also be used to "regularize" the likelihood function by giving the posterior density a more elliptical shape. Finally, a prior distribution can be used to add substantive information about model parameters not contained in the estimation sample θ to the inference problem. Bayesian estimation of DSGE models uses prior distributions mostly to add information contained in data sets other than Y and to smooth out the likelihood function, down-weighing regions of the parameter space in which implications of the structural model contradict nonsample information and the model becomes implausible. An example would be a DSGE model with a likelihood that has a local maximum at which the discount factor is, say, $\beta = 0.5$. Such a value of β would strongly contradict observations of real interest rates. A prior distribution that implies that real interest rates are between 0% and 10% with high probability would squash the undesirable local maximum of the likelihood function.

To the extent that the prior distribution is "informative" and affects the shape of the posterior distribution, it is important that the specification of the prior distribution be carefully documented. Del Negro and Schorfheide (2008) developed a procedure to construct prior distributions based on information contained in presamples or in time series that are not directly used for the estimation of the DSGE model. To facilitate the elicitation of a prior distribution it is useful to distinguish three groups of parameters: steady-state-related parameters, exogenous shock parameters, and endogenous propagation parameters.

In the context of the stylized DSGE model, the steady-state-related parameters are given by β (real interest rate), π^* (inflation), γ (output growth rate), and λ (labor share). A prior for these parameters could be informed by presample averages of these series. The endogenous propagation parameters are ζ_p (Calvo probability of not being able to reoptimize price) and ν (determines the labor supply elasticity). Micro-level information about the frequency of price changes and labor supply elasticities can be used to specify a prior distribution for these two parameters. Finally, the exogenous shock parameters are the autocorrelation parameters ρ and the shock standard deviations σ.

Because the exogenous shocks are latent, it is difficult to specify a prior distribution for these parameters directly. However, it is possible to map beliefs about the persistence and volatility of observables such as output growth, inflation, and interest rates into beliefs about the exogenous shock parameters. This can be done using the formal procedure described in Del Negro and Schorfheide (2008) or, informally, by generating draws of θ from the prior distribution, simulating artificial observations from the DSGE model, and computing the implied sample moments of the observables. If the prior predictive distribution of these sample moments appears implausible, say, in view of sample statistics computed from a presample of actual observations, then one can adjust the prior distribution of the exogenous shock parameters and repeat the simulation until a plausible prior is obtained. Table 7 contains an example of a prior distribution for our stylized DSGE model. The joint distribution for θ is typically generated as a product of marginal distributions for the elements (or some transformations thereof) of the vector θ.[aw] In most applications this product of marginals is truncated to ensure that the model has a unique equilibrium.

12.2 Metropolis–Hastings Algorithm

Direct sampling from the posterior distribution of θ is unfortunately not possible. One widely used algorithm to generate draws from $p(\theta|Y)$ is the Metropolis–Hastings (MH) algorithm, which belongs to the class of MCMC algorithms. MCMC algorithms produce a sequence of serially correlated parameter draws θ^i, $i = 1,\ldots,N$ with the property that the random variables θ^i converge in distribution to the target posterior distribution, which we abbreviate as

[aw] In high-dimensional parameter spaces it might be desirable to replace some of the θ elements by transformations, eg, steady states, that are more plausibly assumed to be independent. This transformation essentially generates nonzero correlations for the original DSGE model parameters. Alternatively, the method discussed in Del Negro and Schorfheide (2008) also generates correlations between parameters.

Table 7 Prior distribution

Name	**Domain**	**Prior**		
		Density	***Para (1)***	***Para (2)***
Steady-state-related parameters $\theta_{(ss)}$				
$100(1/\beta-1)$	$\mathbb{R}^+$	Gamma	0.50	0.50
$100\log\pi^*$	$\mathbb{R}^+$	Gamma	1.00	0.50
$100\log\gamma$	$\mathbb{R}$	Normal	0.75	0.50
λ	$\mathbb{R}^+$	Gamma	0.20	0.20
Endogenous propagation parameters $\theta_{(endo)}$				
ζ_p	[0,1]	Beta	0.70	0.15
$1/(1+\nu)$	$\mathbb{R}^+$	Gamma	1.50	0.75
Exogenous shock parameters $\theta_{(exo)}$				
ρ_ϕ	[0,1)	Uniform	0.00	1.00
ρ_λ	[0,1)	Uniform	0.00	1.00
ρ_z	[0,1)	Uniform	0.00	1.00
$100\sigma_\phi$	$\mathbb{R}^+$	InvGamma	2.00	4.00
$100\sigma_\lambda$	$\mathbb{R}^+$	InvGamma	0.50	4.00
$100\sigma_z$	$\mathbb{R}^+$	InvGamma	2.00	4.00
$100\sigma_r$	$\mathbb{R}^+$	InvGamma	0.50	4.00

Notes: Marginal prior distributions for each DSGE model parameter. Para (1) and Para (2) list the means and the standard deviations for Beta, Gamma, and Normal distributions; the upper and lower bound of the support for the Uniform distribution; s and ν for the Inverse Gamma distribution, where $p_{\mathcal{IG}}(\sigma|\nu,s)\propto\sigma^{-\nu-1}e^{-\nu s^2/2\sigma^2}$. The joint prior distribution of θ is truncated at the boundary of the determinacy region.

$$\pi(\theta)=p(\theta|Y)=\frac{p(Y|\theta)p(\theta)}{p(Y)}, \tag{210}$$

as $N\rightarrow\infty$. More important, under suitable regularity conditions sample averages of draws converge to posterior expectations:

$$\frac{1}{N-N_0}\sum_{i=N_0+1}^{N}h(\theta^i)\xrightarrow{a.s.}\mathbb{E}_\pi[h(\theta)]. \tag{211}$$

Underlying this convergence result is the fact that the algorithm generates a Markov transition kernel $K(\theta^i|\theta^{i-1})$, characterizing the distribution of θ^i conditional on θ^{i-1}, with the invariance property

$$\int K(\theta^i|\theta^{i-1})\pi(\theta^{i-1})d\theta^{i-1}=\pi(\theta^i). \tag{212}$$

Thus, if θ^{i-1} is a draw from the posterior distribution, then so is θ^i. Of course, this invariance property is not sufficient to guarantee the convergence of the θ^i draws. Chib and

Greenberg (1995) provide an excellent introduction to MH algorithms and detailed textbook treatments can be found, for instance, in Robert and Casella (2004) and Geweke (2005).

12.2.1 The Basic MH Algorithm

The key ingredient of the MH algorithm is a proposal distribution $q(\vartheta|\theta^{i-1})$, which potentially depends on the draw θ^{i-1} in iteration $i-1$ of the algorithm. With probability $\alpha(\vartheta|\theta^{i-1})$ the proposed draw is accepted and $\theta^i = \vartheta$. If the proposed draw is not accepted, then the chain does not move and $\theta^i = \theta^{i-1}$. The acceptance probability is chosen to ensure that the distribution of the draws converges to the target posterior distribution. The algorithm takes the following form:

Algorithm 7 (Generic MH Algorithm). For $i = 1$ to N:

1. Draw ϑ from a density $q(\vartheta|\theta^{i-1})$.

2. Set $\theta^i = \vartheta$ with probability

$$\alpha(\vartheta|\theta^{i-1}) = \min\left\{1, \frac{p(Y|\vartheta)p(\vartheta)/q(\vartheta|\theta^{i-1})}{p(Y|\theta^{i-1})p(\theta^{i-1}))/q(\theta^{i-1}|\vartheta)}\right\}$$

and $\theta^i = \theta^{i-1}$ otherwise.

Because $p(\theta|Y) \propto p(Y|\theta)p(\theta)$ we can replace the posterior densities in the calculation of the acceptance probabilities $\alpha(\vartheta|\theta^{i-1})$ with the product of the likelihood and prior, which does not require the evaluation of the marginal data density $p(Y)$.

12.2.2 Random-Walk Metropolis–Hastings Algorithm

The most widely used MH algorithm for DSGE model applications is the *random walk MH* (RWMH) algorithm. The basic version of this algorithm uses a normal distribution centered at the previous θ^i draw as the proposal density:

$$\vartheta|\theta^i \sim N\left(\theta^i, c^2\hat{\Sigma}\right). \tag{213}$$

Given the symmetric nature of the proposal distribution, the acceptance probability becomes

$$\alpha = \min\left\{\frac{p(\vartheta|Y)}{p(\theta^{i-1}|Y)}, 1\right\}.$$

A draw, ϑ, is accepted with probability one if the posterior at ϑ has a higher value than the posterior at θ^{i-1}. The probability of acceptance decreases as the posterior at the candidate value decreases relative to the current posterior.

To implement the RWMH, the user needs to specify c, and $\hat{\Sigma}$. The proposal variance controls the relative variances and correlations in the proposal distribution. The sampler can work very poorly if q is strongly at odds with the target distribution. A good choice for $\hat{\Sigma}$ seeks to incorporate information from the posterior, to potentially capture the *a posteriori* correlations among parameters. Obtaining this information can be difficult. A popular approach, used in Schorfheide (2000), is to set $\hat{\Sigma}$ to be the negative of the inverse Hessian at the mode of the log posterior, $\hat{\theta}$, obtained by running a numerical optimization routine before running the MCMC algorithm. Using this as an estimate for the covariance of the posterior is attractive, because it can be viewed as a large sample approximation to the posterior covariance matrix.

Unfortunately, in many applications, the maximization of the posterior density is tedious and the numerical approximation of the Hessian may be inaccurate. These problems may arise if the posterior distribution is very nonelliptical and possibly multimodal, or if the likelihood function is replaced by a nondifferentiable particle filter approximation. In both cases, a (partially) adaptive approach may work well: First, generate a set of posterior draws based on a reasonable initial choice for $\hat{\Sigma}$, eg, the prior covariance matrix. Second, compute the sample covariance matrix from the first sequence of posterior draws and use it as $\hat{\Sigma}$ in a second run of the RWMH algorithm. In principle, the covariance matrix $\hat{\Sigma}$ can be adjusted more than once. However, $\hat{\Sigma}$ must be fixed eventually to guarantee the convergence of the posterior simulator. Samplers that constantly (or automatically) adjust $\hat{\Sigma}$ are known as adaptive samplers and require substantially more elaborate theoretical justifications.

12.2.3 Numerical Illustration

We generate a single sample of size $T = 80$ from the stylized DSGE model using the parameterization in Table 5. The DSGE model likelihood function is combined with the prior distribution in Table 7 to form a posterior distribution. Draws from this posterior distribution are generated using the RWMH described in the previous section. The chain is initialized with a draw from the prior distribution. The covariance matrix $\hat{\Sigma}$ is based on the negative inverse Hessian at the mode. The scaling constant c is set equal to 0.075, which leads to an acceptance rate for proposed draws of 0.55.

The top panels of Fig. 32 depict the sequences of posterior draws of the Calvo parameter ζ_p^i and preference shock standard deviation σ_ϕ^i. It is apparent from the figure that the draws are serially correlated. The draws for the standard deviation are strongly contaminated by the initialization of the chain, but they eventually settle to a range of 0.8–1.1. The bottom panel depicts recursive means of the form

$$\bar{h}_{N|N_0} = \frac{1}{N - N_0} \sum_{i=N_0+1}^{N} h(\theta^i). \tag{214}$$

To remove the effect of the initialization of the Markov chain, it is common to drop the first N_0 draws from the computation of the posterior mean approximation. In the figure we set $N_0 = 7500$ and $N = 37{,}500$. Both recursive means eventually settle to a limit point.

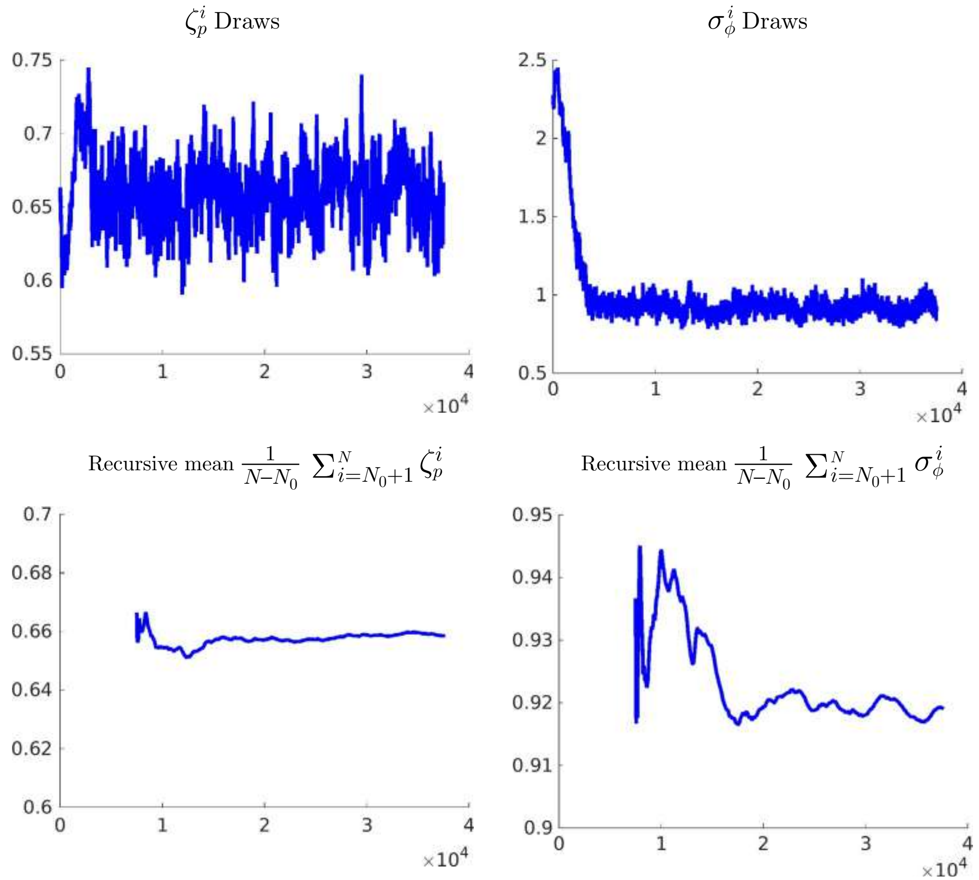


Fig. 32 Parameter draws from MH algorithm. *Notes*: The posterior is based on a simulated sample of observations of size $T = 80$. The *top panel* shows the sequence of parameter draws and the *bottom panel* shows recursive means.

The output of the algorithm is stochastic, which implies that running the algorithm repeatedly will generate different numerical results. Under suitable regularity conditions the recursive means satisfy a CLT. The easiest way to obtain a measure of numerical accuracy is to run the RWMH algorithm, say, fifty times using random starting points, and compute the sample variance of $\bar{h}_{N|N_0}$ across chains. Alternatively, one could compute a heteroskedasticity and autocorrelation consistent (HAC) standard error estimate for $\bar{h}_{N|N_0}$ based on the output of a single chain.

Fig. 33 depicts univariate prior and posterior densities, which are obtained by applying a standard kernel density estimator to draws from the prior and posterior distribution. In addition, one can also compute posterior credible sets based on the output of the

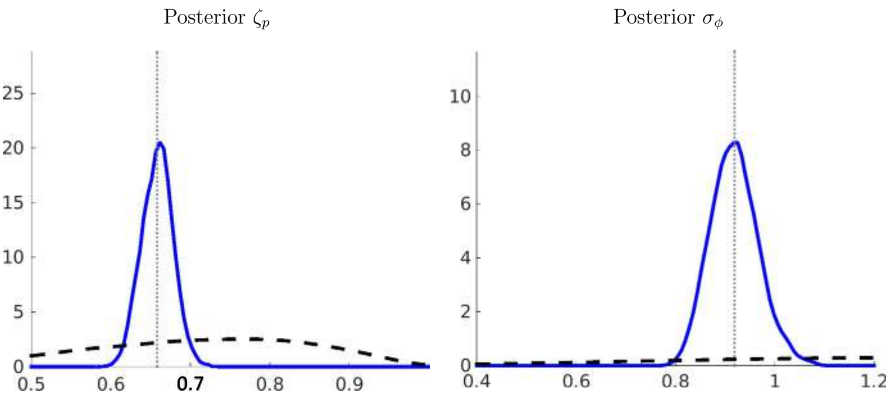


Fig. 33 Prior and posterior densities. *Notes*: The *dashed lines* represent the prior densities, whereas the *solid lines* correspond to the posterior densities of ζ_p and σ_ϕ. The posterior is based on a simulated sample of observations of size $T = 80$. We generate $N = 37{,}500$ draws from the posterior and drop the first $N_0 = 7{,}500$ draws.

posterior sampler. For a univariate parameter, the shortest credible set is given by the highest-posterior-density (HPD) set defined as

$$CS_{HPD}(Y) = \{\theta \, | p(\theta|Y) \geq \kappa_\alpha\}, \tag{215}$$

where κ_α is chosen to ensure that the credible set has the desired posterior coverage probability.

12.2.4 Blocking

Despite a careful choice of the proposal distribution $q(\cdot|\theta^{i-1})$, it is natural that the efficiency of the MH algorithm decreases as the dimension of the parameter vector θ increases. The success of the proposed random walk move decreases as the dimension d of the parameter space increases. One way to alleviate this problem is to break the parameter vector into blocks. Suppose the dimension of the parameter vector θ is d. A partition of the parameter space, B, is a collection of N_{blocks} sets of indices. These sets are mutually exclusive and collectively exhaustive. Call the subvectors that correspond to the index sets θ_b, $b = 1,\ldots,N_{blocks}$. In the context of a sequence of parameter draws, let θ_b^i refer to the b-th block of i-th draw of θ and let $\theta_{<b}^i$ refer to the i-th draw of all of the blocks before b and similarly for $\theta_{>b}^i$. Algorithm 8 describes a generic Block MH algorithm.

Algorithm 8 (Block MH Algorithm). Draw $\theta^0 \in \Theta$ and then for $i = 1$ to N:

1. Create a partition B^i of the parameter vector into N_{blocks} blocks $\theta_1,\ldots,\theta_{N_{blocks}}$ via some rule (perhaps probabilistic), unrelated to the current state of the Markov chain.
2. For $b = 1,\ldots,N_{blocks}$:

(a) Draw $\vartheta_b \sim q(\cdot|[\theta^i_{<b},\, \theta^{i-1}_b,\, \theta^{i-1}_{\geq b}])$.

(b) With probability,

$$\alpha = \max\left\{\frac{p([\theta^i_{<b},\vartheta_b,\theta^{i-1}_{>b}]|Y)q(\theta^{i-1}_b,|\theta^i_{<b},\vartheta_b,\theta^{i-1}_{>b})}{p(\theta^i_{<b},\theta^{i-1}_b,\theta^{i-1}_{>b}|Y)q(\vartheta_b|\theta^i_{<b},\theta^{i-1}_b,\theta^{i-1}_{>b})},1\right\},$$

set $\theta^i_b = \vartheta_b$, otherwise set $\theta^i_b = \theta^{i-1}_b$.

In order to make the Block MH algorithm operational, the researcher has to decide how to allocate parameters to blocks in each iteration and how to choose the proposal distribution $q(\cdot|[\theta^i_{<b},\theta^{i-1}_b,\theta^{i-1}_{>b}])$ for parameters of block b.

A good rule of thumb, however, is that we want the parameters *within* a block, say, θ^b, to be as correlated as possible, while we want the parameters between blocks, say, θ_b and θ_{-b}, to be as independent as possible, according to Robert and Casella (2004). Unfortunately, picking the "optimal" blocks to minimize dependence across blocks requires *a priori* knowledge about the posterior and is therefore often infeasible. Chib and Ramamurthy (2010) propose grouping parameters randomly. Essentially, the user specifies how many blocks to partition the parameter vector into and every iteration a new set of blocks is constructed. Key to the algorithm is that the block configuration be independent of the Markov chain. This is crucial for ensuring the convergence of the chain.

In order to tailor the block-specific proposal distributions, Chib and Ramamurthy (2010) advocate using an optimization routine—specifically, simulated annealing—to find the mode of the conditional posterior distribution. As in the RWMH-V algorithm, the variance of the proposal distribution is based on the inverse Hessian of the conditional log posterior density evaluated at the mode. Unfortunately, the tailoring requires many likelihood evaluations that slow down the algorithm and a simpler procedure, such as using marginal or conditional covariance matrices from an initial approximation of the joint posterior covariance matrix, might be computationally more efficient.

12.2.5 Marginal Likelihood Approximations

The computations thus far do not rely on the marginal likelihood $p(Y)$, which appears in the denominator of Bayes Theorem. Marginal likelihoods play an important role in assessing the relative fit of models because they are used to turn prior model probabilities into posterior probabilities. The most widely used marginal likelihood approximation in the DSGE model literature is the modified harmonic mean estimator proposed by Geweke (1999). This estimator is based on the identity

$$\int \frac{f(\theta)}{p(Y)} d\theta = \int \frac{f(\theta)}{p(Y|\theta)p(\theta)} p(\theta|Y)d\theta, \tag{216}$$

where $f(\theta)$ has the property that $\int f(\theta)d\theta = 1$. The identity is obtained by rewriting Bayes Theorem, multiplying both sides with $f(\theta)$ and integrating over θ. Realizing that the left-hand side simplifies to $1/p(Y)$ and that the right-hand side can be approximated by a Monte Carlo average we obtain

$$\hat{p}_{HM}(Y) = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{f(\theta^i)}{p(Y|\theta^i)p(\theta^i)}\right]^{-1}, \tag{217}$$

where the θ_i's are drawn from the posterior $p(\theta|Y)$. The function $f(\theta)$ should be chosen to keep the variance of $f(\theta^i)/p(Y|\theta^i)p(\theta^i)$ small. Geweke (1999) recommends using for $f(\theta)$ a truncated normal approximation of the posterior distribution for θ that is computed from the output of the posterior sampler. Alternative methods to approximate the marginal likelihood are discussed in Chib and Jeliazkov (2001), Sims et al. (2008), and Ardia et al. (2012). An and Schorfheide (2007) and Herbst and Schorfheide (2015) provide accuracy comparisons of alternative methods.

12.2.6 Extensions

The basic estimation approach for linearized DSGE models has been extended in several dimensions. Typically, the parameter space is restricted to a subspace in which a linearized model has a unique nonexplosive rational expectations solution (determinacy). Lubik and Schorfheide (2004) relax this restriction and also consider the region of the parameter space in which the solution is indeterminate. By computing the posterior probability of parameter values associated with indeterminacy, they are able to conduct a posterior odds assessment of determinacy vs indeterminacy. Justiniano and Primiceri (2008) consider a linearized DSGE model with structural shocks that exhibit stochastic volatility and develop an MCMC algorithm for posterior inference. A further extension is provided by Curdia et al. (2014), who also allow for shocks that, conditional on the volatility process, have a fat-tailed student-t distribution to capture extreme events such as the Great Recession. Schorfheide (2005a) and Bianchi (2013) consider the estimation of linearized DSGE models with regime switching in the coefficients of the state-space representation.

Müller (2012) provides an elegant procedure to assess the robustness of posterior inference to shifts in the mean of the prior distribution. One of the attractive features of his procedure is that the robustness checks can be carried out without having to reestimate the DSGE model under alternative prior distributions. Koop et al. (2013) propose some diagnostics that allow users to determine the extent to which the likelihood function is informative about the DSGE model parameters. In a nutshell, the authors recommend examining whether the variance of marginal posterior distributions shrinks at the rate T^{-1} (in a stationary model) if the number of observations is increased in a simulation experiment.

12.2.7 Particle MCMC

We now turn to the estimation of fully nonlinear DSGE models. As discussed in Section 10, for nonlinear DSGE models the likelihood function has to be approximated by a nonlinear filter. Embedding a particle filter approximation into an MCMC sampler leads to a so-called particle MCMC algorithm. We refer to the combination of a particle-filter approximated likelihood and the MH algorithm as a PFMH algorithm. This idea was first proposed for the estimation of nonlinear DSGE models by Fernández-Villaverde and Rubio-Ramírez (2007). The theory underlying the PFMH algorithm is developed in Andrieu et al. (2010). Flury and Shephard (2011) discuss non-DSGE applications of particle MCMC methods in econometrics. The modification of Algorithm 7 is surprisingly simple: one only has to replace the exact likelihood function $p(Y|\theta)$ with the particle filter approximation $\hat{p}(Y|\theta)$.

Algorithm 9 (PFMH Algorithm). For $i = 1$ to N:

1. Draw ϑ from a density $q(\vartheta|\theta^{i-1})$.

2. Set $\theta^i = \vartheta$ with probability

$$\alpha(\vartheta|\theta^{i-1}) = \min\left\{1, \frac{\hat{p}(Y|\vartheta)p(\vartheta)/q(\vartheta|\theta^{i-1})}{\hat{p}(Y|\theta^{i-1})p(\theta^{i-1})/q(\theta^{i-1}|\vartheta)}\right\}$$

and $\theta^i = \theta^{i-1}$ otherwise. The likelihood approximation $\hat{p}(Y|\vartheta)$ is computed using Algorithm 6.

The surprising implication of the theory developed in Andrieu et al. (2010) is that the distribution of draws generated by Algorithm 9 from the PFMH algorithm that replaces $p(Y|\theta)$ with $\hat{p}(Y|\theta)$ in fact does converge to the exact posterior. The replacement of the exact likelihood function by the particle-filter approximation generally increases the persistence of the Markov chain and makes Monte Carlo approximations less accurate; see Herbst and Schorfheide (2015) for numerical illustrations. Formally, the key requirement is that the particle-filter approximation provide an unbiased estimate of the likelihood function. In practice it has to be ensured that the variance of the numerical approximation is small relative to the expected magnitude of the differential between $p(Y|\theta^{i-1})$ and $p(Y|\vartheta)$ in an ideal version of the algorithm in which the likelihood could be evaluated exactly. Thus, before embedding the particle-filter approximation into a likelihood function, it is important to assess its accuracy for low- and high-likelihood parameter values.

12.3 SMC Methods

Sequential Monte Carlo (SMC) techniques to generate draws from posterior distributions of a static parameter θ are emerging as an attractive alternative to MCMC methods. SMC algorithms can be easily parallelized and, properly tuned, may produce more accurate approximations of posterior distributions than MCMC algorithms. Chopin (2002) showed how to adapt the particle filtering techniques discussed in Section 10.3 to conduct posterior inference for a static parameter vector. Textbook treatments of SMC algorithms can be found, for instance, in Liu (2001) and Cappé et al. (2005).

The first paper that applied SMC techniques to posterior inference in a small-scale DSGE models was Creal (2007). Herbst and Schorfheide (2014) develop the algorithm further, provide some convergence results for an adaptive version of the algorithm building on the theoretical analysis of Chopin (2004), and show that a properly tailored SMC algorithm delivers more reliable posterior inference for large-scale DSGE models with a multimodal posterior than the widely used RWMH-V algorithm. Creal (2012) provides a recent survey of SMC applications in econometrics. Durham and Geweke (2014) show how to parallelize a flexible and self-tuning SMC algorithm for the estimation of time series models on graphical processing units (GPU). The remainder of this section draws heavily from the more detailed exposition in Herbst and Schorfheide (2014, 2015).

SMC combines features of classic importance sampling and modern MCMC techniques. The starting point is the creation of a sequence of intermediate or bridge distributions $\{\pi_n(\theta)\}_{n=0}^{N_\phi}$ that converge to the target posterior distribution, ie, $\pi_{N_\phi}(\theta)=\pi(\theta)$. At any stage the posterior distribution $\pi_n(\theta)$ is represented by a swarm of particles $\{\theta_n^i, W_n^i\}_{i=1}^N$ in the sense that the Monte Carlo average

$$\bar{h}_{n,N}=\frac{1}{N}\sum_{i=1}^{N}W_n^i h(\theta_n^i)\overset{a.s.}{\longrightarrow}\mathbb{E}_{\pi_n}[h(\theta_n)]. \tag{218}$$

The bridge distributions can be generated either by taking power transformations of the entire likelihood function, that is, $[p(Y|\theta)]^{\phi_n}$, where $\phi_n\uparrow 1$, or by adding observations to the likelihood function, that is, $p(Y_{1:t_n}|\theta)$, where $t_n\uparrow T$. We refer to the first approach as likelihood tempering and the second approach as data tempering. Formally, the sequences of bridge distributions are defined as (likelihood tempering)

$$\pi_n(\theta)=\frac{[p(Y|\theta)]^{\phi_n}p(\theta)}{\int[p(Y|\theta)]^{\phi_n}p(\theta)d\theta}\quad n=0,\ldots,N_\phi,\quad \phi_n\uparrow 1, \tag{219}$$

and (data tempering, writing $t_n=\lfloor\phi_n T\rfloor$)

$$\pi_n^{(D)}(\theta)=\frac{p(Y_{1:\lfloor\phi_n T\rfloor})p(\theta)}{\int p(Y_{1:\lfloor\phi_n T\rfloor})p(\theta)d\theta}\quad n=0,\ldots,N_\phi,\quad \phi_n\uparrow 1, \tag{220}$$

respectively. While data tempering is attractive in sequential applications, eg, real-time forecasting, likelihood tempering generally leads to more stable posterior simulators for two reasons: First, in the initial phase it is possible to add information that corresponds to a fraction of an observation. Second, if the latter part of the sample contains influential observations that drastically shift the posterior mass, data tempering may have difficulties adapting to the new information.

12.3.1 The SMC Algorithm

The algorithm can be initialized with draws from the prior density $p(\theta)$, provided the prior density is proper. For the prior in Table 7 it is possible to directly sample independent draws

θ_0^i from the marginal distributions of the DSGE model parameters. One can add an accept-reject step that eliminates parameter draws for which the linearized model does not have a unique stable rational expectations solution. The initial weights W_0^i can be set equal to one. We adopt the convention that the weights are normalized to sum to N.

The SMC algorithm proceeds iteratively from $n = 0$ to $n = N_\phi$. Starting from stage $n-1$ particles $\{\theta_{n-1}^i, W_{n-1}^i\}_{i=1}^N$ each stage n of the algorithm consists of three steps: *correction*, that is, reweighting the stage $n-1$ particles to reflect the density in iteration n; *selection*, that is, eliminating a highly uneven distribution of particle weights (degeneracy) by resampling the particles; and *mutation*, that is, propagating the particles forward using a Markov transition kernel to adapt the particle values to the stage n bridge density.

Algorithm 10 (Generic SMC Algorithm with Likelihood Tempering).

1. **Initialization.** ($\phi_0 = 0$). Draw the initial particles from the prior: $\theta_0^i \overset{iidp}{\sim} p(\theta)$ and $W_0^i = 1$, $i = 1,\ldots,N$.
2. **Recursion.** For $n = 1,\ldots,N_\phi$,
   (a) **Correction.** Reweight the particles from stage $n-1$ by defining the incremental weights
$$\tilde{w}_n^i = [p(Y|\theta_{n-1}^i)]^{\phi_n - \phi_{n-1}} \tag{221}$$
   and the normalized weights
$$\tilde{W}_n^i = \frac{\tilde{w}_n^i W_{n-1}^i}{\frac{1}{N}\sum_{i=1}^N \tilde{w}_n^i W_{n-1}^i,} \qquad i = 1,\ldots,N. \tag{222}$$
   (b) **Selection (Optional).** Resample the particles via multinomial resampling. Let $\{\hat{\theta}\}_{i=1}^N$ denote N *iid* draws from a multinomial distribution characterized by support points and weights $\{\theta_{n-1}^i, \tilde{W}_n^i\}_{i=1}^N$ and set $W_n^i = 1$.
   (c) **Mutation.** Propagate the particles $\{\hat{\theta}_i, W_n^i\}$ via N_{MH} steps of an MH algorithm with transition density $\theta_n^i \sim K_n(\theta_n|\hat{\theta}_n^i; \zeta_n)$ and stationary distribution $\pi_n(\theta)$. An approximation of $\mathbb{E}_{\pi_n}[h(\theta)]$ is given by
$$\bar{h}_{n,N} = \frac{1}{N}\sum_{i=1}^N h(\theta_n^i) W_n^i. \tag{223}$$
3. For $n = N_\phi$ ($\phi_{N_\phi} = 1$) the final importance sampling approximation of $\mathbb{E}_\pi[h(\theta)]$ is given by:
$$\bar{h}_{N_\phi,N} = \sum_{i=1}^N h(\theta_{N_\phi}^i) W_{N_\phi}^i. \tag{224}$$

The correction step is a classic importance sampling step, in which the particle weights are updated to reflect the stage n distribution $\pi_n(\theta)$. Because this step does not change the particle value, it is typically not necessary to reevaluate the likelihood function.

The selection step is optional. On the one hand, resampling adds noise to the Monte Carlo approximation, which is undesirable. On the other hand, it equalizes the particle weights, which increases the accuracy of subsequent importance sampling approximations. The decision of whether or not to resample is typically based on a threshold rule for the variance of the particle weights. As for the particle filter in Section 10.3, we can define an effective particle sample size as:

$$\widehat{ESS}_n = N \Big/ \left(\frac{1}{N} \sum_{i=1}^{N} (\tilde{W}_n^i)^2 \right) \tag{225}$$

and resample whenever $\widehat{ESS}_n$ is less that $N/2$ or $N/4$. In the description of Algorithm 10 we consider multinomial resampling. Other, more efficient resampling schemes are discussed, for instance, in the books by Liu (2001) or Cappé et al. (2005) (and references cited therein).

The mutation step changes the particle values. In the absence of the mutation step, the particle values would be restricted to the set of values drawn in the initial stage from the prior distribution. This would clearly be inefficient, because the prior distribution is a poor proposal distribution for the posterior in an importance sampling algorithm. As the algorithm cycles through the N_ϕ phases, the particle values successively adapt to the shape of the posterior distribution. The key feature of the transition kernel $K_n(\theta_n|\hat{\theta}_n;\zeta_n)$ is the invariance property:

$$\pi_n(\theta_n) = \int K_n(\theta_n|\hat{\theta}_n;\zeta_n)\pi_n(\hat{\theta}_n)d\hat{\theta}_n. \tag{226}$$

Thus, if $\hat{\theta}_n^i$ is a draw from π_n, then so is θ_n^i. The mutation step can be implemented by using one or more steps of the RWMH algorithm described in Section 12.2.2. The probability of mutating the particles can be increased by blocking or by iterating the RWMH algorithm over multiple steps. The vector ζ_n summarizes the tuning parameters, eg, c and $\hat{\Sigma}$ of the RWMH algorithm.

The SMC algorithm produces as a by-product an approximation of the marginal likelihood. It can be shown that

$$\hat{p}_{SMC}(Y) = \prod_{n=1}^{N_\phi} \left(\frac{1}{N} \sum_{i=1}^{N} \tilde{w}_n^i W_{n-1}^i \right)$$

converges almost surely to $p(Y)$ as the number of particles $N \to \infty$.

12.3.2 Tuning the SMC Algorithm

The implementation of the SMC algorithm requires the choice of several tuning constants. The most important choice is the number of particles N. As shown in Chopin (2004), Monte Carlo averages computed from the output of the SMC algorithm satisfy a CLT as the number of particles increases to infinity. This means that the variance of the Monte Carlo approximation decreases at the rate $1/N$. The user has to determine the number of bridge distributions N_ϕ and the tempering schedule ϕ_n. Based on experiments with a small-scale DSGE model, Herbst and Schorfheide (2015) recommend a convex tempering schedule of the form $\phi_n = (n/N_\phi)^\lambda$ with $\lambda \approx 2$. Durham and Geweke (2014) recently developed a self-tuning algorithm that chooses the sequence ϕ_n adaptively as the algorithm cycles through the stages.

The mutation step requires the user to determine the number of MH steps N_{MH} and the number of parameter blocks. The increased probability of mutation raises the accuracy but unfortunately, the number of likelihood evaluations increases as well, which slows down the algorithm. The scaling constant c and the covariance matrix $\hat{\Sigma}$ can be easily chosen adaptively. Based on the MH rejection frequency, c can be adjusted to achieve a target rejection rate of approximately 25–40%. For $\hat{\Sigma}_n$ one can use an approximation of the posterior covariance matrix computed at the end of the stage n correction step.

To monitor the accuracy of the SMC approximations Durham and Geweke (2014) suggest creating H groups of N particles and setting up the algorithm so that there is no communication across groups. This leads to H Monte Carlo approximations of posterior moments of interest. The across-group standard deviation of within-group Monte Carlo averages provides a measure of numerical accuracy. Parallelization of the SMC algorithm is relatively straightforward because the mutation step and the computation of the incremental weights in the correction step can be carried out in parallel on multiple processors, each of which is assigned a group of particles. In principle, the exact likelihood function can be replaced by a particle-filter approximation, which leads to an SMC^2 algorithm, developed by Chopin et al. (2012) and discussed in more detail in the context of DSGE models in Herbst and Schorfheide (2015).

12.3.3 Numerical Illustration

We now illustrate the SMC model in the context of the stylized DSGE models. The setup is similar to the one in Section 12.2.3. We generate $T = 80$ observations using the parameters listed in Table 5 and use the prior distribution given in Table 7. The algorithm is configured as follows. We use $N = 2048$ particles and $N_\phi = 500$ tempering stages. We set $\lambda = 3$, meaning that we add very little information in the initial stages to ensure that the prior draws adapt to the shape of the posterior. We use one step of a single-block RWMH algorithm in the mutation step and choose c and $\hat{\Sigma}_n$ adaptively as described in Herbst and Schorfheide (2014). The target acceptance rate for the mutation step is 0.25. Based on the output of the SMC algorithm, we plot marginal bridge densities $\pi_n(\cdot)$ for the price stickiness parameter ζ_p and the shock standard deviation σ_ϕ in Fig. 34. The initial set of particles is drawn from the prior distribution. As ϕ_n increases to one, the distribution concentrates. The final stage approximates the posterior distribution.

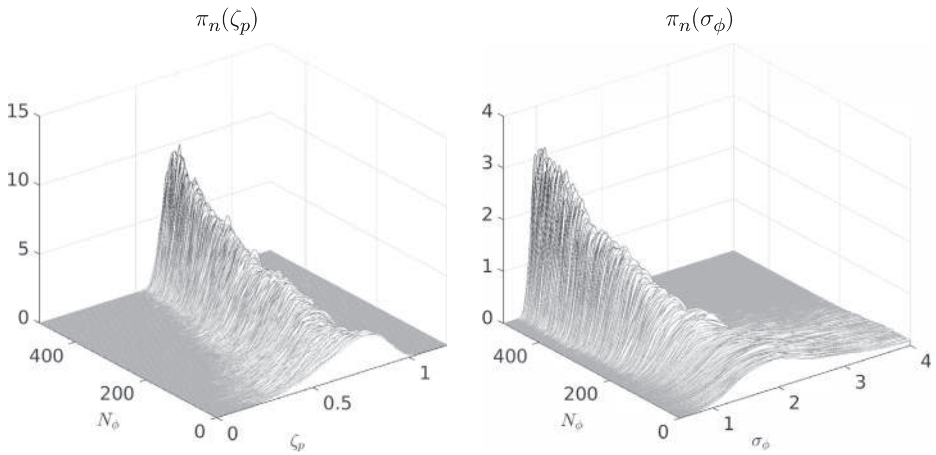


Fig. 34 SMC bridge densities. *Notes*: The posterior is based on a simulated sample of observations of size $T = 80$. The two panels show the sequence of posterior (bridge) densities $\pi_n(\cdot)$.

12.4 Model Diagnostics

DSGE models provide stylized representations of the macroeconomy. To examine whether a specific model is able to capture salient features of the data Y from an *a priori* perspective, prior predictive checks provide an attractive diagnostic. Prior (and posterior) predictive checks are discussed in general terms in the textbooks by Lancaster (2004) and Geweke (2005). The first application of a prior predictive check in the context of DSGE models is Canova (1994).

Let $Y_{1:T}^*$ be an artificial sample of length T. The predictive distribution for $Y_{1:T}^*$ based on the time t information set $\mathcal{F}_t$ is

$$p(Y_{1:T}^*|\mathcal{F}_t) = \int p(Y_{1:T}^*|\theta)p(\theta|\mathcal{F}_t)d\theta. \tag{227}$$

We used a slightly more general notation (to accommodate posterior predictive checks below) with the convention that $\mathcal{F}_0$ corresponds to prior information. The idea of a predictive check is to examine how far the actual realization $Y_{1:T}$ falls into the tail of the predictive distribution. If $Y_{1:T}$ corresponds to an unlikely tail event, then the model is regarded as poorly specified and should be adjusted before it is estimated.

In practice, the high-dimensional vector $Y_{1:T}$ is replaced by a lower-dimensional statistic $\mathcal{S}(Y_{1:T})$, eg, elements of the sample autocovariance matrix $vech(\hat{\Gamma}_{yy}(h))$, for which it is easier to calculate or visualize tail probabilities. While it is not possible to directly evaluate the predictive density of sample statistics, it is straightforward to generate draws. In the case of a prior predictive check, let $\{\theta^i\}_{i=1}^N$ be a sequence of parameter draws from the prior. For each draw, simulate the DSGE model, which leads to the trajectory

$Y^{*i}_{1:T}$. For each of the simulated trajectories, compute the sample statistic $\mathcal{S}(\cdot)$, which leads to a draw from the predictive density.

For a posterior predictive check one equates $\mathcal{F}_t$ with the sample $Y_{1:T}$. The posterior predictive check examines whether the estimated DSGE model captures the salient features of the sample. A DSGE model application can be found in Chang et al. (2007), who examine whether versions of an estimated stochastic growth model are able to capture the variance and the serial correlation of hours worked.

12.5 Limited Information Bayesian Inference

Bayesian inference requires a likelihood function $p(Y|\theta)$. However, as discussed in Section 11, many of the classical approaches to DSGE model estimation, eg, (generalized) methods of moments and impulse response function matching, do not utilize the likelihood function of the DSGE model, in part because there is some concern about misspecification. These methods are referred to as limited-information (instead of full information) techniques. This subsection provides a brief survey of Bayesian approaches to limited-information inference.

12.5.1 Single-Equation Estimation

Lubik and Schorfheide (2005) estimate monetary policy rules for small open economy models by augmenting the policy rule equation with a vector-autoregressive law of motion for the endogenous regressors, eg, the output gap and inflation in the case of our stylized model. This leads to a VAR for output, inflation, and interest rates, with cross-coefficient restrictions that are functions of the monetary policy rule parameters. The restricted VAR can be estimated with standard MCMC techniques. Compared to the estimation of a fully specified DSGE model, the limited-information approach robustifies the estimation of the policy rule equation against misspecification of the private sector's behavior. Kleibergen and Mavroeidis (2014) apply a similar technique to the estimation of a New Keynesian Phillips curve. Their work focuses on the specification of prior distributions that regularize the likelihood function in settings in which the sample only weakly identifies the parameters of interest, eg, the slope of the New Keynesian Phillips curve.

12.5.2 Inverting a Sampling Distribution

Suppose one knows the sampling distribution $p(\hat{\theta}|\theta)$ of an estimator $\hat{\theta}$. Then, instead of updating beliefs conditional on the observed sample Y, one could update the beliefs about θ based on the realization of $\hat{\theta}$:

$$p(\theta|\hat{\theta}) = \frac{p(\hat{\theta}|\theta)p(\theta)}{\int p(\hat{\theta}|\theta)p(\theta)}. \tag{228}$$

This idea dates back at least to Pratt et al. (1965) and is useful in situations in which a variety of different distributions for the sample Y lead to the same distribution of the estimator $\hat{\theta}$. The drawback of this approach is that a closed-form representation of the density $p(\hat{\theta}|\theta)$ is typically not available.

In practice one could use a simulation-based approximation of $p(\hat{\theta}|\theta)$, which is an idea set forth by Boos and Monahan (1986). Alternatively, one could replace the finite-sample distribution with a limit distribution, eg,

$$\sqrt{T}(\hat{\theta}_T - \theta_T)|\theta_T \Longrightarrow N(0, V(\theta)), \tag{229}$$

where the sequence of "true" parameters θ_T converges to θ. This approach is considered by Kwan (1999). In principle $\hat{\theta}_T$ could be any of the frequentist estimators studied in Section 11 for which we derived an asymptotic distribution, including the MD estimator, the IRF matching estimator, or the GMM estimator. However, in order for the resulting limited-information posterior to be meaningful, it is important that the convergence to the asymptotic distribution be uniform in θ, which requires (229) to hold for each sequence $\theta_T \longrightarrow \theta$. A uniform convergence to a normal distribution is typically not attainable as θ_T approaches the boundary of the region of the parameter space in which the time series $Y_{1:T}$ is stationary.

Rather than making statements about the approximation of the limited-information posterior distribution $p(\theta|\hat{\theta})$, Müller (2013) adopts a decision-theoretic framework and shows that decisions based on the quasi-posterior that is obtained by inverting the limit distribution of $\hat{\theta}_T|\theta$ are asymptotically optimal (in the sense that they minimize expected loss) under fairly general conditions. Suppose that the likelihood function of a DSGE model is misspecified. In this case the textbook analysis of the ML estimator in Section 11.1 has to be adjusted as follows. The information matrix equality that ensures that $\| -\nabla^2_\theta \ell_T(\theta|Y) - \mathcal{I}(\theta_0) \|$ converges to zero is no longer satisfied. If we let $D = \text{plim}_{T\rightarrow\infty} - \nabla^2_\theta \ell_T(\theta|Y)$, then the asymptotic variance of the ML estimator takes the sandwich form $D\mathcal{I}(\theta_0)D'$. Under the limited-information approach coverage sets for individual DSGE model parameters would be computed based on the diagonal elements of $D\mathcal{I}(\theta_0)D'$, whereas under a full-information Bayesian approach with misspecified likelihood function, the coverage sets would (asymptotically) be based on $\mathcal{I}^{-1}(\theta_0)$. Thus, the limited-information approach robustifies the coverage sets against model misspecification.

Instead of inverting a sampling distribution of an estimator, one could also invert the sampling distribution of some auxiliary sample statistic $\hat{\varphi}(Y)$. Not surprisingly, the main obstacle is the characterization of the distribution $\hat{\varphi}|\theta$. A collection of methods referred to as approximate Bayesian computations (ABC) use a simulation approximation of $p(\hat{\varphi}|\theta)$ and they could be viewed as a Bayesian version of indirect inference. These algorithms target

$$p^{\delta}_{ABC}(\theta, \hat{\varphi}^*|\hat{\varphi}) \propto p(\hat{\varphi}^*|\theta)p(\theta)\mathbb{I}\{\| \hat{\varphi}^* - \hat{\varphi} \| \leq \delta\}, \tag{230}$$

where $\hat{\varphi}$ refers to the auxiliary statistic computed from the observed data, $\hat{\varphi}^*$ is the auxiliary statistic computed from data simulated from the model conditional on a parameter θ, and δ is the level of tolerance for discrepancies between model-simulated and observed statistics. To date, there are few applications of ABC in econometrics. Forneron and Ng (2015) discuss the relationship between ABC and the simulated MD estimators introduced in Section 11.2 and Scalone (2015) explores a DSGE model application.

12.5.3 Limited-Information Likelihood Functions

Kim (2002) constructs a limited-information likelihood function from the objective function of an extremum estimator. For illustrative purposes we consider the GMM estimator discussed in Section 11.4, but the same idea can also be applied to the MD estimator and the IRF matching estimator. Suppose the data are generated under the probability measure $\mathbb{P}$ and at $\theta = \theta_0$ the following GMM moment condition is satisfied: $\mathbb{E}_{\mathbb{P}}[g(y_{t-p:t}|\theta_0)] = 0$. The sample objective function $Q_T(\theta|Y)$ for the resulting GMM estimator based on a weight matrix W was given in (207). Assuming uniform integrability of the sample objective function

$$\lim_{T\to\infty} \mathbb{E}_{\mathbb{P}}[Q_T(\theta_0|Y)] = r \tag{231}$$

where r is the number of overidentifying moment conditions (meaning the difference between the number of moments stacked in the vector $g(\cdot)$ and the number of elements of the parameter vector θ).

Let $\mathcal{P}(\theta)$ denote the collection of probability distributions that satisfy the moment conditions in the following sense:

$$\mathcal{P}(\theta) = \left\{ P \mid \lim_{T\to\infty} \mathbb{E}_P[TQ_T(\theta|Y)] = r \right\}. \tag{232}$$

$\mathcal{P}(\theta)$ cannot be used directly for likelihood-based inference because it comprises a collection of probability distributions indexed by θ. To obtain a unique distribution for each θ, Kim (2002) projects the "true" distribution $\mathbb{P}$ onto the set $\mathcal{P}(\theta)$ using the Kullback–Leibler discrepancy as the metric:

$$P^*(Y|\theta) = \text{argmin}_{P\in\mathcal{P}(\theta)} \int \log(dP/d\mathbb{P})dP. \tag{233}$$

The solution takes the convenient form

$$p^*(Y|\theta) \propto \exp\left\{ -\frac{1}{2} Q_T(\theta|Y) \right\}, \tag{234}$$

where $p^*(Y|\theta) = dP/d\mathbb{P}$ is the Radon–Nikodym derivative of P with respect to $\mathbb{P}$.

Kim's (2002) results suggest that the frequentist objective functions of Sections 11.2–11.4 can be combined with a prior density and used for (limited information) Bayesian inference. The posterior mean

$$\hat{\theta} = \frac{\int \theta \exp\left\{-\frac{1}{2}Q_T(\theta|Y)\right\}p(\theta)d\theta}{\int \exp\left\{-\frac{1}{2}Q_T(\theta|Y)\right\}p(\theta)d\theta} \tag{235}$$

resembles the LT estimator discussed in Section 11.2. The main difference is that the LT estimator was interpreted from a frequentist perspective, whereas the quasi-posterior based on $p^*(Y|\theta)$ and statistics such as the posterior mean are meant to be interpreted from a Bayesian perspective. This idea has been recently exploited by Christiano et al. (2010) to propose a Bayesian IRF matching estimator. An application to an asset pricing model is presented in Gallant (2015) and an extension to models with latent variables is provided in Gallant et al. (2013). Inoue and Shintani (2014) show that the limited information marginal likelihood

$$p^*(Y|M) = \int p^*(Y|\theta, M)p(\theta)d\theta$$

can be used as a model selection criterion that asymptotically is able to select a correct model specification.

12.5.4 Nonparametric Likelihood Functions

There is also a literature on nonparametric likelihood functions that are restricted to satisfy model-implied moment conditions. Lazar (2003) and Schennach (2005) use empirical likelihood functions, which, roughly speaking assign probability p_t to observation y_t such that the likelihood function is written as $\prod_{t=1}^{T} p_t$, at least if the data are *iid*. One then imposes the side constraint $\sum_{t=1}^{T} p_t g(y_{t-p:t}|\theta) = 0$ and concentrates out p_t probabilities to obtain a profile objective function that only depends on θ. This method is designed for *iid* data and possible models in which $g(y_{t-p:t}|\theta)$ is a martingale difference sequence. Kitamura and Otsu (2011) propose to using a Dirichlet process to generate a prior for the distribution of $Y_{1:T}$ and then project this distribution on the set of distributions that satisfies the moment restrictions. Shin (2014) uses a Dirichlet process mixture and provides a time series extension.

13. CONCLUSION

Over the past two decades the development and application of solution and estimation methods for DSGE models have experienced tremendous growth. Part of this growth has been spurred by central banks, which have included DSGE models in their suites of models used for forecasting and policy analysis. The rapid rise of computing power has enabled researchers to study more and more elaborate model specifications. As we have been writing this chapter, new methods have been developed and novel applications have been explored. While it is impossible to provide an exhaustive treatment of such a dynamic field, we hope that this chapter provides a thorough training for those who are interested in working in this area, offers a good overview of the state of the art as of 2015, and inspires innovative research that expands the frontier of knowledge.

ACKNOWLEDGMENTS

J.F.-V. and J.F.R.-R. gratefully acknowledges financial support from the National Science Foundation under Grant SES 1223271. F.S. gratefully acknowledges financial support from the National Science Foundation under Grant SES 1424843. Minsu Chang, Eugenio Rojas, and Jacob Warren provided excellent research assistance. We thank our discussant Serguei Maliar, the editors John Taylor and Harald Uhlig, John Cochrane, Michael Curran, Oren Levintal, Ben Moll, Eric Swanson, and the participants at the Handbook Conference hosted by the Hoover Institution for helpful comments and suggestions. Computer code to implement the solution and estimation techniques discussed in this chapter is available on the authors' websites.

REFERENCES

Aldrich, E.M., 2014. GPU computing in economics. In: Schmedders, K., Judd, K.L. (Eds.), Handbook of Computational Economics, vol. 3. Elsevier, Amsterdam, pp. 557–598.

Aldrich, E.M., Kung, H., 2011. Computational methods for production-based asset pricing models with recursive utility. Economic Research Initiatives at Duke (ERID) Working Paper Series 87.

Aldrich, E.M., Fernández-Villaverde, J., Gallant, A.R., Rubio-Ramírez, J.F., 2011. Tapping the supercomputer under your desk: solving dynamic equilibrium models with graphics processors. J. Econ. Dyn. Control 35, 386–393.

Algan, Y., Allais, O., Den Haan, W.J., Rendahl, P., 2014. Solving and simulating models with heterogeneous agents and aggregate uncertainty. In: Schmedders, K., Judd, K.L. (Eds.), Handbook of Computational Economics, vol. 3. Elsevier, Amsterdam, pp. 277–324.

Altig, D., Christiano, L., Eichenbaum, M., Linde, J., 2011. Firm-specific capital, nominal rigidities and the business cycle. Rev. Econ. Dyn. 14 (2), 225–247. http://ideas.repec.org/a/red/issued/09-191.html.

Altug, S., 1989. Time-to-build and aggregate fluctuations: some new evidence. Int. Econ. Rev. 30 (4), 889–920.

Alvarez, F., Jermann, U.J., 2004. Using asset prices to measure the cost of business cycles. J. Polit. Econ. 112, 1223–1256.

An, S., Schorfheide, F., 2007. Bayesian analysis of DSGE models. Econ. Rev. 26 (2-4), 113–172. http://dx.doi.org/10.1080/07474930701220071.

Anderson, E.W., Hansen, L.P., McGrattan, E.R., Sargent, T.J., 1996. On the mechanics of forming and estimating dynamic linear economies. In: Amman, H.M., Kendrick, D.A., Rust, J. (Eds.), Handbook of Computational Economics, vol. 1. Elsevier, Amsterdam, pp. 171–252.

Andreasen, M.M., 2013. Non-linear DSGE models and the central difference Kalman filter. J. Appl. Econ. 28 (6), 929–955.

Andreasen, M.M., Fernández-Villaverde, J., Rubio-Ramírez, J.F., 2013. The pruned state-space system for nonlinear DSGE models: theory and empircal applications. NBER Working 18983.

Andrews, I., Mikusheva, A., 2015. Maximum likelihood inference in weakly identified DSGE models. Quant. Econ 6 (1), 123–152.

Andrieu, C., Doucet, A., Holenstein, R., 2010. Particle Markov chain Monte Carlo methods. J. R. Stat. Soc. Ser. B 72 (3), 269–342.

Ardia, D., Bastürk, N., Hoogerheide, L., van Dijk, H.K., 2012. A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. Comput. Stat. Data Anal. 56 (11), 3398–3414.

Arulampalam, S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-gaussian Bayesian tracking. IEEE Trans. Signal Proc. 50 (2), 174–188.

Aruoba, S.B., Schorfheide, F., 2015. Inflation during and after the zero lower bound. Manuscript, University of Maryland.

Aruoba, S.B., Fernández-Villaverde, J., Rubio-Ramírez, J.F., 2006. Comparing solution methods for dynamic equilibrium economies. J. Econ. Dyn. Control 30 (12), 2477–2508.

Barthelmann, V., Novak, E., Ritter, K., 2000. High dimensional polynomial interpolation on sparse grids. Adv. Comput. Math. 12, 273–288.

Bastani, H., Guerrieri, L., 2008. On the application of automatic differentiation to the likelihood function for dynamic general equilibrium models. In: Bischof, C.H., Bücker, H.M., Hovland, P., Naumann, U., Utke, J. (Eds.), Advances in Automatic Differentiation: Lecture Notes in Computational Science and Engineering, vol. 64. Springer, pp. 303–313.

Bender, C.M., Orszag, S.A., 1999. Advanced Mathematical Methods for Scientists and Engineers: Asymptotic Methods and Perturbation Theory. Springer.

Benigno, P., Woodford, M., 2004. Optimal monetary and fiscal policy: a linear-quadratic approach. In: NBER Macroeconomics Annual 2003, vol. 18. MIT Press, Cambridge, MA, pp. 271–364.

Bernanke, B.S., Gertler, M., 1989. Agency costs, net worth, and business fluctuations. Am. Econ. Rev. 79 (1), 14–31.

Bernanke, B.S., Gertler, M., Gilchrist, S., 1999. The financial accelerator in a quantitative business cycle framework. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics, vol. 1. Elsevier, pp. 1341–1393.

Bianchi, F., 2013. Regime switches, agents' beliefs, and post-world war II U.S. macroeconomic dynamics. Rev. Econ. Stud. 80 (2), 463–490.

Blanchard, O.J., Kahn, C.M., 1980. The solution of linear difference models under rational expectations. Econometrica 48 (5), 1305–1312.

Bloom, N., 2009. The impact of uncertainty shocks. Econometrica 77, 623–685.

Bocola, L., 2015. The pass-through of sovereign risk. Manuscript, Northwestern University.

Boos, D.D., Monahan, J.F., 1986. Bootstrap methods using prior information. Biometrika 73 (1), 77–83.

Boyd, J.P., 2000. Chebyshev and Fourier Spectral Methods. Dover.

Boyd, J.P., Petschek, R.G., 2014. The relationships between Chebyshev, Legendre and Jacobi polynomials: the generic superiority of Chebyshev polynomials and three important exceptions. J. Sci. Comput. 59, 1–27.

Brenner, S., Scott, R., 2008. The Mathematical Theory of Finite Element Methods. Springer Verlag.

Brown, D.B., Smith, J.E., Peng, S., 2010. Information relaxations and duality in stochastic dynamic programs. Operat. Res. 58 (4), 785–801.

Brumm, J., Scheidegger, S., 2015. Using adaptive sparse grids to solve high-dimensional dynamic models. Manuscript, University of Zurich.

Bungartz, H.J., Griebel, M., 2004. Sparse grids. Acta Numer. 13, 147–269.

Burnside, C., Eichenbaum, M., Rebelo, S., 1993. Labor hoarding and the business cycle. J. Polit. Econ. 101 (2), 245–273. http://ideas.repec.org/a/ucp/jpolec/v101y1993i2p245-73.html.

Cai, Y., Judd, K.L., 2014. Advances in numerical dynamic programming and new applications. In: Schmedders, K., Judd, K.L. (Eds.), Handbook of Computational Economics, vol. 3. Elsevier, pp. 479–516.

Caldara, D., Fernández-Villaverde, J., Rubio-Ramírez, J.F., Yao, W., 2012. Computing DSGE models with recursive preferences and stochastic volatility. Rev. Econ. Dyn. 15, 188–206.

Canova, F., 1994. Statistical inference in calibrated models. J. Appl. Econ. 9, S123–S144.

Canova, F., 2007. Methods for Applied Macroeconomic Research. Princeton University Press.

Canova, F., 2014. Bridging cyclical DSGE models and the raw data. J. Monet. Econ. 67, 1–15. http://www.crei.cat/people/canova/pdf/%20files/dsge_trend.pdf.

Canova, F., De Nicoló, G., 2002. Monetary disturbances matter for business fluctuations in the G-7. J. Monet. Econ. 49 (4), 1131–1159. http://ideas.repec.org/a/ijc/ijcjou/y2007q4a4.html.

Canova, F., Ferroni, F., Matthes, C., 2014. Choosing the variables to estimate singular DSGE models. J. Appl. Econ. 29 (7), 1099–1117.

Cappé, O., Moulines, E., Ryden, T., 2005. Inference in Hidden Markov Models. Springer Verlag.

Cappé, O., Godsill, S.J., Moulines, E., 2007. An overview of existing methods and recent advances in sequential Monte Carlo. Proc. IEEE 95 (5), 899–924.

Carlstrom, C., Fuerst, T.S., 1997. Agency costs, net worth, and business fluctuations: a computable general equilibrium analysis. Am. Econ. Rev. 87, 893–910.

Chang, Y., Doh, T., Schorfheide, F., 2007. Non-stationary hours in a DSGE model. J. Money, Credit, Bank. 39 (6), 1357–1373.

Chari, V.V., Kehoe, P.J., McGrattan, E.R., 2008. Are structural VARs with long-run restrictions useful in developing business cycle theory? J. Monet. Econ. 55 (8), 1337–1352.

Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. In: Heckman, J.J., Leamer, E. E. (Eds.), Handbook of Econometrics, vol. 6. Elsevier, pp. 5549–5632.

Chernozhukov, V., Hong, H., 2003. An MCMC approach to classical estimation. J. Econ. 115, 293–346.

Chib, S., Greenberg, E., 1995. Understanding the metropolis-hastings algorithm. Am. Stat. 49, 327–335.

Chib, S., Jeliazkov, I., 2001. Marginal likelihoods from the metropolis hastings output. J. Am. Stat. Assoc. 96 (453), 270–281.

Chib, S., Ramamurthy, S., 2010. Tailored randomized block MCMC methods with application to DSGE models. J. Econ. 155 (1), 19–38.

Cho, J.O., Cooley, T.F., Kim, H.S.E., 2015. Business cycle uncertainty and economic welfare. Rev. Econ. Dyn. 18, 185–200.

Chopin, N., 2002. A sequential particle filter for static models. Biometrika 89 (3), 539–551.

Chopin, N., 2004. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. Ann. Stat. 32 (6), 2385–2411.

Chopin, N., Jacob, P.E., Papaspiliopoulos, O., 2012. SMC^2: an efficient algorithm for sequential analysis of state-space models. ArXiv:1101.1528.

Christiano, L.J., 1990. Linear-quadratic approximation and value-function iteration: a comparison. J. Bus. Econ. Stat. 8, 99–113.

Christiano, L.J., Eichenbaum, M., 1992. Current real-business-cycle theories and aggregate labor-market fluctuations. Am. Econ. Rev. 82 (3), 430–450. http://ideas.repec.org/a/aea/aecrev/v82y1992i3p430-50.html.

Christiano, L.J., Fisher, J.D.M., 2000. Algorithms for solving dynamic models with occasionally binding constraints. J. Econ. Dyn. Control 24, 1179–1232.

Christiano, L.J., Vigfusson, R.J., 2003. Maximum likelihood in the frequency domain: the importance of time-to-plan. J. Monet. Econ. 50 (4), 789–815. http://ideas.repec.org/a/eee/moneco/v50y2003i4p789-815.html.

Christiano, L.J., Eichenbaum, M., Evans, C.L., 1999. Monetary policy shocks: what have we learned and to what end. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics, vol. 1a. North Holland, Amsterdam, pp. 65–148.

Christiano, L.J., Eichenbaum, M., Evans, C.L., 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. J. Polit. Econ. 113 (1), 1–45.

Christiano, L.J., Eichenbaum, M., Vigfusson, R., 2007. Assessing structural VARs. In: Acemoglu, D., Rogoff, K., Woodford, M. (Eds.), NBER Macroeconomics Annual 2006, vol. 21. MIT Press, Cambridge, pp. 1–72.

Christiano, L.J., Trabandt, M., Walentin, K., 2010. Dsge models for monetary policy analysis. In: Friedman, B.M., Woodford, M. (Eds.), Handbook of Monetary Economics, vol. 3. Elsevier, pp. 285–367. http://ideas.repec.org/h/eee/monchp/3-07.html.

Christiano, L.J., Eihenbaum, M., Rebelo, S.T., 2011. When is the government spending multiplier large? J. Polit. Econ. 119 (1), 78–121.

Christiano, L.J., Motto, R., Rostagno, M., 2014. Risk shocks. Am. Econ. Rev. 104, 27–65.

Clough, R.W., 1960. The finite element method in plane stress analysis. In: Proceedings of the 2nd ASCE Conference on Electronic Computation.

Clough, R.W., Wilson, E.L., 1999. Early finite element research at Berkeley. Manuscript, University of California, Berkeley.

Cochrane, J.H., 1994. Shocks. Carnegie Rochester Conf. Ser. Publ. Pol. 41 (4), 295–364. http://ideas.repec.org/a/ijc/ijcjou/y2007q4a4.html.

Cochrane, J.H., 2011. Determinacy and identification with Taylor rules. J. Polit. Econ. 119 (3), 565–615. http://ideas.repec.org/a/ucp/jpolec/doi10.1086-660817.html.

Creal, D., 2007. Sequential Monte Carlo samplers for Bayesian DSGE models. Manuscript, Chicago Booth.

Creal, D., 2012. A survey of sequential Monte Carlo methods for economics and finance. Econ. Rev. 31 (3), 245–296.

Crisan, D., Rozovsky, B., (Eds.)2011. The Oxford Handbook of Nonlinear Filtering. Oxford University Press.

Curdia, V., Del Negro, M., Greenwald, D.L., 2014. Rare shocks, great recessions. J. Appl. Econ. 29 (7), 1031–1052.

DeJong, D.N., Dave, C., 2007. Structural Macroeconometrics. Princeton University Press.

DeJong, D.N., Ingram, B.F., Whiteman, C.H., 2000. A Bayesian approach to dynamic macroeconomics. J. Econ. 98 (2), 203–223.

Del Moral, P., 2004. Feynman-Kac Formulae. Springer Verlag.

Del Moral, P., 2013. Mean Field Simulation for Monte Carlo Integration. Chapman & Hall/CRC.

Del Negro, M., Schorfheide, F., 2004. Priors from general equilibrium models for VARs. Int. Econ. Rev. 45 (2), 643–673.

Del Negro, M., Schorfheide, F., 2008. Forming priors for DSGE models (and how it affects the assessment of nominal rigidities). J. Monet. Econ. 55 (7), 1191–1208. ISSN 0304-3932. http://dx.doi.org/10.1016/j.jmoneco.2008.09.006. http://www.sciencedirect.com/science/article/B6VBW-4TKPVGT-3/2/508d89fdb8eb927643250b7f36aab161.

Del Negro, M., Schorfheide, F., 2009. Monetary policy with potentially misspecified models. Am. Econ. Rev. 99 (4), 1415–1450. http://www.econ.upenn.edu/schorf/papers/mpol_p11.pdf.

Del Negro, M., Schorfheide, F., 2011. Bayesian macroeconometrics. In: van Dijk, H., Koop, G., Geweke, J. (Eds.), Handbook of Bayesian Econometrics. Oxford University Press, pp. 293–389.

Del Negro, M., Schorfheide, F., 2013. DSGE model-based forecasting. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting, vol. 2. North Holland, Amsterdam, pp. 57–140.

Del Negro, M., Schorfheide, F., Smets, F., Wouters, R., 2007. On the fit of new Keynesian models. J. Bus. Econ. Stat. 25 (2), 123–162.

Del Negro, M., Hasegawa, R., Schorfheide, F., 2014. Dynamic prediction pools: an investigation of financial frictions and forecasting performance. NBER Working Paper 20575.

Delvos, F.J., 1982. d-Variate Boolean interpolation. J. Approx. Theory 34, 99–114.

Demkowicz, L., 2007. Computing with hp-Adaptive Finite Elements, Volume 1. Chapman & Hall/CRC.

Den Haan, W.J., De Wind, J., 2012. Nonlinear and stable perturbation-based approximations. J. Econ. Dyn. Control 36, 1477–1497.

Den Haan, W.J., Marcet, A., 1990. Solving the stochastic growth model by parameterizing expectations. J. Bus. Econ. Stat. 8 (1), 31–34.

Den Haan, W.J., Marcet, A., 1994. Accuracy in simulations. Rev. Econ. Stud. 61, 3–17.

Díaz-Giménez, J., 1999. Linear-quadratic approximations: an introduction. In: Marimon, R., Scott, A. (Eds.), Computational Methods for the Study of Dynamic Economies. Oxford University Press.

Diebold, F.X., Ohanian, L.E., Berkowitz, J., 1998. Dynamic equilibrium economies: a framework for comparing models and data. Rev. Econ, Stud. 65 (3), 433–452.

Doucet, A., Johansen, A.M., 2011. A tutorial on particle filtering and smoothing: fifteen years later. In: Crisan, D., Rozovsky, B. (Eds.), Handook of Nonlinear Filtering. Oxford University Press.

Doucet, A., de Freitas, N., Gordon, N., 2001. Sequential Monte Carlo Methods in Practice. Springer Verlag.

Dridi, R., Guay, A., Renault, E., 2007. Indirect inference and calibration of dynamic stochastic general equilibrium models. J. Econ. 136 (2), 397–430.

Dufour, J.M., Khalaf, L., Kichian, M., 2013. Identification-robust analysis of DSGE and structural macroeconomic models. J. Monet. Econ. 60, 340–350.

Durbin, J., Koopman, S.J., 2001. Time Series Analysis by State Space Methods. Oxford University Press.

Durham, G., Geweke, J., 2014. Adaptive sequential posterior simulators for massively parallel computing environments. Adv. Econ. 34, 1–44.

Eggertsson, G.B., Woodford, M., 2003. The zero bound on interest rates and optimal monetary policy. Brook. Pap. Econ. Act. 34, 139–235.

Epstein, L.G., Zin, S.E., 1989. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: a theoretical framework. Econometrica 57, 937–969.

Erdös, P., Turán, P., 1937. On interpolation I. Quadrature and mean convergence in the Lagrange interpolation. Ann. Math. 38, 142–155.

Fair, R.C., Taylor, J.B., 1983. Solution and maximum likelihood estimation of dynamic nonlinear rational expectations models. Econometrica 51, 1169–1185.

Faust, J., 1998. The robustness of identified VAR conclusions about money. Carnegie Rochester Conf. Ser. Publ. Pol. 49 (4), 207–244. http://ideas.repec.org/a/ijc/ijcjou/y2007q4a4.html.

Fernández-Villaverde, J., 2010. Fiscal policy in a model with financial frictions. Am. Econ. Rev. Pap. Proc. 100, 35–40.

Fernández-Villaverde, J., Levintal, O., 2016. Solution methods for models with rare disasters. Manuscript, University of Pennsylvania.

Fernández-Villaverde, J., Rubio-Ramírez, J.F., 2004. Comparing dynamic equilibrium models to data: a Bayesian approach. J. Econ. 123 (1), 153–187.

Fernández-Villaverde, J., Rubio-Ramírez, J.F., 2006. Solving DSGE models with perturbation methods and a change of variables. J. Econ. Dyn. Control 30, 2509–2531.

Fernández-Villaverde, J., Rubio-Ramírez, J.F., 2007. Estimating macroeconomic models: a likelihood approach. Rev. Econ. Stud. 74 (4), 1059–1087.

Fernández-Villaverde, J., Rubio-Ramírez, J.F., 2008. How structural are structural parameters? In: Acemoglu, D., Rogoff, K., Woodford, M. (Eds.), NBER Macroeconomics Annual 2007, vol. 22. University of Chicago Press, Chicago, IL.

Fernández-Villaverde, J., Rubio-Ramírez, J.F., Santos, M.S.S., 2006. Convergence properties of the likelihood of computed dynamic models. Econometrica 74 (1), 93–119. http://dx.doi.org/10.1111/j.1468-0262.2006.00650.x.

Fernández-Villaverde, J., Rubio-Ramírez, J.F., Sargent, T.J., Watson, M.W., 2007. ABCs (and Ds) of understanding VARs. Am. Econ. Rev. 97 (3), 1021–1026.

Fernández-Villaverde, J., Guerrón-Quintana, P.A., Rubio-Ramírez, J.F., Uribe, M., 2011. Risk matters: the real effects of volatility shocks. Am. Econ. Rev. 101, 2530–2561.

Fernández-Villaverde, J., Guerrón-Quintana, P.A., Rubio-Ramírez, J.F., 2014. Supply-side policies and the zero lower bound. IMF Econ. Rev. 62, 248–260.

Fernández-Villaverde, J., Gordon, G., Guerrón-Quintana, P.A., Rubio-Ramírez, J.F., 2015a. Nonlinear adventures at the zero lower bound. J. Econ. Dyn. Control 57, 182–204.

Fernández-Villaverde, J., Guerrón-Quintana, P.A., Rubio-Ramírez, J.F., 2015b. Estimating dynamic equilibrium models with stochastic volatility. J. Econ. 185, 216–229.

Flury, T., Shephard, N., 2011. Bayesian inference based only on simulated likelihood: particle filter analysis of dynamic economic models. Econ. Theory 27, 933–956.

Fornberg, B., 1996. A Practical Guide to Pseudospectral Methods. Cambridge University Press.

Forneron, J.J., Ng, S., 2015. The ABC of simulation estimation with auxiliary statistics. Manuscript, Columbia University.

Gallant, A.R., 2015. Reflections on the probability space induced by moment conditions with implications for Bayesian inference. J. Fin. Econ. Forthcoming. http://jfec.oxfordjournals.org/content/early/2015/05/28/jjnec.nbv008.abstract.

Gallant, A.R., Giacomini, R., Ragusa, G., 2013. Generalized method of moments with latent variables. CEPR Discussion Papers DP9692.

Galor, O., 2007. Discrete Dynamical Systems. Springer.

Gaspar, J., Judd, K.L., 1997. Solving large-scale rational-expectations models. Macroecon. Dyn. 1, 45–75.

Geweke, J., 1999. Using simulation methods for Bayesian econometric models: inference, development, and communication. Econ. Rev. 18 (1), 1–126.

Geweke, J., 2005. Contemporary Bayesian Econometrics and Statistics. John Wiley & Sons, Inc.

Geweke, J., 2010. Complete and Incomplete Econometric Models. Princeton University Press, Princeton, NJ.

Geweke, J., Amisano, G., 2011. Optimal prediction pools. J. Econ. 164, 130–141.

Geweke, J., Amisano, G., 2012. Prediction with misspecified models. Am. Econ. Rev. Pap. Proc. 103 (3), 482–486.

Gordon, G., 2011. Computing dynamic heterogeneous-agent economies: tracking the distribution. PIER Working Paper 11-018, University of Pennsylvania.

Gorodnichenko, Y., Ng, S., 2010. Estimation of DSGE models when the data are persistent. J. Monet. Econ. 57 (3), 325–340.

Gourieroux, C., Monfort, A., Renault, E., 1993. Indirect inference. J. Appl. Econ. 8, S85–S118.

Gourieroux, C., Phillips, P.C.B., Yu, J., 2010. Indirect inference for dynamic panel models. J. Econ. 157 (1), 68–77.

Guerrón-Quintana, P.A., 2010. What you match does matter: the effects of observable variables on DSGE estimation. J. Appl. Econ. 25, 774–804.

Guerrón-Quintana, P.A., Inoue, A., Kilian, L., 2013. Frequentist inference in weakly identified dynamic stochastic general equilibrium models. Quant. Econ. 4, 197–229.

Guerrón-Quintana, P.A., Inoue, A., Kilian, L., 2014. Impulse response matching estimators for DSGE models. In: Center for Financial Studies (Frankfurt am Main): CFS working paper series, No. 498, CFS working paper seriesWirtschaftswissenschaften URL http://ssrn.com/abstract=2533453.

Guo, D., Wang, X., Chen, R., 2005. New sequential Monte Carlo methods for nonlinear dynamic systems. Stat. Comput. 15, 135–147.

Gust, C., Herbst, E., López-Salido, J.D., Smith, M.E., 2016. The empirical implications of the interest-rate lower bound. Federal Reserve Board.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50 (4), 1029–1054. http://ideas.repec.org/a/ecm/emetrp/v50y1982i4p1029-54.html.

Hansen, G.D., Prescott, E.C., 1995. Recursive methods for computing equilibria of business cycle models. In: Cooley, T.F. (Ed.), Frontiers of Business Cycle Research. Princeton University Press, pp. 39–64.

Hansen, L.P., Sargent, T.J., 2013. Recursive Models of Dynamic Linear Economies. Princeton Press.

Hansen, L.P., Heaton, J.C., Li, N., 2008. Consumption strikes back? Measuring long-run risk. J. Polit. Econ. 116 (2), 260–302.

Herbst, E., Schorfheide, F., 2014. Sequential Monte Carlo sampling for DSGE models. J. Appl. Econ. 29 (7), 1073–1098.

Herbst, E., Schorfheide, F., 2015. Bayesian Estimation of DSGE Models. Princeton University Press.

Hnatkosvaka, V., Marmer, V., Tang, Y., 2012. Comparison of misspecified calibrated models: the minimum distance approach. J. Econ. 169 (1), 131–138.

Hughes, T.J.R., 2000. The Finite Element Method: Linear Static and Dynamic Finite Element Analysis. Dover.

Hurwicz, L., 1962. On the structural form of interdependent systems. In: Nagel, E., Tarski, A. (Eds.), Logic, Methodology and Philosophy of Science. Stanford University Press.

Ingram, B., Whiteman, C., 1994. Supplanting the minnesota prior-forecasting macroeconomic time series using real business cycle model priors. J. Monet. Econ. 49 (4), 1131–1159. http://ideas.repec.org/a/ijc/ijcjou/y2007q4a4.html.

Inoue, A., Shintani, M., 2014. Quasi-Bayesian model selection. Manuscript, Vanderbilt University.

Iskrev, N., 2010. Local identification of DSGE models. J. Monet. Econ. 2, 189–202. http://dx.doi.org/10.1016/j.jmoneco.2009.12.007.

Jin, H.H., Judd, K.L., 2002. Perturbation methods for general dynamic stochastic models. Manuscript, Hoover Institution.

Judd, K., 1998. Numerical Methods in Economics. MIT Press, Cambridge.

Judd, K.L., 1992. Projection methods for solving aggregate growth models. J. Econ. Theory 58, 410–452.

Judd, K.L., 2003. Perturbation methods with nonlinear changes of variables. Manuscript, Hoover Institution.

Judd, K.L., Guu, S.M., 1993. Perturbation solution methods for economic growth models. In: Varian, H. (Ed.), Economic and Financial Modeling with Mathematica. Springer Verlag, pp. 80–103.

Judd, K.L., Guu, S.M., 1997. Asymptotic methods for aggregate growth models. J. Econ. Dyn. Control 21, 1025–1042.

Judd, K.L., Guu, S.M., 2001. Asymptotic methods for asset market equilibrium analysis. Econ. Theory 18, 127–157.

Judd, K.L., Maliar, L., Maliar, S., 2011. How to solve dynamic stochastic models computing expectations just once. NBER Working Paper 17418.

Judd, K.L., Maliar, L., Maliar, S., 2011. Numerically stable and accurate stochastic simulation methods for solving dynamic models. Quant. Econ. 2, 173–210.

Judd, K.L., Maliar, L., Maliar, S., 2014. Lower bounds on approximation errors: testing the hypothesis that a numerical solution is accurate. Manuscript, Hoover Institution.

Judd, K.L., Maliar, L., Maliar, S., Valero, R., 2014. Smolyak method for solving dynamic economic models: Lagrange interpolation, anisotropic grid and adaptive domain. J. Econ. Dyn. Control 44, 92–123.

Justiniano, A., Primiceri, G.E., 2008. The time-varying volatility of macroeconomic fluctuations. Am. Econ. Rev. 98 (3), 604–641.

Kantas, N., Doucet, A., Singh, S., Maciejowski, J., Chopin, N., 2014. On particle methods for parameter estimation in state-space models. ArXiv Working Paper 1412.8659v1.

Kilian, L., 1998. Small-sample confidence intervals for impulse response functions. Rev. Econ. Stat. 80 (2), 218–230. http://dx.doi.org/10.1162/003465398557465.

Kilian, L., 1999. Finite-sample properties of percentile and percentile-t bootstrap confidence intervals for impulse responses. Rev. Econ. Stat. 81 (4), 652–660.

Kim, J.Y., 2002. Limited information likelihood and Bayesian analysis. J. Econ. 107 (1-2), 175–193. http://dx.doi.org/10.1016/S0304-4076(01)00119-1.

Kim, J., Kim, S.H., 2003. Spurious welfare reversals in international business cycle models. J. Int. Econ. 60, 471–500.

Kim, J., Kim, S.H., Schaumburg, E., Sims, C.A., 2008. Calculating and using second-order accurate solutions of discrete time dynamic equilibrium models. J. Econ. Dyn. Control 32, 3397–3414.

Kimball, M.S., 1990. Precautionary saving in the small and in the large. Econometrica 58, 53–73.

King, R.G., Watson, M.W., 1998. The solution of singular linear difference systems under rational expectations. Int. Econ. Rev. 39, 1015–1026.

King, R.G., Plosser, C.I., Rebelo, S., 1988. Production, growth, and business cycles: I the basic neoclassical model. J. Monet. Econ. 21 (2-3), 195–232.

King, R.G., Plosser, C.I., Rebelo, S.T., 2002. Production, growth and business cycles: technical appendix. Comput. Econ. 20, 87–116.

Kitamura, Y., Otsu, T., 2011. Bayesian analysis of moment condition models using nonparametric priors. Manuscript, Yale University and LSE.

Kleibergen, F., Mavroeidis, S., 2009. Weak instrument robust tests in GMM and the New Keynesian Phillips curve. J. Bus. Econ. Stat. 27 (3), 293–311. http://ideas.repec.org/a/bes/jnlbes/v27i3y2009p293-311.html.

Kleibergen, F., Mavroeidis, S., 2014. Identification issues in limited-information Bayesian analysis of structural macroeconomic models. J. Appl. Econ. 29, 1183–1209.

Klein, P., 2000. Using the generalized Schur form to solve a multivariate linear rational expectations model. J. Econ. Dyn. Control 24 (10), 1405–1423. http://dx.doi.org/10.1016/S0165-1889(99)00045-7.

Kociecki, A., Kolasa, M., 2015. Global identification of linearized DSGE models. Manuscript, Bank of Poland.

Kogan, L., Mitra, I., 2014. Accuracy verification for numerical solutions of equilibrium models. Manuscript, MIT.

Kollmann, R., 2015. Tractable latent state filtering for non-linear DSGE models using a second-order approximation and pruning. Comput. Econ. 45, 239–260.

Komunjer, I., Ng, S., 2011. Dynamic identification of DSGE models. Econometrica 79 (6), 1995–2032.

Koop, G., Pesaran, H.M., Potter, S.M., 1996. Impulse response analysis in nonlinear multivariate models. J. Econ. 74, 119–147.

Koop, G., Pesaran, H.M., Smith, R.P., 2013. On identification of Bayesian DSGE models. J. Bus. Econ. Stat. 31 (3), 300–314.

Kopecky, K.A., Suen, R.M.H., 2010. Finite state Markov-chain approximations to highly persistent processes. Rev. Econ. Dyn. 13, 701–714.

Kormilitsina, A., Nekipelov, D., 2012. Approximation properties of Laplace-type estimators. Adv. Econ. 28, 291–318.

Kormilitsina, A., Nekipelov, D., 2016. Consistent variance of the Laplace type estimators: application to DSGE models. Int. Econ. Rev. 57 (2), 603–622.

Krüger, D., Kubler, F., 2004. Computing equilibrium in OLG models with stochastic production. J. Econ. Dyn. Control 28, 1411–1436.

Krusell, P., Smith, A.A., 1998. Income and wealth heterogeneity in the macroeconomy. J. Polit. Econ. 106, 867–896.

Kwan, Y.K., 1999. Asymptotic Bayesian analysis based on a limited information estimator. J. Econ. 88, 99–121.

Kydland, F.E., Prescott, E.C., 1982. Time to build and aggregate fluctuations. Econometrica 50 (6), 1345–1370.

Lancaster, T., 2004. An Introduction to Modern Bayesian Econometrics. Blackwell Publishing.

Lanczos, C., 1938. Trigonometric interpolation of empirical and analytical functions. J. Math. Phys. 17, 123–199.

Lazar, N.A., 2003. Bayesian empirical likelihood. Biometrika 90 (2), 319–326.

Lee, B.S., Ingram, B.F., 1991. Simulation estimation of time-series models. J. Econ. 47 (2-3), 197–205. http://ideas.repec.org/a/eee/econom/v47y1991i2-3p197-205.html.

Leeper, E.M., 1991. Equilibria under 'active' and 'passive' monetary and fiscal policies. J. Monet. Econ. 27, 129–147.

Leeper, E.M., Sims, C.A., 1995. Toward a modern macroeconomic model usable for policy analysis. In: Fischer, S., Rotemberg, J.J. (Eds.), NBER Macroeconomics Annual 1994. MIT Press, Cambridge, pp. 81–118.

Leland, H.E., 1968. Saving and uncertainty: the precautionary demand for saving. Q. J. Econ. 82, 465–473.

Levin, A., Onatski, A., Williams, J.C., Williams, N., 2006. Monetary policy under uncertainty in micro-founded macroeconometric models. In: Gertler, M., Rogoff, K. (Eds.), NBER Macroeconomics Annual 2005, vol. 20. MIT Press, Cambridge, pp. 229–287. http://www.columbia.edu/%7Eao2027/LOWW.pdf.

Levintal, O., 2015a. Fifth-order perturbation solution to DSGE models. Manuscript, Interdisciplinary Center Herzliya.

Levintal, O., 2015b. Taylor projection: a new solution method to dynamic general equilibrium models. Manuscript, Interdisciplinary Center Herzliya.

Liu, J.S., 2001. Monte Carlo Strategies in Scientific Computing. Springer Verlag.

Lubik, T., Schorfheide, F., 2003. Computing sunspot equilibria in linear rational expectations models. J. Econ. Dyn. Control 28 (2), 273–285.

Lubik, T., Schorfheide, F., 2005. Do central banks respond to exchange rate movements? A structural investigation. J. Monet. Econ. 54 (4), 1069–1087.

Lubik, T., Schorfheide, F., 2006. A Bayesian look at the new open macroeconomics. NBER Macroeconomics Annual 2005.

Lubik, T.A., Schorfheide, F., 2004. Testing for indeterminancy: an application to U.S. monetary policy. Am. Econ. Rev. 94 (1), 190–217.

Lucas Jr., R.E., 1987. Models of Business Cycles. Basil Blackwell, Oxford.

Lütkepohl, H., 1990. Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. Rev. Econ. Stat. 72 (1), 116–125. http://ideas.repec.org/a/tpr/restat/v72y1990i1p116-25.html.

Maliar, L., Maliar, S., 2014. Numerical methods for large scale dynamic economic models. In: Schmedders, K., Judd, K.L. (Eds.), Handbook of Computational Economics. vol. 3. Elsevier, pp. 325–477.

Maliar, L., Maliar, S., 2015. Merging simulation and projection approaches to solve high-dimensional problems with an application to a New Keynesian model. Quant. Econ. 6, 1–47.

Maliar, L., Maliar, S., Judd, K.L., 2011. Solving the multi-country real business cycle model using Ergodic set methods. J. Econ. Dyn. Control 35, 207–228.

Maliar, L., Maliar, S., Taylor, J.B., Tsener, I., 2015. A tractable framework for analyzing a class of nonstationary Markov models. NBER Working Paper 21155.

Maliar, L., Maliar, S., Villemot, S., 2013. Taking perturbation to the accuracy frontier: a hybrid of local and global solutions. Comput. Econ. 42, 307–325.

Malik, S., Pitt, M.K., 2011. Particle filters for continuous likelihood evaluation and maximization. J. Econ. 165, 190–209.

Malin, B.A., Krüger, D., Kubler, F., 2011. Solving the multi-country real business cycle model using a Smolyak-collocation method. J. Econ. Dyn. Control 35, 229–239.

Marcet, A., Lorenzoni, G., 1999. Parameterized expectations approach: some practical issues. In: Marimon, R., Scott, A. (Eds.), Computational Methods for the Study of Dynamic Economies. Oxford University Press.

Marcet, A., Marshall, D.A., 1994. Solving nonlinear rational expectations models by parameterized expectations: convergence to stationary solutions. .

Marmer, V., Otsu, T., 2012. Optimal comparison of misspecified moment restriction models under a chosen measure of fit. J. Econ. 170 (2), 538–550.

Mason, J.C., Handscomb, D., 2003. Chebyshev Polynomials. CRC Press.

Mavroeidis, S., 2005. Identification issues in forward-looking models estimated by GMM, with an application to the Phillips curve. J. Money Credit Bank. 37 (3), 421–448. http://ideas.repec.org/a/mcb/jmoncb/v37y2005i3p421-48.html.

Mavroeidis, S., 2010. Monetary policy rules and macroeconomic stability: some new evidence. Am. Econ. Rev. 100 (1), 491–503. http://www.ingentaconnect.com/content/aea/aer/2010/00000100/00000001/art00018.

Mavroeidis, S., Plagborg-Moller, M., Stock, J.H., 2014. Empirical evidence on inflation expectations in the New Keynesian Phillips curve. J. Econ. Lit. 52 (1), 124–188. http://ideas.repec.org/a/aea/jeclit/v52y2014i1p124-88.html.

McGrattan, E.R., 1994. The macroeconomic effects of distortionary taxation. J. Monet. Econ. 33 (3), 573–601.

McGrattan, E.R., 1996. Solving the stochastic growth model with a finite element method. J. Econ. Dyn. Control 20, 19–42.

Mittnik, S., Zadrozny, P.A., 1993. Asymptotic distributions of impulse responses, step responses, and variance decompositions of estimated linear dynamic models. Econometrica 61 (4), 857–870. http://ideas.repec.org/a/ecm/emetrp/v61y1993i4p857-70.html.

Müller, U., 2012. Measuring prior sensitivity and prior informativeness in large Bayesian models. J. Monet. Econ. 59, 581–597.

Müller, U., 2013. Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. Econometrica 81 (5), 1805–1849.

Nishiyama, S., Smetters, K., 2014. Analyzing fiscal policies in a heterogeneous-agent overlapping-generations economy. In: Schmedders, K., Judd, K.L. (Eds.), Handbook of Computational Economics, vol. 3. Elsevier, pp. 117–160.

Nobile, F., Tempone, R., Webster, C.G., 2008. An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. 46, 2411–2442.

Nocedal, J., Wright, S.J., 2006. Numerical Optimization. Springer Verlag.

Otrok, C., 2001. On measuring the welfare costs of business cycles. J. Monet. Econ. 47 (1), 61–92.

Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57 (5), 1027–1057. http://EconPapers.repec.org/RePEc:ecm:emetrp:v:57:y:1989:i:5:p:1027-57.

Parra-Alvarez, J.C., 2015. Solution methods and inference in continuous-time dynamic equilibrium economies. Aarhus University.

Pesavento, E., Rossi, B., 2007. Impulse response confidence intervals for persistent data: what have we learned? J. Econ. Dyn. Control 31 (7), 2398–2412. ISSN 0165-1889. http://dx.doi.org/10.1016/j.jedc.2006.07.006.

Phillips, P.C.B., 1998. Impulse response and forecast error variance asymptotics in nonstationary vars. J. Econ. 83 (1-2), 21–56. http://ideas.repec.org/a/eee/econom/v83y1998i1-2p21-56.html.

Phillips, P.C., Solo, V., 1992. Asymptotics for linear processes. Ann. Stat. 20 (2), 971–1001.

Piazzesi, M., Schneider, M., 2006. Equilibrium yield curves. NBER Macroeconomics Annual 2006.

Pitt, M.K., Silva, R.d.S., Giordani, P., Kohn, R., 2012. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. J. Econ. 171, 134–151.

Pratt, J.W., Raiffa, H., Schlaifer, R., 1965. Introduction to Statistical Decision Theory. Wiley, New York, NY.

Preston, B., Roca, M., 2007. Incomplete markets, heterogeneity and macroeconomic dynamics. NBER Working Paper 13260.

Priestley, M.B., 1981. Spectral Analysis and Time Series. Academic Press.

Qu, Z., 2014. Inference in dynamic stochastic general equilibrium models with possible weak identification. Quant. Econ. 5, 457–494.

Qu, Z., 2015. A composite likelihood framework for analyzing singular DSGE models. Manuscript, Boston University.

Qu, Z., Tkachenko, D., 2012. Identification and frequency domain quasi-maximum likelihood estimation of linearized DSGE models. Quant. Econ. 3, 95–132.

Qu, Z., Tkachenko, D., 2014. Local and global parameter identification in DSGE models: allowing for indeterminancy. Manuscript, Boston University.

Rabanal, P., Rubio-Ramírez, J.F., 2005. Comparing New Keynesian models of the business cycle: a Bayesian approach. J. Monet. Econ. 52 (6), 1151–1166.

Ramey, V.A., 2016. Macroeconomic shocks and their propagation. In: Taylor, J.B., Uhlig, H. (Eds.), Handbook of Macroeconomics, vol. 2A. Elsevier, Amsterdam, Netherlands, pp. 71–162.

Ríos-Rull, J.V., Schorfheide, F., Fuentes-Albero, C., Kryshko, M., Santaeulalia-Llopis, R., 2012. Methods versus substance: measuring the effects of technology shocks. J. Monet. Econ. 59 (8), 826–846.

Robert, C.P., Casella, G., 2004. Monte Carlo Statistical Methods. Springer.

Rossi, B., Pesavento, E., 2006. Small-sample confidence intervals for multivariate impulse response functions at long horizons. J. Appl. Econ. 21 (8), 1135–1155. http://ideas.repec.org/a/jae/japmet/v21y2006i8p1135-1155.html.

Rotemberg, J.J., Woodford, M., 1997. An optimization-based econometric framework for the evaluation of monetary policy. In: Bernanke, B.S., Rotemberg, J.J. (Eds.), NBER Macroeconomics Annual 1997. MIT Press, Cambridge.

Rouwenhorst, K.G., 1995. Asset pricing implications of equilibrium business cycle models. In: Cooley, T.F. (Ed.), Frontiers of Business Cycle Research. Princeton University Press, pp. 294–330.

Rudebusch, G., Swanson, E., 2011. Examining the bond premium puzzle with a DSGE model. J. Monet. Econ. 55.

Rudebusch, G., Swanson, E., 2012. The bond premium in a DSGE model with long-run real and nominal risks. Am. Econ. J. Macroecon. 4, 105–143.

Rudin, W., 1976. Principles of Mathematical Analysis. McGraw and Hill, New York, NY.

Ruge-Murcia, F., 2012. Estimating nonlinear DSGE models by the simulated method of moments: with an application to business cycles. J. Econ. Dyn. Control 36, 914938.

Ruge-Murcia, F., 2014. Indirect inference estimation of nonlinear dynamic general equilibrium models: with an application to asset pricing under skewness risk. McGill University, Working Paper.

Ruge-Murcia, F.J., 2007. Methods to estimate dynamic stochastic general equilibrium models. J. Econ. Dyn. Control 31 (8), 2599–2636. http://ideas.repec.org/a/eee/dyncon/v31y2007i8p2599-2636.html.

Rust, J., 1996. Numerical dynamic programming in economics. In: Amman, H.M., Kendrick, D.A., Rust, J. (Eds.), Handbook of Computational Economics, vol. 1. Elsevier, pp. 619–729.

Sala, L., 2015. DSGE models in the frequency domain. J. Appl. Econ. 30, 219–240. http://ideas.repec.org/p/igi/igierp/504.html.

Samuelson, P.A., 1970. The fundamental approximation theorem of portfolio analysis in terms of means, variances and higher moments. Rev. Econ. Stud. 37, 537–542.

Sandmo, A., 1970. The effect of uncertainty on saving decisions. Rev. Econ. Stud. 37, 353–360.

Santos, M.S., 1992. Differentiability and comparative analysis in discrete-time infinite-horizon optimization. J. Econ. Theory 57, 222–229.

Santos, M.S., 1993. On high-order differentiability of the policy function. Econ. Theory 2, 565–570.

Santos, M.S., 2000. Accuracy of numerical solutions using the Euler equation residuals. Econometrica 68, 1337–1402.

Santos, M.S., Peralta-Alva, A., 2005. Accuracy of simulations for stochastic dynamic models. Econometrica 73 (6), 1939–1976.http://EconPapers.repec.org/RePEc:ecm:emetrp:v:73:y:2005:i:6:p:1939-1976.

Santos, M.S., Peralta-Alva, A., 2014. Analysis of numerical errors. In: Schmedders, K., Judd, K.L. (Eds.), Handbook of Computational Economics, vol. 3. Elsevier, pp. 517–556.

Santos, M.S., Rust, J., 2004. Convergence properties of policy iteration. SIAM J. Control Optim. 42, 2094–2115.

Santos, M.S., Vigo-Aguiar, J., 1998. Analysis of a numerical dynamic programming algorithm applied to economic models. Econometrica 66, 409–426.

Scalone, V., 2015. Estimating non-linear DSGEs with approximate Bayesian computations. Manuscript, University of Rome La Sapienza.

Schennach, S.M., 2005. Bayesian exponential tilted empirical likelihood. Biometrika 92, 31–46.

Schmitt-Grohé, S., Uribe, M., 2004. Solving dynamic general equilibrium models using a second-order approximation to the policy function. J. Econ. Dyn. Control 28, 755–775.

Schorfheide, F., 2000. Loss function-based evaluation of DSGE models. J. Appl. Econ. 15, 645–670.

Schorfheide, F., 2005. Learning and monetary policy shifts. Rev. Econ. Dyn. 8 (2), 392–419.

Schorfheide, F., 2005. VAR forecasting under misspecification. J. Econ. 128 (1), 99–136.

Schorfheide, F., 2013. Estimation and evaluation of DSGE models: progress and challenges. In: Acemoglu, D., Arellano, M., Dekel, E. (Eds.), Advances in Economics and Econometrics: Tenth World Congress, vol. III. Cambridge University Press, pp. 184–230.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.

Shin, M., 2014. Bayesian GMM. PhD Thesis, University of Pennsylvania.

Sikorski, K., 1985. Optimal solution of nonlinear equations. J. Complex. 1, 197–209.

Simmonds, J.G., Mann, J.E.J., 1997. A First Look at Perturbation Theory. Dover.

Simon, H.A., 1956. Dynamic programming under uncertainty with a quadratic criterion function. Econometrica 24, 74–81.

Sims, C.A., 2002. Solving linear rational expectations models. Comput. Econ. 20, 1–20.

Sims, C.A., Waggoner, D., Zha, T., 2008. Methods for inference in large multiple-equation Markov-switching models. J. Econ. 146 (2), 255–274. http://www.frbatlanta.org/filelegacydocs/wp0622.pdf.

Smets, F., Wouters, R., 2003. An estimated dynamic stochastic general equilibrium model of the euro area. J. Eur. Econ. Assoc. 1 (5), 1123–1175.

Smets, F., Wouters, R., 2007. Shocks and frictions in US business cycles: a Bayesian DSGE approach. Am. Econ. Rev. 97, 586–608.

Smith Jr., A., 1993. Estimating nonlinear time-series models using simulated vector autoregressions. J. Appl. Econ. 8, S63–S84. http://ideas.repec.org/a/jae/japmet/v8y1993isps63-84.html.

Smolyak, S.A., 1963. Quadrature and interpolation formulas for tensor products of certain classes of functions. Sov. Math. 4, 240–243.

Solín, P., Segeth, K., Doležel, I., 2004. Higher-Order Finite Elements Method. Chapman & Hall/CRC.

Stachurski, J., Martin, V., 2008. Computing the distributions of economic models via simulation. Econometrica 76, 443–450.

Stock, J.H., Watson, M.W., 2001. Vector autoregressions. J. Econ. Perspect. 15 (4), 101–115. http://ideas.repec.org/a/ijc/ijcjou/y2007q4a4.html.

Stokey, N.L., Lucas Jr., R.E., Prescott, E.C., 1989. Recursive Methods in Economic Dynamics. Harvard University Press.

Swanson, E.T., Anderson, G.S., Levin, A.T., 2006. Higher-order perturbation solutions to dynamic, discrete-time rational expectations models. Federal Reserve Bank of San Francisco, Working Paper Series, 2006-01.

Tallarini, T.D.J., 2000. Risk-sensitive real business cycles. J. Monet. Econ. 45, 507–532.

Tauchen, G., 1986. Finite state Markov-chain approximations to univariate and vector autoregressions. Econ. Lett. 20, 177–181.

Theil, H., 1957. A note on certainty equivalence in dynamic planning. Econometrica 25, 346–349.

Thompson, J.F., Warsi, Z., Mastin, C.W., 1985. Numerical Grid Generation: Foundations and Applications. North-Holland, New York, NY.

Uhlig, H., 1999. A toolkit for analysing nonlinear dynamic stochastic models easily. In: Marimon, R., Scott, A. (Eds.), Computational Methods for the Study of Dynamic Economies. Oxford University Press, pp. 30–61.

Uhlig, H., 2005. What are the effects of monetary policy on output? Results from an agnostic identification procedure. J. Monet. Econ. 52 (2), 381–419.

van Binsbergen, J.H., Fernández-Villaverde, J., Koijen, R.S., Rubio-Ramírez, J.F., 2012. The term structure of interest rates in a DSGE model with recursive preferences. J. Monet. Econ. 59, 634–648.

Waggoner, D., Zha, T., 2012. Confronting model misspecification in macroeconomics. J. Econ. 171 (2), 167184.

White, H., 1994. Estimation, Inference, and Specification Analysis. Cambridge University Press.

Woodford, M., 2003. Optimal interest-rate smoothing. Rev. Econ. Stud. 70, 861–886.

CHAPTER 10

Recursive Contracts and Endogenously Incomplete Markets

M. Golosov*, **A. Tsyvinski**†, **N. Werquin**‡
*Princeton University, Princeton, NJ, United States
†Yale University, New Haven, CT, United States
‡Toulouse School of Economics, Toulouse, France

Contents

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.007

Abstract

In this chapter we study dynamic incentive models in which risk sharing is endogenously limited by the presence of informational or enforcement frictions. We comprehensively overview one of the most important tools for the analysis such problems—the theory of recursive contracts. Recursive formulations allow us to reduce often complex models to a sequence of essentially static problems that are easier to analyze both analytically and computationally. We first provide a self-contained treatment of the basic theory: the Revelation Principle, formulating and simplifying the incentive constraints, using promised utilities as state variables, and analyzing models with persistent shocks using the first-order approach. We then discuss more advanced topics: duality theory and Lagrange multiplier techniques, models with lack of commitment, and martingale methods in continuous time. Finally, we show how a variety of applications in public economics, corporate finance, development and international economics featuring incomplete risk sharing can be analyzed using the tools of the theory of recursive contracts.

Keywords

Principal–agent model, Dynamic mechanism design, Recursive contracts, Private information, Limited commitment, Incomplete markets, Revelation Principle, Promised utility, First-order approach, Hidden storage, Lagrangian, Continuous time contracts

JEL Classification Codes

A33, C61, D52, D82, D86, H21

1. INTRODUCTION

Dynamic incentive problems are ubiquitous in macroeconomics. The design of social insurance programs by governments, long-run relationships between banks and entrepreneurs, informal insurance contracts against idiosyncratic shocks provided in village economies, sovereign borrowing and lending between countries can all be understood using the theory of dynamic incentives. These models have been widely used in macroeconomics, public economics, international macroeconomics, finance, development, and political economy, both for explaining existing patterns in the data and for normative policy analysis. The unifying feature of these models is that, at their essence, they study endogenously incomplete markets, ie, environments in which risk sharing is constrained by (informational or enforcement) frictions, and where insurance arrangements arise endogenously.

One of the most important tools used for studying dynamic incentive problems is the theory of recursive contracts. Recursive formulations allow one to reduce often complex models to a sequence of essentially static problems that are easier to analyze both analytically and computationally. This substantially simplifies the analysis and the characterization of the optimal insurance arrangements in rich and realistic environments. The goal of this chapter is to provide an overview of the theory of recursive contracts and give a number of examples of application. The analysis in the theoretical part is self-contained; whenever a textbook approach is not directly applicable (eg, when the assumptions needed to apply the recursive techniques in Stokey et al., 1989 are not met), we provide the necessary mathematical background. We also discuss the strengths and weaknesses of several alternative approaches to solving dynamic incentive problems that emerged in the literature. In the last part of the chapter we show how the methods of recursive contracts can be used in a variety of applications.

Our paper is organized as follows. Section 2 considers a prototypical dynamic incentive problem—insurance against privately observable idiosyncratic taste shocks under perfect commitment by the principal. The goal of this section is to provide an example of a self-contained, rigorous, and relatively general treatment of a dynamic incentive problem. We also use this economy in subsequent sections to illustrate other approaches to the analysis of dynamic incentive problems. In Section 2 we highlight the three main steps in the analysis: first, applying the Revelation Principle to set up a mechanism design problem with incentive constraints; second, simplifying this problem by focusing on one-shot incentive constraints; and third, writing this problem recursively using "promised utilities" as state variables. We then show how this recursive formulation can be used to characterize the properties of the optimal insurance arrangements in our economy. We derive general features of the optimal insurance contract and characterize the long-run behavior of the economy in Section 2.4. We show how to overcome the technical difficulties that arise when the idiosyncratic shocks are persistent in Section 2.5. Next we discuss in a simple version of the framework how the optimal insurance arrangement is affected when the agent can unobservably save in Section 2.6. We conclude by showing how the same techniques can be applied to other dynamic incentive problems, such as moral hazard in Section 2.7.

Section 3 considers more advanced topics. We focus on three of them: using Lagrange multiplier tools in recursive formulations, studying dynamic insurance problems in economies in which the principal has imperfect commitment, and applying martingale techniques to study recursive contracts in continuous time. Section 3.1 discusses the Lagrangian techniques. Using Lagrangians together with the recursive methods of Section 2 greatly expands the class of problems that can be characterized. We first provide an overview of the theory of constrained optimization using Lagrange multipliers, with a particular focus on showing how to use them in the infinite dimensional settings that frequently arise in macroeconomic applications. We then show how to apply these theoretical techniques to incentive problems to obtain several alternative recursive formulations having some advantages relative to those discussed in Section 2. A number of results in this section are new to the dynamic contracts literature. In Section 3.2 we show how to analyze dynamic insurance problems in settings where the principal cannot commit to the contracts. The arguments used to prove simple versions of the Revelation Principle under commitment fail in such an environment; we discuss several ways to generalize it and write a recursive formulation of the mechanism design problem. Our characterization of such problems relies heavily on our analysis of Section 3.1. Finally, in Section 3.3, we show how to analyze a dynamic contracting problem in continuous time using martingale methods and the dynamic programming principle. To keep the analysis self-contained, we start by stating the stochastic calculus results that we use. Continuous-time methods often simplify the characterization of optimal contracts, allowing for analytical comparative statics and easier numerical analysis of the solution.

Section 4 gives a number of applications of the recursive techniques discussed in Sections 2 and 3 to various environments. We show that these diverse applications share three key features: (i) insurance is endogenously limited by the presence of a friction; (ii) the problem is dynamic; and (iii) the recursive contract techniques that we develop in the theoretical sections allow us to derive deep characterizations of these problems. We explain how theoretical constructs such as the incentive constraints and promised utilities can be mapped into concrete economic concepts, and how the predictions of dynamic incentive models can be tested empirically and used for policy analysis. In Section 4.1, we apply the techniques and results of Section 2 to public finance where the endogenous market incompleteness and the limited social insurance arise due to the unobservability of the shocks that agents receive. We derive several central results characterizing the optimal social insurance mechanisms and show how to implement the optimal allocations with a tax and transfer system that arises endogenously, without restricting the system exogenously to a specific functional form. In Section 4.2 we show how recursive techniques can be applied to corporate finance problems to study the effects of informational frictions on firm dynamics and the optimal capital structure. Section 4.3 presents applications of these techniques to study insurance in village economies in developing countries where contracts are limited by enforcement and informational frictions. Section 4.4 discusses applications to international borrowing and lending.

2. A SIMPLE MODEL OF DYNAMIC INSURANCE

In this section we study a prototypical model of dynamic insurance against privately observed idiosyncratic shocks. Our goal is to explain the key steps in the analysis and the main insights in the simplest setting. The mathematical techniques that we use as well as the economic insights that we obtain extend to many richer and more realistic environments. We discuss examples of such environments in the following sections.

2.1 Environment

We consider a discrete-time economy that lasts T periods, where T may be finite or infinite. The economy is populated by a continuum of ex ante identical agents whose preferences over period-t consumption $c_t \geq 0$ are given by $\theta_t U(c_t)$, where $\theta_t \in \Theta \subset \mathbb{R}_+$ is an idiosyncratic "taste shock" that the individual receives in period t, and U is a utility function.

Assumption 1 The utility function $U : \mathbb{R}_+ \to \mathbb{R}$ is an increasing, strictly concave, differentiable function that satisfies the Inada conditions $\lim_{c \to 0} U'(c) = \infty$ and $\lim_{c \to \infty} U'(c) = 0$.

All agents have the same discount factor $\beta \in (0,1)$. In each period the economy receives e units of endowment which can be freely transferred between periods at rate β.

The idiosyncratic taste shocks are stochastic. We use the notations $\theta^t = (\theta_1, \ldots, \theta_t) \in \Theta^t$ to denote a history of realizations of shocks up to period t and $\pi_t(\theta^t)$ to denote the probability of realization of history θ^t. We assume that the law of large numbers holds so that $\pi_t(\theta^t)$ is also the measure of individuals who experienced history θ^t.[a] An individual privately learns his taste shock θ_t at the beginning of period t. Thus, at the beginning of period t an agent knows his history θ^t of current and past shocks, but not his future shocks. This implies that his choices in period t, and more generally all the period-t random variables x_t that we encounter, can only be a function of this history.

Some parts of our analysis use results from probability theory and require us to be more formal about the probability spaces that we use. A standard way to formalize these stochastic processes is as follows.[b] Let Θ^T be the space of all histories θ^T and let π_T be a probability measure over the Borel subsets $\mathcal{B}(\Theta^T)$ of Θ^T. Thus, $(\Theta^T, \mathcal{B}(\Theta^T), \pi_T)$ forms a probability space. Any period-t random variable is required to be measurable with respect to $\mathcal{B}(\Theta^t)$, that is, for any Borel subset M of $\mathbb{R}$, $x_t^{-1}(M) = B \times \Theta^{T-t}$, where B is a Borel subset of Θ^t. This formalizes the intuition that the realization of shocks in future periods is not known as of period t.

Until Section 2.5.2 we make the following assumptions about the idiosyncratic taste shocks:

Assumption 2 The set $\Theta \subset \mathbb{R}_+$ of taste shocks is discrete and finite with cardinality $|\Theta|$. Agents' shocks evolve according to a first-order Markov process, that is, the probability of drawing type θ_t in period t depends only on the period-$(t-1)$ type:

$$\pi_t\left(\theta_t \middle| \theta^{t-1}\right) = \pi(\theta_t | \theta_{t-1}), \forall \theta^{t-1} \in \Theta^{t-1}, \theta_t \in \Theta,$$

where θ_{t-1} is the last component of θ^{t-1}.

We use the notation $\pi_t(\theta^t | \theta^s)$ for $t > s$ to denote the probability of realization of history θ^t up to period t conditional on the realization of history θ^s up to period s, with

[a] The assumption that the law of large numbers holds can be justified formally (see Uhlig, 1996; Sun, 2006).

[b] See Stokey et al. (1989, Chapter 7) for a review of the measure-theoretic apparatus.

a convention that $\pi_t(\theta^t|\theta^s)=0$ if the first s elements of θ^t are not θ^s (history θ^t in period t cannot occur if θ^s was not realized up to period s). We use θ_s^t to denote $(\theta_s,\ldots,\theta_t)$. Finally we index the elements of Θ by the subscript (j) for $j\in\{1,\ldots,|\Theta|\}$, and assume that $\theta_{(1)}<\theta_{(2)}<\ldots<\theta_{(|\Theta|)}$.

We consider the problem of a social planner that chooses consumption allocations[c] $c_t:\Theta^t\to\mathbb{R}_+$ to maximize agents' ex ante expected utility and has the ability to commit to such allocations in period 0. At this stage we are agnostic about who or what this planner is. One can think of it as a government that provides insurance to agents, or as some decentralized market arrangement. We study the optimal insurance contract that such a planner can provide given a feasibility constraint and informational constraints. We use the shortcut $\mathbf{c}$ to denote the consumption plan $\{c_t(\theta^t)\}_{t\geq 1,\theta^t\in\Theta^t}$.

The ex ante, "period-0" expected utility of all agents is denoted by $U_0(\mathbf{c})$ and is given by

$$U_0(\mathbf{c})\equiv\mathbb{E}_0\left[\sum_{t=1}^{T}\beta^{t-1}\theta_t U(c_t)\right]=\sum_{t=1}^{T}\sum_{\theta^t\in\Theta^t}\beta^{t-1}\pi_t(\theta^t)\,\theta_t U\left(c_t(\theta^t)\right). \tag{1}$$

Here $\mathbb{E}_0$ represents the (unconditional) expectation at time 0, before the first-period type θ_1 is known. Under our assumption that resources can be freely transferred between periods, the resource constraint is

$$\mathbb{E}_0\left[\sum_{t=1}^{T}\beta^{t-1}c_t\right]\leq\frac{1-\beta^T}{1-\beta}e. \tag{2}$$

Note that to write the left hand side of this feasibility constraint we again implicitly invoked the law of large numbers.

When the realizations of the taste shocks are observable by the planner, this problem can easily be solved explicitly. Let $\zeta>0$ be the Lagrange multiplier on the feasibility constraint. The optimal allocation $\mathbf{c}^{\text{fb}}$ in the case where shocks are observable (the "first best" allocation) is a solution to

$$\theta_t U'\left(c_t^{\text{fb}}(\theta^t)\right)=\zeta,\ \forall t\geq 1,\ \forall\theta^t\in\Theta^t. \tag{3}$$

It is immediate to see that this equation implies that $c_t^{\text{fb}}(\theta^t)$ is independent of period t or the past history of shocks θ^{t-1}, and only depends on the current realization of the shock θ_t. That is, the informationally unconstrained optimal insurance in this economy gives to agents with a higher realization of the shock θ in any period (hence with a higher current marginal utility) more consumption than to agents with a lower realization of a taste shock.

[c] Formally, c_t is a random variable over the probability space $\left(\Theta^T,\mathcal{B}\left(\Theta^T\right),\pi_T\right)$ that is measurable with respect to $\mathcal{B}(\Theta^t)$.

2.2 The Revelation Principle and Incentive Compatibility

We are interested in understanding the properties of the best insurance arrangements that a planner can provide in the economy with private information. This insurance can be provided by many different mechanisms: the agents may be required to live in autarky and consume their endowment, or may be allowed to trade assets, or may be provided with more sophisticated arrangements by the planner. A priori it is not obvious how to set up the problem of finding the best mechanism to provide the highest utility to agents. This problem simplifies once we apply the results of the mechanism design literature, in particular the Revelation Principle. Textbook treatments of the Revelation Principle are widely available (see, eg, Chapter 23 in Mas-Colell et al., 1995). Here we outline the main arguments behind the Revelation Principle in our context. This overview is useful both to keep the analysis self-contained and to emphasize subtleties that emerge in using the Revelation Principle once additional frictions, such as lack of commitment by the planner, are introduced.

Hurwicz (1960, 1972) provided a general framework to study various arrangements of allocation provision in environments with private information. He showed that such arrangements can be represented as abstract communication mechanisms. Consider an arbitrary message space M that consists of a collection of messages m. Each agent observes his shock θ_t and sends a (possibly random) message $m_t \in M$ to the principal. The agent's reporting strategy in period t is a map $\tilde{\sigma}_t : \Theta^t \to \Delta(M)$. The planner in turn chooses a (possibly stochastic) allocation rule $\tilde{c}_t : M^t \to \Delta(\mathbb{R}_+)$, where $\Delta(\mathbb{R}_+)$ denotes the space of probability measures on $\mathbb{R}_+$. The strategies $\tilde{\boldsymbol{\sigma}} = \left\{\tilde{\sigma}_t(\theta^t)\right\}_{t\geq 1, \theta^t \in \Theta^t}$ and $\tilde{\boldsymbol{c}} = \left\{\tilde{c}_t(m^t)\right\}_{t\geq 1, m^t \in M^t}$ induce a measure over the consumption paths $\{c_t\}_{t\geq 1} \in \mathbb{R}_+^T$, which we denote by $\tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{\sigma}}$. The expected utility of each agent is then equal to $\mathbb{E}^{\tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{\sigma}}}\left[\sum_{t=1}^{T} \beta^{t-1} U(c_t)\right]$, where the superscript in $\mathbb{E}^{\tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{\sigma}}}$ means that the expectation is computed using the probability distribution $\tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{\sigma}}$ over the paths $\{c_t\}_{t\geq 1}$. The strategy $\tilde{\boldsymbol{\sigma}}$ is *incentive compatible* for the agent if[d,e]

$$\mathbb{E}^{\tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{\sigma}}}\left[\sum_{t=1}^{T} \beta^{t-1} \theta_t U(c_t)\right] - \mathbb{E}^{\tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{\sigma}}'}\left[\sum_{t=1}^{T} \beta^{t-1} \theta_t U(c_t)\right] \geq 0, \forall \tilde{\boldsymbol{\sigma}}'. \tag{4}$$

A mechanism $\tilde{\Gamma} = \left(M, \tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{\sigma}}\right)$ is incentive compatible if it satisfies (4), and *feasible* if it satisfies

[d] When T is allowed to be infinite, these sums may not be well defined for all $\boldsymbol{\sigma}$, and we require (4) to hold as $\limsup_{T\to\infty}$.

[e] Note that the constraints (4) also include all the constraints that ensure that $\tilde{\boldsymbol{\sigma}}$ is optimal after any history t, θ^t, ie, $\mathbb{E}^{\tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{\sigma}}}\left[\sum_{s=t}^{T} \beta^{s-t} \theta_s U(c_s) | \theta^t\right] - \mathbb{E}^{\tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{\sigma}}'}\left[\sum_{s=t}^{T} \beta^{s-t} \theta_s U(c_s) | \theta^t\right] \geq 0, \forall \tilde{\boldsymbol{\sigma}}'$.

$$\mathbb{E}^{\tilde{\boldsymbol{c}}\circ\tilde{\boldsymbol{\sigma}}}\left[\sum_{t=1}^{T}\beta^{t-1}c_t\right]\leq\frac{1-\beta^T}{1-\beta}e. \tag{5}$$

The key insight behind the Revelation Principle is that any outcome $\tilde{\boldsymbol{c}}\circ\tilde{\boldsymbol{\sigma}}$ of an incentive-compatible and feasible mechanism can be achieved as the outcome of a *direct-truthtelling* mechanism, in which agents report their types directly to the principal. Define a direct mechanism as a reporting strategy $\sigma_t:\Theta^t\to\Theta$. Define a truthtelling strategy $\boldsymbol{\sigma}^{truth}$ as $\sigma_t^{truth}\left(\theta^{t-1},\theta\right)=\theta$ for all θ^{t-1}, θ. The key observation is that there exists $\boldsymbol{c}=\{c_t\}_{t\geq1}$, with $c_t:\Theta^t\to\Delta(\mathbb{R}_+)$ for each t, such that the (induced) measure $\boldsymbol{c}\circ\boldsymbol{\sigma}^{truth}$ replicates the measure $\tilde{\boldsymbol{c}}\circ\tilde{\boldsymbol{\sigma}}$.[f]

Theorem 1 (Revelation Principle) *The outcome of any incentive-compatible and feasible mechanism* $\tilde{\Gamma}=\left(M,\tilde{\boldsymbol{c}}\circ\tilde{\boldsymbol{\sigma}}\right)$ *is also the outcome of an incentive-compatible and feasible direct truthful mechanism* $\Gamma=\left(\Theta,\boldsymbol{c}\circ\boldsymbol{\sigma}^{truth}\right)$.

Proof By construction, we have

$$\mathbb{E}^{\boldsymbol{c}\circ\boldsymbol{\sigma}^{truth}}\left[\sum_{t=1}^{T}\beta^{t-1}c_t\right]=\mathbb{E}^{\tilde{\boldsymbol{c}}\circ\tilde{\boldsymbol{\sigma}}}\left[\sum_{t=1}^{T}\beta^{t-1}c_t\right],$$

so that the truthtelling strategy satisfies (5). Any alternative strategy $\boldsymbol{\sigma}'$ induces a measure $\boldsymbol{c}\circ\boldsymbol{\sigma}'$ which replicates the measure $\tilde{\boldsymbol{c}}\circ\tilde{\boldsymbol{\sigma}}'$ for some strategy $\tilde{\boldsymbol{\sigma}}'$ in the original mechanism. Therefore

$$\mathbb{E}^{\boldsymbol{c}\circ\boldsymbol{\sigma}^{truth}}\left[\sum_{t=1}^{T}\beta^{t-1}\theta_tU(c_t)\right]-\mathbb{E}^{\boldsymbol{c}\circ\boldsymbol{\sigma}'}\left[\sum_{t=1}^{T}\beta^{t-1}\theta_tU(c_t)\right]\geq0,\forall\boldsymbol{\sigma}'. \tag{6}$$

This concludes the proof. □

We can simplify our analysis further by showing that there is no loss of generality in focusing on *deterministic* direct mechanisms, where each history of reports yields a deterministic consumption allocation (rather than a measure) $\boldsymbol{c}_t^{\det}:\Theta^t\to\mathbb{R}_+$. We show:

[f] The proof of this observation is straightforward. For simplicity, assume that $\tilde{\boldsymbol{\sigma}}$ and $\tilde{\boldsymbol{c}}$ involve randomization over a finite number of elements after each history and let $\tilde{\sigma}^t(m^t|\theta^t)$ be the probability that agent θ^t sends a history of messages m^t, and $\tilde{c}_t(x|m^t)$ be the probability that the principal delivers consumption x to an agent with a reported history m^t. Then c_t is simply defined by $c_t(x|\theta^t)\equiv\sum_{m^t}\tilde{c}_t(x|m^t)\tilde{\sigma}^t(m^t|\theta^t)$. Given this definition of $\boldsymbol{c}$ the payoff of any strategy $\tilde{\boldsymbol{\sigma}}$ in the original mechanism $\left(M,\tilde{\boldsymbol{c}}\circ\tilde{\boldsymbol{\sigma}}\right)$ can be replicated by a strategy $\boldsymbol{\sigma}$ in the truthtelling mechanism.

Proposition 1 *For any incentive-compatible and feasible direct mechanism* $\Gamma = \left(\Theta, \boldsymbol{c} \circ \boldsymbol{\sigma}^{truth}\right)$ *there exists an incentive-compatible, feasible, deterministic direct mechanism* $\left(\Theta, \boldsymbol{c}^{det} \circ \boldsymbol{\sigma}^{truth}\right)$ *that achieves the same ex ante utility.*

Proof Consider any incentive-compatible and feasible, but possibly stochastic, direct mechanism $\Gamma = \left(\Theta, \boldsymbol{c} \circ \boldsymbol{\sigma}^{truth}\right)$. Define a deterministic consumption allocation $\boldsymbol{c}_t^{\det} : \Theta^t \to \mathbb{R}_+$ implicitly by

$$U\left(\boldsymbol{c}_t^{\det}(\theta^t)\right) = \mathbb{E}^{\boldsymbol{c} \circ \boldsymbol{\sigma}^{truth}}\left[U(c_t)|\theta^t\right], \quad \forall t \geq 1, \theta^t \in \Theta^t, \tag{7}$$

where the right hand side is the expected consumption given at time t under the mechanism Γ to the agent who reports the history θ^t. Since U is concave by Assumption 1, Jensen's inequality implies that

$$\mathbb{E}^{\boldsymbol{c} \circ \boldsymbol{\sigma}^{truth}}\left[c_t|\theta^t\right] \geq \boldsymbol{c}_t^{\det}(\theta^t), \forall \theta^t,$$

hence the mechanism $\left(\Theta, \boldsymbol{c}^{\det} \circ \boldsymbol{\sigma}^{truth}\right)$ is feasible. By construction, we have that for all t, θ^t, $\mathbb{E}^{\boldsymbol{c} \circ \boldsymbol{\sigma}^{truth}}\left[U(c_t)|\theta^t\right] = \mathbb{E}^{\boldsymbol{c}^{\det} \circ \boldsymbol{\sigma}^{truth}}\left[U(c_t)|\theta^t\right]$, since the conditional expectation in (7) implies that for any report the agent receives the same expected utility under $\boldsymbol{c}$ and under $\boldsymbol{c}^{\det}$. Hence the mechanism is incentive compatible. This concludes the proof. □

With a slight abuse of notation we will use $\mathbf{c} = \{c_t(\theta^t)\}_{t \geq 1, \theta^t \in \Theta^t}$ instead of $\boldsymbol{c}_t^{\det}$. The incentive constraint in the deterministic direct mechanism can be written simply as

$$\sum_{t=1}^{T} \sum_{\theta^t \in \Theta^t} \beta^{t-1} \pi_t(\theta^t) \theta_t \left[U(c_t(\theta^t)) - U(c_t(\sigma'^t(\theta^t)))\right] \geq 0, \quad \forall \boldsymbol{\sigma}'. \tag{8}$$

The proof of the Revelation Principle requires very few assumptions except the ability of the social planner to commit to the long-term contract in period 0. Theorem 1 and Proposition 1 are very powerful results that provide a simple way to find informationally constrained optimal allocations. In particular, such allocations are a solution to the problem

$$\begin{gathered} V(e) \equiv \sup_{\mathbf{c}} \sum_{t=1}^{T} \sum_{\theta^t \in \Theta^t} \beta^{t-1} \pi_t(\theta^t) \theta_t U(c_t(\theta^t)) \\ \text{subject to } (2),\ (8). \end{gathered} \tag{9}$$

If the supremum of this problem is attained by some vector $\mathbf{c}^*$, any insurance arrangement in which agents consume $\mathbf{c}^*$ in equilibrium is efficient.

In Sections 2.3–2.5 we focus on describing general methods to solve the maximization problem defined in (9). We give examples of specific insurance arrangements when discussing various applications in Section 4.

2.3 Recursive Formulation with i.i.d. Shocks

The analysis of the solution to problem (9) is significantly simplified if shocks are independently and identically distributed (i.i.d.). In more general Markov settings, many of the same arguments continue to hold but they are more cumbersome, and analytical results are more difficult to obtain. For this reason we first focus on i.i.d. shocks and discuss general Markov shocks in Section 2.5.

Assumption 3 Types $\{\theta_t\}_{t\geq 1}$ are independent and identically distributed, that is, $\pi_t\left(\theta_t|\theta^{t-1}\right)=\pi(\theta_t)$. Without loss of generality we assume that $\mathbb{E}[\theta]=\sum_{\theta\in\Theta}\pi(\theta)\theta=1$.

2.3.1 Main Ideas in a Finite-Period Economy

In an economy with a finite number of periods, the maximization problem (9) is defined over a closed and bounded set, because the feasibility constraint imposes that for all t,θ^t, we have $0\leq c_t(\theta^t)\leq \beta\dfrac{1-\beta^T}{1-\beta}(\beta\min_{\theta\in\Theta}\pi(\theta))^{-T}e$. In finite dimensions closed and bounded sets are compact and therefore by Weierstrass' theorem the maximum of problem (9) is achieved, so that we can replace the "sup" with a "max." Moreover, it is easy to see that at the optimum the feasibility constraint must hold with equality.

We want to simplify the set of the incentive constraints in problem (9). Eq. (8) should hold for all possible reporting strategies $\boldsymbol{\sigma}'$. The set of such strategies is large; it consists of all strategies in which an agent misreports his type in some (possibly all) states only in period 1, all strategies in which he misreports his types in some states in periods 1 and 2, and so on. Most of these constraints are redundant. We say that $\boldsymbol{\sigma}''$ is a one-shot deviation strategy if $\sigma''_t\left(\theta^{t-1},\theta_t\right)\neq\theta_t$ for only one θ^t. It turns out that if (8) is satisfied for one-shot deviations, it is satisfied for all deviations in a finite period economy. Formally, we can write a *one-shot* incentive constraint (see Green, 1987) as: for all $\theta^{t-1},\theta,\hat{\theta}$,

$$\begin{aligned}
&\theta U\left(c_t\left(\theta^{t-1},\theta\right)\right)+\beta\sum_{s=1}^{T-t}\sum_{\theta^{t+s}\in\Theta^{t+s}}\beta^{s-1}\pi_{t+s}\left(\theta^{t+s}|\theta^{t-1},\theta\right)\theta_{t+s}U\left(c_{t+s}\left(\theta^{t-1},\theta,\theta_{t+1}^{t+s}\right)\right)\\
\geq\ &\theta U\left(c_t\left(\theta^{t-1},\hat{\theta}\right)\right)+\beta\sum_{s=1}^{T-t}\sum_{\theta^{t+s}\in\Theta^{t+s}}\beta^{s-1}\pi_{t+s}\left(\theta^{t+s}|\theta^{t-1},\theta\right)\theta_{t+s}U\left(c_{t+s}\left(\theta^{t-1},\hat{\theta},\theta_{t+1}^{t+s}\right)\right).
\end{aligned}\tag{10}$$

Proposition 2 *Suppose that T is finite and Assumption 3 is satisfied. An allocation* **c** *satisfies (8) if and only if it satisfies (10).*

Proof That (8) implies (10) is clear, since (10) considers a strict subset of the possible deviations. To show the converse, consider any reporting strategy $\boldsymbol{\sigma}'$. Suppose that the last period in which the agent misreports his type is period t. By (10), for any θ_t the agent gets higher utility from reporting his type truthfully in that period than from deviating. Therefore, the strategy $\boldsymbol{\sigma}''$ which coincides with $\boldsymbol{\sigma}'$ in the first $t-1$ periods and reveals types truthfully from period t onward gives higher utility to the agent than $\boldsymbol{\sigma}'$. Backward induction then implies that truthtelling gives higher utility than $\boldsymbol{\sigma}'$, establishing the result. □

Proposition 2 simplifies the maximization problem (9) by replacing the constraint set (8) with a smaller number of constraints (10). This simplified problem is still too complicated to be solved directly. We next show how to rewrite this problem recursively to reduce it to a sequence of essentially static problems which can be easily analyzed analytically and computationally.

We take several intermediate steps to rewrite constraints (2) and (10). First, observe that the constraint set defined by Eqs. (2) and (10) is not convex. Although much of the analysis can be done for a nonconvex maximization problem, we can obtain convexity by a simple change of variables: instead of choosing consumption $c_t(\theta^t)$ we can choose utils $u_t(\theta^t) \equiv U(c_t(\theta^t))$. The resource cost of providing u units of utils is $C(u) = U^{-1}(u)$, where the cost function C, defined on the range of U, is increasing, differentiable, and strictly convex by Assumption 1. Let $\underline{u}$ and $\bar{u}$ be the (possibly infinite) greatest lower bound and smallest upper bound of U. Observe that $\lim_{u \to \underline{u}} C(u) = 0$ and $\lim_{u \to \bar{u}} C(u) = \infty$. We use $\mathbb{U}$ to denote the domain of C, which is $(\underline{u}, \bar{u})$ if the utility function is unbounded below and $[\underline{u}, \bar{u})$ if it is bounded below. Given this change of variables, the incentive constraint (10) becomes linear in $\mathbf{u} = \{u_t(\theta^t)\}_{t,\theta^t}$, while the resource constraint becomes

$$\mathbb{E}_0\left[\sum_{t=1}^{T} \beta^{t-1} C(u_t)\right] \leq \frac{1-\beta^T}{1-\beta} e,$$

which defines a convex set of feasible $\mathbf{u}$.

The second simplification is to define a *continuation (or promised) utility* variable

$$v_t(\theta^t) \equiv \sum_{s=1}^{T-t} \sum_{\theta^{t+s} \in \Theta^{t+s}} \beta^{s-1} \pi_{t+s}(\theta^{t+s}|\theta^t) \theta_{t+s} u_{t+s}\left(\theta^t, \theta_{t+1}^{t+s}\right). \tag{11}$$

Using repeated substitution we get

$$v_t(\theta^t) = \sum_{\theta \in \Theta} \pi(\theta)[\theta u_{t+1}(\theta^t, \theta) + \beta v_{t+1}(\theta^t, \theta)], \forall \theta^t, \tag{12}$$

where we use the convention $v_T = 0$. Given this definition we can rewrite the incentive constraints (10) as

$$\theta u_t(\theta^{t-1},\theta) + \beta v_t(\theta^{t-1},\theta) \geq \theta u_t(\theta^{t-1},\hat{\theta}) + \beta v_t(\theta^{t-1},\hat{\theta}), \forall \theta^{t-1},\theta,\hat{\theta}. \tag{13}$$

We are now ready to simplify our analysis by observing that while the original maximization problem does not have an obvious recursive structure, its *dual* does. Our arguments imply that the maximization problem (9) can be rewritten as the maximization of the planner's objective over $(\mathbf{u},\mathbf{v}) = \left(\{u_t(\theta^t)\}_{t,\theta^t}, \{v_t(\theta^t)\}_{t,\theta^t}\right)$ subject to the constraints (2), (12), and (13). Let $(\mathbf{u}^*,\mathbf{v}^*)$ be the solution to that problem and v_0 be the value of the maximum. Then, by standard duality arguments, $(\mathbf{u}^*,\mathbf{v}^*)$ also minimizes the cost of providing $(\mathbf{u},\mathbf{v})$ subject to the incentive-compatibility constraints and the "promise-keeping constraint"

$$\mathbb{E}_0\left[\sum_{t=1}^{T}\beta^{t-1}\theta_t u_t(\theta^t)\right] = v_0. \tag{14}$$

Using the definition of $v_1(\theta^1)$, this constraint can be rewritten as

$$v_0 = \sum_{\theta\in\Theta}\pi(\theta)[\theta u_1(\theta) + \beta v_1(\theta)]. \tag{15}$$

Define the set $\Gamma(v_0)$ as

$$\Gamma(v_0) = \left\{(\mathbf{u},\mathbf{v}) : (12), (13), (15) \text{ hold}\right\}. \tag{16}$$

We thus obtain that $(\mathbf{u}^*,\mathbf{v}^*)$ is the solution to

$$K_0(v_0) \equiv \max_{(\mathbf{u},\mathbf{v})\in\Gamma(v_0)} \mathbb{E}_0\left[-\sum_{t=1}^{T}\beta^{t-1}C(u_t)\right]. \tag{17}$$

The key simplification allowed by this formulation is that it can be easily solved using recursive techniques. Let $K_{T-1}(\cdot) \equiv -C(\cdot)$, which has domain $\mathbb{V}_{T-1} = \mathbb{U}$. Define the functions K_t for $0 \leq t \leq T-2$ and their domains $\mathbb{V}_t$ recursively by

$$K_t(v) \max_{\{(u(\theta),w(\theta))\}_{\theta\in\Theta}} \sum_{\theta\in\Theta}\pi(\theta)[-C(u(\theta)) + \beta K_{t+1}(w(\theta))] \tag{18}$$

subject to the promise-keeping constraint:

$$v = \sum_{\theta\in\Theta}\pi(\theta)[\theta u(\theta) + \beta w(\theta)], \tag{19}$$

and the incentive-compatibility constraint:

$$\theta u(\theta) + \beta w(\theta) \geq \theta u(\hat{\theta}) + \beta w(\hat{\theta}), \forall \theta,\hat{\theta}, \tag{20}$$

and

$$u(\theta)\in\mathbb{U},\quad w(\theta)\in\mathbb{V}_{t+1}.$$

Eq. (18) defines the domain of K_t, denoted by $\mathbb{V}_t$. It is easy to verify that it is either $\left[\frac{1-\beta^{T-t}}{1-\beta}\underline{u},\frac{1-\beta^{T-t}}{1-\beta}\bar{u}\right)$ or $\left(\frac{1-\beta^{T-t}}{1-\beta}\underline{u},\frac{1-\beta^{T-t}}{1-\beta}\bar{u}\right)$, depending on whether the utility function is bounded below or not. It is easy to see that the function K_0 defined in (17) satisfies (18) for $t = 0$. Standard arguments establish that K_t is a continuous, strictly decreasing, strictly concave, and differentiable function. For any value $v\in\mathbb{V}_{t-1}$ for $t\geq 1$, let $\vec{u}_{v,t}=\{u_{v,t}(\theta)\}_{\theta\in\Theta}$ and $\vec{w}_{v,t}=\{w_{v,t}(\theta)\}_{\theta\in\Theta}$ denote the solution (ie, the argmax) of the Bellman equation (18). We call $\left(\vec{u}_{v,t},\vec{w}_{v,t}\right)$ the *policy functions* of the Bellman equation. Given our assumption that C is strictly convex, these policy functions are unique for each v,t.

We can now describe how to find the solution to (17). The main simplification comes from the fact that if we know the optimal value $v_t^*(\theta^t)$ after any history θ^t, we can find the optimal allocations in the nodes following θ^t without having to know the optimal allocations in any other node. We start with $t = 1$. Since $K_0(v)$ is (minus) the amount of resources required to achieve the expected utility v, the initial value v_0 must satisfy $K_0(v_0)=-\frac{1-\beta^T}{1-\beta}e$. The constrained-optimal utility allocation in period 1 for an agent with shock θ_1 is then given by $u_1^*(\theta_1)=u_{v_0,1}(\theta_1)$, and his expected utility starting from period 2 is $v_1^*(\theta_1)=w_{v_0,1}(\theta_1)$. The optimal utility allocation in period two for a history of shocks (θ_1,θ_2) is then given by $u_2^*(\theta_1,\theta_2)=u_{w_{v_0,1}(\theta_1),2}(\theta_2)$, and similarly $v_2^*(\theta_1,\theta_2)=w_{w_{v_0,1}(\theta_1),2}(\theta_2)$. This way we can use forward induction to find the solution to (9), $(\mathbf{u}^*,\mathbf{v}^*)$. We say that the solution $(\mathbf{u}^*,\mathbf{v}^*)$ is *generated by the policy functions* of the Bellman equation (18) given v_0.

2.3.2 Extension to an Infinite Period Economy

In the previous section we showed a simple way to characterize the solution to a dynamic contracting problem recursively when the number of periods is finite. For many applications it is more convenient to work with infinite periods for at least two reasons. The first is that many problems do not have a natural terminal period so that the assumption of infinite periods is more convenient. The second reason is that the assumption of infinite periods allows us to obtain sharp insights about the economic forces behind the optimal provision of incentives that are more difficult to see in finite-period economies.

The key step in the analysis of Section 2.3.1 consisted of setting up the dual problem (17) and its recursive representation (18). In the finite-horizon setting, we were able to obtain the formulation (17) by proving the one-shot deviation principle (Proposition 2). Here, we start by *assuming* that the one-shot deviation principle holds, and solve a relaxed problem where the incentive constraints (8) are replaced with (10). We then show later in Proposition 4 that under some conditions, the solution to the relaxed problem is also a solution to the original problem. The infinite period analogue of the (relaxed) sequential dual problem is

$$K(v_0) \equiv \sup_{\mathbf{u}} \mathbb{E}_0 \left[-\sum_{t=1}^{\infty} \beta^{t-1} C(u_t) \right] \tag{21}$$
$$\text{subject to (11), (13), (14).}$$

We now show that the value function K defined in (21) can be written recursively, and that the solution to this recursive formulation can, under some conditions, recover the maximum to our primal problem (9). Let $\bar{v} = \frac{1}{1-\beta}\bar{u}$, $\underline{v} = \frac{1}{1-\beta}\underline{u}$, and let $\mathbb{V} = [\underline{v}, \bar{v})$ if the utility is bounded below and $\mathbb{V} = (\underline{v}, \bar{v})$ otherwise.[g] We denote by $B(v)$ the set of pairs $(\vec{u}, \vec{w}) = \left(\{u(\theta)\}_{\theta\in\Theta}, \{w(\theta)\}_{\theta\in\Theta}\right)$ that satisfy the constraints of the recursive problem, ie,

$$B(v) \equiv \left\{ (\vec{u}, \vec{w}) \in \mathbb{U}^{|\Theta|} \times \mathbb{V}^{|\Theta|} : (19), (20) \text{ hold} \right\}. \tag{22}$$

We first prove an infinite period analogue of the Bellman equation (18). Some of the arguments are based on those in Farhi and Werning (2007).

Proposition 3 *Suppose that the utility function satisfies Assumption 1, shocks satisfy Assumptions 2 and 3, and $T = \infty$. Then K satisfies the Bellman equation*

$$K(v) = \max_{(\vec{u}, \vec{w}) \in B(v)} \sum_{\theta\in\Theta} \pi(\theta) [-C(u(\theta)) + \beta K(w(\theta))]. \tag{23}$$

Proof We first show that the maximum in problem (23) is well defined. That is, for any $v \in \mathbb{V}$, there exist $(\vec{u}_v, \vec{w}_v)$ that maximize the right hand side of (23) within the set $B(v)$ defined in (22). To do so we restrict the optimization over $(\vec{u}, \vec{w})$ to a compact set.

[g] In our benchmark taste shock model it is easy to find the domain of K that we denote by $\mathbb{V}$. Any constant consumption sequence is incentive compatible. Since the consumption set is bounded below by 0, the greatest lower bound for the set $\mathbb{V}$ must be $\underline{v} = \sum_{\theta\in\Theta} \pi(\theta)\theta[U(0) + \beta\underline{v}] = \frac{1}{1-\beta}\underline{u}$, where we used the normalization $\mathbb{E}\theta = 1$. If $U(0)$ is finite, so is $\underline{v}$. Similarly, since the consumption set is unbounded above, $\bar{v} = \frac{1}{1-\beta}\bar{u}$ is the least upper bound of $\mathbb{V}$. Since (13) and (14) define a convex set, any $v_0 \in (\underline{v}, \bar{v})$ can be attained by incentive-compatible allocations, which establishes that $\mathbb{V} = [\underline{v}, \bar{v})$ if the utility is bounded below and $\mathbb{V} = (\underline{v}, \bar{v})$ otherwise. It is not always possible to characterize the domain of the value function in such a simple way. The general way to characterize the set $\mathbb{V}$ is described in Proposition 8.

Since the right hand side of (23) is a continuous function of $\left(\vec{u}, \vec{w}\right)$, this implies that it reaches its maximum.

The allocation $\left(\vec{u}\,', \vec{w}\,'\right)$ defined by $u'(\theta) = (1-\beta)v$ and $w'(\theta) = v$ for all $\theta \in \Theta$ satisfies the constraints (19) and (20) and yields value $-C((1-\beta)v) + \beta K(v) \equiv \underline{K}_v$. Therefore the r.h.s. of the Bellman equation is larger than $\underline{K}_v$. Now suppose that for some θ, $w(\theta)$ is such that $\beta\pi(\theta)K(w(\theta)) < \underline{K}_v$. Then we have

$$\sum_{\theta\in\Theta} \pi(\theta)[-C(u(\theta)) + \beta K(w(\theta))] < \underline{K}_v,$$

a contradiction. Thus we can restrict the search to $\left\{w(\theta) \text{ s.t. } \beta\pi(\theta)K(w(\theta)) \geq \underline{K}_v\right\}$ and, similarly, to $\left\{u(\theta) \text{ s.t. } -\pi(\theta)C(u(\theta)) \geq \underline{K}_v\right\}$. Moreover, we have $\lim_{u\to\bar{u}} - C(u) = -\infty$ and $\lim_{v\to\bar{v}} K(v) = -\infty$. To show the latter, consider the function $\bar{K}(v)$ which maximizes the objective function (21) subject to delivering lifetime utility $v_0 = v$, without the incentive constraints. Obviously $\bar{K}(v) \geq K(v)$. We easily obtain that the solution to this relaxed problem is $\bar{K}(v) = -\dfrac{1}{1-\beta}\mathbb{E}\left[C\left(C'^{-1}(\gamma_v\theta)\right)\right]$ where $\gamma_v > 0$ is the multiplier on the promise-keeping constraint. We have $\mathbb{E}\left[\theta C'^{-1}(\gamma_v\theta)\right] = (1-\beta)v$, so $\lim_{v\to\bar{v}}\gamma_v = \infty$ and hence $\lim_{v\to\bar{v}}\bar{K}(v) = -\infty$. This implies that $\lim_{v\to\bar{v}} K(v) = -\infty$, and therefore the previous arguments lead to upper bounds $\bar{\bar{u}}_\theta, \bar{\bar{v}}_\theta$ for $u(\theta)$ and $w(\theta)$, respectively. Moreover, $\mathbb{E}[\theta u(\theta) + \beta w(\theta)]$ goes to $-\infty$ if $u(\theta)$ or $w(\theta)$ go to $-\infty$ because of the upper bounds $\bar{\bar{u}}_\theta$, $\bar{\bar{w}}_\theta$. This contradicts the promise-keeping constraint and thus gives us lower bounds $\underline{\underline{u}}_\theta$, $\underline{\underline{w}}_\theta$ for all θ. Therefore, we can restrict the search for $\{u(\theta), w(\theta)\}_{\theta\in\Theta}$ to the compact set $\prod_{\theta\in\Theta}\left[\underline{\underline{u}}_\theta, \bar{\bar{u}}_\theta\right] \times \left[\underline{\underline{w}}_\theta, \bar{\bar{w}}_\theta\right]$. This concludes the proof that the maximum in the right hand side of (23) is attained.

Next, we show that K, the solution to (21), satisfies the Bellman equation (23). We start by showing that the left hand side is weakly smaller than the right hand side. Suppose that for some v, we have

$$K(v) > \max_{\left(\vec{u}, \vec{w}\right)\in B(v)} \sum_{\theta\in\Theta} \pi(\theta)[-C(u(\theta)) + \beta K(w(\theta))].$$

Thus there exists $\varepsilon > 0$ such that

$$K(v) \geq \mathbb{E}[-C(u(\theta)) + \beta K(w(\theta))] + \varepsilon, \quad \forall\left(\vec{u}\,, \vec{w}\right) \in B(v).$$

Now consider any allocation $\mathbf{u} = \{u_t(\theta^t)\}_{t\geq 1, \theta^t\in\Theta^t}$ that satisfies incentive compatibility (10) and delivers lifetime utility v. We can write $\mathbf{u} = \left(\{u_1(\theta_1)\}_{\theta_1\in\Theta}, \{\mathbf{u}_2(\theta_1)\}_{\theta_1\in\Theta}\right)$, where for all θ_1, $\mathbf{u}_2(\theta_1) = \left\{u_t\left(\theta_1, \theta_2^t\right)\right\}_{t\geq 2, \theta_2^t\in\Theta^{t-1}}$. Let $w_2(\theta_1)$ denote the lifetime utility achieved by $\mathbf{u}_2(\theta_1)$.[h] The pair $\left(\vec{u}_1, \vec{w}_2\right) = \left(\{u_1(\theta_1)\}_{\theta_1\in\Theta}, \{w_2(\theta_1)\}_{\theta_1\in\Theta}\right)$ satisfies (19) and (20), ie, $\left(\vec{u}_1, \vec{w}_2\right) \in B(v)$. Thus, the previous inequality implies that

$$\begin{aligned} K(v) &\geq \mathbb{E}[-C(u_1(\theta_1)) + \beta K(w_2(\theta_1))] + \varepsilon \\ &\geq \mathbb{E}\left[-C(u_1(\theta_1)) + \beta \mathbb{E}_1\left[-\sum_{t=2}^{\infty} \beta^{t-2} C\left(u_t\left(\theta_1, \theta_2^t\right)\right)\right]\right] + \varepsilon \\ &= \mathbb{E}_0\left[-\sum_{t=1}^{\infty} \beta^{t-1} C(u_t(\theta^t))\right] + \varepsilon, \end{aligned}$$

where the second inequality follows from the definition (21) of $K(w_2(\theta_1))$, since the allocation $\mathbf{u}_2(\theta_1)$ satisfies (10) and yields $w_2(\theta_1)$. Since this reasoning holds for any allocation $\mathbf{u}$ that satisfies (10) and delivers v, we get a contradiction.

Next we show the reverse inequality. Note that by definition of the supremum in (21), for all v and $\varepsilon > 0$ there exists an allocation $\tilde{\mathbf{u}}^{v,\varepsilon} = \left\{\tilde{u}_t^{v,\varepsilon}(\theta^t)\right\}$ that satisfies (10) and delivers v with cost

$$\mathbb{E}_0\left[-\sum_{t=1}^{\infty} \beta^{t-1} C\left(\tilde{u}_t^{v,\varepsilon}(\theta^t)\right)\right] > K(v) - \varepsilon.$$

Let

$$\left(\vec{u}_v, \vec{w}_v\right) \in \arg \max_{\left(\vec{u}, \vec{w}\right) \in B(v)} \mathbb{E}[-C(u(\theta)) + \beta K(w(\theta))].$$

Consider the incentive-compatible allocation $\mathbf{u}$ defined by $u_1\left(\theta^1\right) = u_v(\theta_1)$ for all $\theta_1 \in \Theta$ and $u_t\left(\theta_1, \theta_2^t\right) = \tilde{u}_t^{w_v(\theta_1),\varepsilon}(\theta^t)$ for all $t \geq 2$, $\theta^t \in \Theta^t$. We have

[h] Note that the continuation utilities, and in particular $w_2(\theta)$ for all θ, are well defined. Indeed, if not, then for some $s \geq 0$, $U_s^+ \equiv \lim_{T\to\infty} \mathbb{E}_s\left[\sum_{t=1}^{T} \beta^{t-1}\{\theta_t u_t(\theta^t) \vee 0\}\right] = \infty$. Since the cost function is convex, we have $C(u) \geq -B + A\{(\max_\Theta \theta) u \vee 0\}$ for some $A, B > 0$, and hence $\lim_{T\to\infty} \mathbb{E}_s\left[\sum_{t=1}^{T} \beta^{t-1} C(u_t(\theta^t))\right] \geq -\frac{B}{1-\beta} + AU_s^+ = \infty$. This implies

$$\lim_{T\to\infty} \mathbb{E}_0\left[\sum_{t=1}^{T} \beta^{t-1} C(u_t(\theta^t))\right] = \sum_{\theta^s \in \Theta^s} \pi_s(\theta^s) \lim_{T\to\infty} \mathbb{E}_s\left[\sum_{t=1}^{T} \beta^{t-1} C(u_t(\theta^t))\right] = \infty,$$

which contradicts the feasibility constraint (2).

$$
\begin{aligned}
K(v) &\geq \mathbb{E}_0\left[-\sum_{t=1}^{\infty}\beta^{t-1}C(u_t(\theta^t))\right]\\
&= \mathbb{E}_0\left[-C(u_v(\theta_1)) + \beta\mathbb{E}_1\left[-\sum_{t=2}^{\infty}\beta^{t-2}C\left(\tilde{u}_t^{w_v(\theta_1),\varepsilon}(\theta^t)\right)\right]\right]\\
&\geq \mathbb{E}_0[-C(u_v(\theta_1)) + \beta K(w_v(\theta_1))] - \beta\varepsilon\\
&= \max_{\left(\vec{u},\vec{w}\right)\in B(v)} \mathbb{E}[-C(u(\theta)) + \beta K(w(\theta))] - \beta\varepsilon.
\end{aligned}
$$

Since $\varepsilon > 0$ was arbitrary, we can let $\varepsilon \to 0$ in this inequality. We have thus shown that the value function (21) of the dual planner's problem satisfies the Bellman equation (23). □

The function K inherits the same properties as the function K_t in the finite period version of this economy.

Lemma 1 *Suppose that the utility function satisfies Assumption 1, shocks satisfy Assumptions 2 and 3, and $T = \infty$. Then K is continuous on $\mathbb{V}$, strictly concave, strictly decreasing, and differentiable, with $\lim_{v\to\underline{v}}K(v) = \lim_{v\to\underline{v}}K'(v) = 0$ and $\lim_{v\to\bar{v}}K(v) = \lim_{v\to\bar{v}}K'(v) = -\infty$.*

Proof The objective function in (21) is concave and the constraint set is convex; therefore, K is weakly concave. To show the strict concavity of K, pick any $v^a, v^b \in \mathbb{V}$ such that $v^a \neq v^b$, and let $\left(\vec{u}_{v^a}, \vec{w}_{v^a}\right)$ and $\left(\vec{u}_{v^b}, \vec{w}_{v^b}\right)$ be the corresponding policy functions that maximize the right hand side of (23). The incentive constraint (20) implies that $\vec{u}_{v^a} \neq \vec{u}_{v^b}$. Let, for $\alpha \in [0,1]$, $v^\alpha \equiv \alpha v^a + (1-\alpha)v^b$, and $\left(\vec{u}_{v^\alpha}, \vec{w}_{v^\alpha}\right)$ be the corresponding policy function. Since (19) and (20) are linear in $u(\theta)$ and $w(\theta)$, we obtain that

$$\left(\alpha\vec{u}_{v^a} + (1-\alpha)\vec{u}_{v^b}, \alpha\vec{w}_{v^a} + (1-\alpha)\vec{w}_{v^b}\right) \in B(v^\alpha).$$

Thus K satisfies

$$
\begin{aligned}
K(v^\alpha) &= \sum_{\theta\in\Theta}\pi(\theta)[-C(u_{v^\alpha}(\theta)) + \beta K(w_{v^\alpha}(\theta))]\\
&\geq \sum_{\theta\in\Theta}\pi(\theta)[-C(\alpha u_{v^a}(\theta) + (1-\alpha)u_{v^b}(\theta)) + \beta K(\alpha w_{v^a}(\theta) + (1-\alpha)w_{v^b}(\theta))].
\end{aligned}
$$

so that by the strict concavity of $-C$ and the weak concavity of K we get

$$K(v^{\alpha}) > \alpha \sum_{\theta \in \Theta} \pi(\theta)[-C(u_{v^a}(\theta)) + \beta K(w_{v^a}(\theta))]$$
$$+ (1-\alpha) \sum_{\theta \in \Theta} \pi(\theta)[-C(u_{v^b}(\theta)) + \beta K(w_{v^b}(\theta))] = \alpha K(v^a) + (1-\alpha) K\left(v^b\right),$$

Therefore K is strictly concave.

The concavity of K implies that it is continuous in the interior of $\mathbb{V}$ (Exercise 4.23 in Rudin, 1976). To show the continuity of K on $\mathbb{V}$ it remains to show that $\lim_{v \to \underline{v}} K(v) = K(\underline{v})$ when the utility is bounded below. Since the only feasible solution that delivers $\underline{v}$ has $u_t(\theta^t) = \underline{u}$ for all t, θ^t, we have in this case $K(\underline{v}) = -\frac{1}{1-\beta} C(\underline{u}) = 0$. Therefore showing the continuity at $\underline{v}$ is equivalent to showing that $\lim_{v \to \underline{v}} K(v) = 0$. Let $\underline{K}(v) = -\frac{1}{1-\beta} C((1-\beta)v)$ be the cost of delivering $u_t(\theta^t) = (1-\beta)v$ independently of θ^t. Since this allocation is incentive compatible, we have $0 \geq K(v) \geq \underline{K}(v)$ for all v. $\underline{K}$ is continuous on $\mathbb{V}$ with $\lim_{v \to \underline{v}} \underline{K}(v) = 0$; therefore, $\lim_{v \to \underline{v}} K(v) = 0$.

We already showed that $\lim_{v \to \bar{v}} K(v) = -\infty$ in the proof of Proposition 3.

To show the strict monotonicity, for any $v_0^a < v_0^b$ pick $v \in \left(\underline{v}, v_0^a\right)$ and $\alpha_v \in [0, 1)$ such that $v_0^a = \alpha_v v + (1-\alpha_v) v_0^b$. Since K is strictly concave, we have $K\left(v_0^a\right) > \alpha_v K(v) + (1-\alpha_v) K\left(v_0^b\right)$. Letting $v \to \underline{v}$ in this inequality, we obtain $K\left(v_0^a\right) \geq \left(1-\alpha_{\underline{v}}\right) K\left(v_0^b\right) \geq K\left(v_0^b\right)$, and hence K is weakly decreasing. But then using $K(v) \geq K\left(v_0^b\right)$ in the previous inequality leads to $K\left(v_0^a\right) > \alpha_v K\left(v_0^b\right) + (1-\alpha_v) K\left(v_0^b\right) = K\left(v_0^b\right)$, so that K is strictly decreasing.

Next we show the differentiability of the cost function K in the case where the utility is unbounded. A slightly different perturbational argument can be used to establish the differentiability when the utility function is bounded, taking care of the situations when the optimum is at the corners (see, eg, Farhi and Werning, 2007). Fix an interior v and define, for all $x \in (-\varepsilon, \varepsilon)$ for some small $\varepsilon > 0$,

$$L_v(x) = \sum_{\theta \in \Theta} \pi(\theta)[-C(u_v(\theta) + x) + \beta K(w_v(\theta))].$$

The allocation $\left(\vec{u}_x, \vec{w}_x\right)$ with $u_x(\theta) = u_v(\theta) + x$ and $w_x(\theta) = w_v(\theta)$ for all θ is incentive compatible and delivers lifetime utility $v + x$. Therefore, for all x we have $L_v(x) \leq K(v+x)$, with equality if $x = 0$. Since $L_v(\cdot)$ is concave and differentiable on $(-\varepsilon, \varepsilon)$ (because $-C(\cdot)$ is), the Benveniste–Scheinkman theorem (Benveniste and Scheinkman, 1979, or Theorem 4.10 in Stokey et al., 1989) implies that K is differentiable at v and we have $K'(v) = L_v'(0)$. Direct calculation of $L_v'(0)$ shows that

$$K'(v) = \sum_{\theta \in \Theta} \pi(\theta)[-C'(u_v(\theta))] \leq 0. \tag{24}$$

The bounds $\underline{K}(v) \le K(v) \le \bar{K}(v)$ (see the proof of Proposition 3) and the limits $\lim_{v\to \underline{v}} \underline{K}'(v) = 0$ and $\lim_{v\to\bar{v}} \bar{K}'(v) = -\gamma_v = -\infty$ imply that $\lim_{v\to\underline{v}} K'(v) = 0$ and $\lim_{v\to\bar{v}} K'(v) = -\infty$. □

Finally, we are ultimately interested in recovering a solution to problem (9). Analogous to the finite period case, we call the solution to (23) a *policy function* and denote it by $\left(\vec{u}_v, \vec{w}_v\right)$. For any initial v_0 these functions *generate* $(\mathbf{u}, \mathbf{v})$ as in Section 2.3.1.

Proposition 4 *Suppose that the utility function satisfies Assumption 1, shocks satisfy Assumptions 2 and 3, and* $T = \infty$. *Let* v_0 *be defined by* $K(v_0) = -\frac{e}{1-\beta}$. *If the sequence* $(\mathbf{u},\ \mathbf{v})$ *generated by the policy functions to the Bellman equation (23) given* v_0 *satisfies*

$$\lim_{t\to\infty} \mathbb{E}_0[\beta^t v_t(\theta^t)] = 0 \tag{25}$$

and

$$\lim\sup_{t\to\infty} \mathbb{E}_0[\beta^t v_t(\sigma^t(\theta^t))] \ge 0, \forall \sigma \tag{26}$$

then $(\mathbf{u},\ \mathbf{v})$ *achieves the supremum of the primal maximization problem (9).*

Proof Let $(\mathbf{u}, \mathbf{v})$ denote the allocations generated by the policy functions $\left(\vec{u},\ \vec{w}\right)$ starting at v_0. First, we show that $(\mathbf{u}, \mathbf{v})$ achieves the supremum of the sequential *dual* problem (21) with the full set of incentive constraints (8) (rather than only the constraints (10) of the relaxed problem), ie, that $(\mathbf{u}, \mathbf{v})$ satisfies the constraints (8) and (14) and attains $K(v_0)$. To see that constraint (14) is satisfied, note that by repeated substitution, $(\mathbf{u}, \mathbf{v})$ satisfies

$$v_0 = \mathbb{E}_0\left[\sum_{t=1}^{T} \beta^{t-1} \theta_t u_t(\theta^t)\right] + \beta^T \mathbb{E}_0\left[v_T\left(\theta^T\right)\right].$$

If $(\mathbf{u}, \mathbf{v})$ satisfies (25), then taking limits as $T \to \infty$ (see Footnote h for the existence of the limit on the right hand side) leads to

$$v_0 = \mathbb{E}_0\left[\sum_{t=1}^{\infty} \beta^{t-1} \theta_t u_t(\theta^t)\right].$$

To see that $(\mathbf{u}, \mathbf{v})$ satisfies the incentive-compatibility constraint (8), consider any reporting strategy σ. Since the policy functions $\left(\vec{u},\ \vec{w}\right)$ that generate $(\mathbf{u}, \mathbf{v})$ satisfy (20), repeated substitution implies that $(\mathbf{u}, \mathbf{v})$ satisfies

$$v_0 \ge \mathbb{E}_0\left[\sum_{t=1}^{T} \beta^{t-1} \theta_t u(\sigma^t(\theta^t))\right] + \beta^T \mathbb{E}_0\left[v_T\left(\sigma^T\left(\theta^T\right)\right)\right].$$

If the condition (26) is satisfied, taking limits implies that

$$\limsup_{T\to\infty}\left\{v_0-\mathbb{E}_0\left[\sum_{t=1}^{T}\beta^{t-1}\theta_t U(c_t(\sigma^t(\theta^t)))\right]\right\}\geq 0\,\forall\sigma,$$

establishing that $(\mathbf{u},\mathbf{v})$ satisfies (8).

We next show that $(\mathbf{u},\mathbf{v})$ attains $K(v_0)$. Repeatedly applying the Bellman equation (23) yields

$$K(v_0)=\mathbb{E}_0\left[-\sum_{t=1}^{T}\beta^{t-1}C(u_t(\theta^t))\right]+\beta^T\mathbb{E}_0\left[K\left(v_T\left(\theta^T\right)\right)\right].$$

Since $\limsup_{T\to\infty}\beta^T\mathbb{E}_0\left[K\left(v_T\left(\theta^T\right)\right)\right]\leq 0$ we obtain

$$K(v_0)\leq\mathbb{E}_0\left[-\sum_{t=1}^{\infty}\beta^{t-1}C(u_t(\theta^t))\right].$$

But $(\mathbf{u},\mathbf{v})$ satisfies the constraints of problem (21), thus $K(v_0)\geq\mathbb{E}_0\left[-\sum_{t=1}^{\infty}\beta^{t-1}C(u_t(\theta^t))\right]$. Therefore $(\mathbf{u},\mathbf{v})$ achieves the supremum of the dual problem (21).

Second, we show that the maximum to the dual problem (21) is also a maximum to the primal problem (9). Since $(\mathbf{u},\mathbf{v})$ delivers v_0 which satisfies $-K(v_0)=\frac{e}{1-\beta}$, $\mathbf{u}$ satisfies the feasibility constraint (2) and therefore $V(e)\geq v_0$. Suppose that this inequality is strict, so that there exists $(\mathbf{u}',\mathbf{v}')$ that delivers lifetime utility $v_0{}'>v_0$, is incentive compatible, and satisfies $\mathbb{E}_0\left[\sum_{t=1}^{\infty}\beta^{t-1}C\left(u_t'\right)\right]\leq\frac{e}{1-\beta}$. The continuity and strict monotonicity of K (Lemma 1) imply that $-K(v_0')>-K(v_0)=\frac{e}{1-\beta}$. Since $\mathbb{E}_0\left[\sum_{t=1}^{\infty}\beta^{t-1}C\left(u_t'\right)\right]\geq -K(v_0')$, this establishes a contradiction. □

If the utility function is bounded, then the limiting conditions (25) and (26) are automatically satisfied and Proposition 4 implies simultaneously that the supremum to problem (21) is attained and that it can be recovered from the policy functions of the Bellman equation (23). When the utility function is unbounded, an extra step is needed to verify that the policy functions generate a solution that satisfies the conditions (25) and (26). We show in an example in Section 2.4 how to verify ex post these conditions with unbounded utilities.

The analysis above can be simplified if we modify Assumption 1 and assume that the domain of U is compact. In this case $C(\cdot)$ is a bounded function on a compact set $[\underline{u},\bar{u}]$. The results of Propositions 3 and 4 can be proven immediately using standard contraction mapping arguments (see Chapter 9 in Stokey et al., 1989). Moreover, the results of Stokey et al. (1989) show that the function K that satisfies the functional equation

(23) is the unique fixed point of the Bellman operator defined on the space of continuous and bounded functions by

$$\mathscr{B}(k)(v) = \max_{(\vec{u},\vec{w}) \in B(v)} \sum_{\theta \in \Theta} \pi(\theta)[-C(u(\theta)) + \beta k(w(\theta))],$$

and that for all bounded and continuous k_0 the sequence $\{k_n\}_{n \geq 0}$ defined by $k_n = \mathscr{B}^n k_0$ for all n converges to K. This characterization can be used to compute the solution to the problem numerically.

2.4 Characterization of the Solution with i.i.d. Shocks

2.4.1 Optimal Incentive Provision

In this section we characterize the solution to the Bellman equation (23). At the end of this section we provide a simple example showing how to verify the limiting conditions (25) and (26) when the utility function is unbounded.

For simplicity, we assume that θ can take only two values, $\Theta = \{\theta_{(1)}, \theta_{(2)}\}$, with $\theta_{(1)} < \theta_{(2)}$. The incentive constraints (20) with two shocks reduce to

$$\theta_{(1)} u(\theta_{(1)}) + \beta w(\theta_{(1)}) \geq \theta_{(1)} u(\theta_{(2)}) + \beta w(\theta_{(2)}), \tag{27}$$

and

$$\theta_{(2)} u(\theta_{(2)}) + \beta w(\theta_{(2)}) \geq \theta_{(2)} u(\theta_{(1)}) + \beta w(\theta_{(1)}). \tag{28}$$

Proposition 5 *Suppose that the utility function satisfies Assumption 1, shocks satisfy Assumptions 2 and 3, $|\Theta| = 2$, and $T = \infty$. The constraint (27) binds, and the constraint (28) is slack for all interior v. Moreover $u_v(\theta_{(1)}) \leq u_v(\theta_{(2)})$ and $w_v(\theta_{(1)}) \geq v \geq w_v(\theta_{(2)})$, with strict inequalities for all interior v. The policy functions $u_v(\theta)$, $w_v(\theta)$ are continuous in v for all $\theta \in \Theta$. If $w_v(\theta_{(2)})$ is interior, the policy functions satisfy*

$$K'(v) = \mathbb{E}[K'(w_v)] = \mathbb{E}[-C'(u_v)], \forall v. \tag{29}$$

Proof The proof proceeds by guessing that the constraint (28) is slack and solving a relaxed problem (23) in which this constraint is dropped. We then verify ex post that (28) is satisfied. The strict concavity of the objective function in (23) and the convexity of the constraint set then implies that the solution to the relaxed problem is the unique solution to the original problem.

Let $\xi_v \geq 0$ and $\gamma_v \geq 0$ be the Lagrange multipliers on the incentive-compatibility constraint (27) and the promise-keeping constraint (19) in the relaxed problem. The first-order conditions with respect to $u(\theta_{(1)})$ and $u(\theta_{(2)})$ are

$$\pi\left(\theta_{(1)}\right)C'\left(u_v\left(\theta_{(1)}\right)\right)-\xi_v\theta_{(1)} \ \geq \gamma_v\pi\left(\theta_{(1)}\right)\theta_{(1)}, \tag{30}$$

$$\pi\left(\theta_{(2)}\right)C'\left(u_v\left(\theta_{(2)}\right)\right)+\xi_v\theta_{(1)} \ \geq \gamma_v\pi\left(\theta_{(2)}\right)\theta_{(2)}, \tag{31}$$

where these constraints hold with equality if $u_v\left(\theta_{(1)}\right)>\underline{u}$ and $u_v\left(\theta_{(2)}\right)>\underline{u}$, respectively. Similarly, the first-order conditions with respect to $w\left(\theta_{(1)}\right)$ and $w\left(\theta_{(2)}\right)$ are

$$-\pi\left(\theta_{(1)}\right)K'\left(w_v\left(\theta_{(1)}\right)\right)-\xi_v \ \geq \gamma_v\pi\left(\theta_{(1)}\right), \tag{32}$$

$$-\pi\left(\theta_{(2)}\right)K'\left(w_v\left(\theta_{(2)}\right)\right)+\xi_v \ \geq \gamma_v\pi\left(\theta_{(2)}\right), \tag{33}$$

where these constraints hold with equality if $w_v\left(\theta_{(1)}\right)>\underline{v}$ and $w_v\left(\theta_{(2)}\right)>\underline{v}$, respectively.

We first show that $u_v\left(\theta_{(1)}\right), w_v\left(\theta_{(1)}\right)$ are interior for all interior v. (We show below that $u_v\left(\theta_{(2)}\right)$ is also interior.) Suppose that $u_v\left(\theta_{(1)}\right)=\underline{u}$. Since $C'(\underline{u})=0$, (30) implies that $\xi_v = \gamma_v = 0$. If $u_v\left(\theta_{(2)}\right)>\underline{u}$, then (31) would hold with equality, implying $C'\left(u_v\left(\theta_{(2)}\right)\right)=0$, a contradiction. Thus we have $u_v\left(\theta_{(1)}\right)=u_v\left(\theta_{(2)}\right)=\underline{u}$, and the same reasoning implies that $w_v\left(\theta_{(1)}\right)=w_v\left(\theta_{(2)}\right)=\underline{v}$, which contradicts the promise-keeping constraint (19) when v is interior. Therefore we must have $u_v\left(\theta_{(1)}\right)>\underline{u}$ so that (30) holds with equality. An identical reasoning implies that $w_v\left(\theta_{(1)}\right)>\underline{v}$, so that (32) holds with equality.

We now show that $\xi_v > 0$ for all interior v. If $\xi_v = 0$, then (32) and (33) imply that $w_v\left(\theta_{(2)}\right)\geq w_v\left(\theta_{(1)}\right)$ by the concavity of K. Moreover (30) and (31) with $\theta_{(2)}>\theta_{(1)}$ imply that $u_v\left(\theta_{(2)}\right)>u_v\left(\theta_{(1)}\right)$. This violates the incentive constraint (27), and hence $\xi_v > 0$ if $v>\underline{v}$. This implies that the constraint (27) holds with equality for all $v>\underline{v}$, and it also trivially holds as an equality for $v=\underline{v}$.

We show next that the solution to the relaxed problem satisfies (28). Suppose not, ie,

$$\theta_{(2)}u_v\left(\theta_{(2)}\right)+\beta w_v\left(\theta_{(2)}\right)<\theta_{(2)}u_v\left(\theta_{(1)}\right)+\beta w_v\left(\theta_{(1)}\right).$$

Sum this equation with (27) which holds with equality, to obtain $u_v\left(\theta_{(2)}\right)<u_v\left(\theta_{(1)}\right)$, and thus $w_v\left(\theta_{(2)}\right)>w_v\left(\theta_{(1)}\right)>\underline{v}$. This implies that (33) holds with equality. But (32) and (33) with $\xi_v\geq 0$ then imply that $w_v\left(\theta_{(2)}\right)\leq w_v\left(\theta_{(1)}\right)$, a contradiction. Therefore the incentive constraint (28) is satisfied in the relaxed problem for all v. Moreover, if v is interior, the same reasoning with $\xi_v > 0$ implies that (28) is slack.

Summing the incentive constraints (27) and (28) implies $u_v\left(\theta_{(2)}\right)\geq u_v\left(\theta_{(1)}\right)$ and hence $w_v\left(\theta_{(1)}\right)\geq w_v\left(\theta_{(2)}\right)$. In particular, $u_v\left(\theta_{(2)}\right)$ is interior for all interior v, and (31) holds with equality. Now suppose $v>\underline{v}$. If $w_v\left(\theta_{(2)}\right)=\underline{v}$, then $w_v\left(\theta_{(1)}\right)>w_v\left(\theta_{(2)}\right)$. If $w_v\left(\theta_{(2)}\right)>\underline{v}$, then (33) holds with equality, and (32) with $\xi_v > 0$ yields $w_v\left(\theta_{(1)}\right)>w_v\left(\theta_{(2)}\right)$ by the strict concavity of K. We then obtain $u_v\left(\theta_{(1)}\right)<u_v\left(\theta_{(2)}\right)$ from (27). When v is interior, we saw that $u_v(\theta)$ is interior for all θ and therefore Benveniste–Scheinkman arguments (see the arguments leading to Eq. (24)), or using the envelope

theorem and summing (30) and (31), establish that $K'(v)=\mathbb{E}[-C'(u_v)]=-\gamma_v$. This equation also holds at the boundary $v=\underline{v}$ since in this case both sides of this expression are equal to zero. Therefore, Eqs. (32) and (33), assuming that $w_v(\theta_{(2)})$ is interior, imply that $-K'(w_v(\theta_{(1)}))>-K'(v)$ and $-K'(w_v(\theta_{(2)}))<-K'(v)$, respectively, so that $w_v(\theta_{(2)})<v<w_v(\theta_{(1)})$.

Next we show that the policy functions are continuous in v. The objective function in (23) is continuous and strictly concave on $\mathbb{U}^2\times\mathbb{V}^2$. Following the same steps as in the proof of Proposition 3, we restrict the optimization over $(\vec{u},\vec{w})$ to a compact set $\mathbb{X}\subset\mathbb{U}^2\times\mathbb{V}^2$. The constraint set $B(\cdot):\mathbb{V}\to\mathbb{X}$ defined in (22) is then a continuous, compact-valued, and convex-valued correspondence. Thus, by the Theorem of the Maximum (see, eg, Theorem 3.6 and Exercise 3.11a in Stokey et al., 1989), the function $(\vec{u}_v,\vec{w}_v)$ is continuous in v.

We now prove Eq. (29). We saw above that for all $v\geq\underline{v}$, $K'(v)=\mathbb{E}[-C'(u_v)]=-\gamma_v$. Moreover, when v is interior, summing the conditions (30) and (31) (which both hold with equality), and the conditions (32) and (33) (the former holds with equality, the latter does as well if $w_v(\theta_{(2)})>\underline{v}$) yields $\mathbb{E}[K'(w_v)]\leq\mathbb{E}[-C'(u_v)]$, with equality if $w_v(\theta_{(2)})$ is interior. Finally, this equation holds with equality if $v=\underline{v}$.

Note finally that from (30) and (31), we obtain that

$$-\theta_{(1)}K'(v)<C'(u_v(\theta_{(1)}))<-K'(v)<C'(u_v(\theta_{(2)}))<-\theta_{(2)}K'(v), \tag{34}$$

for all interior v. □

Proposition 5 highlights the main principle underlying the optimal provision of incentives in dynamic economies. Consider the unconstrained first-best allocation, given by

$$\frac{1}{\theta_{(1)}}C'\left(u_v^{\text{fb}}(\theta_{(1)})\right)=\frac{1}{\theta_{(2)}}C'\left(u_v^{\text{fb}}(\theta_{(2)})\right). \tag{35}$$

In this case the future continuation allocations are independent of the current realization of the shock: the social planner redistributes resources from agents with shock $\theta_{(1)}$ to agents with shock $\theta_{(2)}$. As we discussed above, this allocation is not incentive compatible when shocks are private information. To provide incentives, the social planner spreads out future promised utilities w_v, rewarding agents who report a lower shock and punishing those who report a higher shock. In exchange, a higher reported shock gives the agent a higher utility today. The bounds (34) imply

$$\frac{1}{\theta_{(2)}}C'(u_v(\theta_{(2)}))<\frac{1}{\theta_{(1)}}C'(u_v(\theta_{(1)})).$$

Therefore the spread in the current period utilities (or consumption allocations) is not as large as in the first best allocation. This reflects the fact that private information makes redistribution more costly. The resources are still being redistributed away from the $\theta_{(1)}$ type, but only up to the point where his incentive constraint binds.

Eq. (29) shows how the planner allocates the costs of providing incentives over time. Fluctuations in promised utilities are costly due to the concavity of the cost function, and it is optimal to smooth these costs over time. The best smoothing can be achieved if the forecast of future marginal costs of providing incentives based on the current information, $\mathbb{E}_t[K'(v_{t+s})]$, is equal to the current period marginal cost, $K'(v_t)$, so that $K'(v_t)$ is a random walk. This result is a manifestation of the same general principle that underlies consumption smoothing in the permanent income hypothesis (see Friedman, 1957; Hall, 1978) or tax smoothing in public finance (see Barro, 1979).

2.4.2 Long-Run Immiseration

Analogous to other environments with cost smoothing, the random walk nature of $K'(v_t)$ has powerful implications about the long-run properties of the solution.[i] To derive these implications, we first introduce the notion of *martingale* and the Martingale Convergence Theorem (see Billingsley, 2008, Section 35):

Definition 1 Let $X_t(\theta^t)$ be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The sequence $\{X_t, \mathcal{F}_t\}_{t=1,2,\ldots}$ is a *martingale* if:

(i) $\{\mathcal{F}_t\}_{t\geq 1}$ is an increasing sequence of σ-algebras,
(ii) X_t is measurable with respect to $\mathcal{F}_t$,
(iii) $\mathbb{E}_0[|X_t|] < \infty$, and
(iv) $\mathbb{E}_t[X_{t+1}] = X_t$ with probability 1.

A *submartingale* is defined as above except that the condition (iv) is replaced by $\mathbb{E}_t[X_{t+1}] \geq X_t$. Any martingale is a submartingale. We have the following important result (Theorem 35.5 in Billingsley, 2008):

Theorem 2 (Martingale Convergence Theorem) *Let $\{X_t\}_{t=0}^{\infty}$ be a submartingale. If $M \equiv \sup_t \mathbb{E}_0[|X_t|] < \infty$, then $X_t \to X$ with probability 1, where X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ satisfying $\mathbb{E}_0[|X|] \leq M$.*

To apply this result in our context, observe that the policy functions induce a law of motion for the distribution of promised utilities over time. For any probability distribution Ψ on $\mathbb{V}$, define an operator $\mathscr{T}$ as follows. For all Borel sets $A \subset \mathbb{V}$, let

$$(\mathscr{T}\Psi)(A) \equiv \int_{\mathbb{V}} \left[\sum_{\theta \in \Theta} \pi(\theta) \mathbb{I}_{\{w_v(\theta) \in A\}} \right] \Psi(dv). \tag{36}$$

$\mathscr{T}\Psi$ defines another probability distribution on $\mathbb{V}$. This operator allows us to study the dynamics of the distribution of lifetime utilities in our economy. In particular, let Ψ_0 be a

[i] See Chamberlain and Wilson (2000) for an analogous result in consumption smoothing models, and Aiyagari et al. (2002) for tax smoothing.

probability distribution on $\mathbb{V}$ that assigns probability 1 to v_0 and define $\boldsymbol{\Psi}_t$ recursively as $\boldsymbol{\Psi}_t = \mathscr{T}\boldsymbol{\Psi}_{t-1}$. In other words, initially in period 0 everyone is identical with the same lifetime utility v_0. Over time idiosyncratic shocks lead to inequality in lifetime promises captured by the distribution $\boldsymbol{\Psi}_t$ in period t. A distribution $\boldsymbol{\Psi}$ is *invariant* if it satisfies $\boldsymbol{\Psi} = \mathscr{T}\boldsymbol{\Psi}$.

Suppose that the utility function is unbounded below, so that $w_v\left(\theta_{(2)}\right)$ is always interior. Consider the random variable $K'(v_t(\theta^t))$ defined recursively on the probability space $(\Theta^\infty, \mathcal{B}(\Theta^\infty), \pi_\infty)$ starting at $K'(v_0)$. The sequence $\{K'(v_t(\theta^t)), \mathcal{B}(\Theta^t)\}_{t=1,2,\ldots,}$ is a martingale. Indeed, $\{\mathcal{B}(\Theta^t)\}_{t\geq 1}$ is an increasing sequence of σ-algebras, $K'(v_t)$ is measurable with respect to $\mathcal{B}(\Theta^t)$, $\mathbb{E}_0[|K'(v_t)|] = -\mathbb{E}_0[K'(v_t)] = -K'(v_0) < \infty$, and $\mathbb{E}_t[K'(v_{t+1})] = K'(v_t)$ follows from Proposition 5. Hence all the conditions of Definition 1 are satisfied, and Theorem 2 implies that $K'(v_t)$ converges almost surely to a random variable X. That is, for almost all histories $\theta^\infty \in \Theta^\infty$, we have $K'(v_t(\theta^t)) \to X(\theta^\infty)$. The following proposition, the proof of which follows Thomas and Worrall (1990), further characterizes the limit of the sequence:

Proposition 6 *Suppose that the utility function satisfies Assumption 1, shocks satisfy Assumptions 2 and 3, $|\Theta| = 2$, and $T = \infty$.*[j] *If the utility function is unbounded below, then $v_t(\theta^t) \to -\infty$ as $t \to \infty$ with probability 1. If the utility function is bounded below by $\underline{v}$, then the unique invariant distribution of continuation utilities on $\mathbb{V}$ assigns mass 1 to the lower bound $\underline{v}$.*

Proof Suppose that the utility function is unbounded below. Then $K'(v_t(\theta^t))$ is a martingale, and Theorem 2 implies that for almost all θ^∞, $K'(v_t(\theta^t))$ converges to some random variable $X(\theta^\infty)$. We now show that its limit $X(\theta^\infty)$ is equal to 0 almost surely.

Consider a path θ^t such that $K'(v_t(\theta^t)) \to \kappa < 0$. The sequence $v_t(\theta^t)$ thus converges to $\hat{v} > \underline{v}$, solution to $K'(\hat{v}) = \kappa$. With probability 1, the state $\theta_{(2)}$ occurs infinitely often on this path. Take the subsequence composed of the dates $\{t_n\}_{n=1,2,\ldots}$ where the state $\theta_{(2)}$ occurs. We have $\lim_{n\to\infty} v_{t_n-1}\left(\theta^{t_n-1}\right) = \hat{v}$ and $\lim_{n\to\infty} v_{t_n}\left(\theta^{t_n}\right) = \hat{v}$. Since the policy function $w_v(\theta)$ is continuous in v for all $\theta \in \Theta$, we obtain

$$\lim_{n\to\infty} w_{v_{t_n-1}\left(\theta^{t_n-1}\right)}\left(\theta_{(2)}\right) = w_{\hat{v}}\left(\theta_{(2)}\right).$$

But $w_{v_{t_n-1}\left(\theta^{t_n-1}\right)}\left(\theta_{(2)}\right) = v_{t_n}\left(\theta^{t_n-1}, \theta_{(2)}\right) = v_{t_n}\left(\theta^{t_n}\right)$, hence we also have

$$\lim_{n\to\infty} w_{v_{t_n-1}\left(\theta^{t_n-1}\right)}\left(\theta_{(2)}\right) = \hat{v}.$$

This implies that $w_{\hat{v}}\left(\theta_{(2)}\right) = \hat{v}$, which contradicts the inequality $w_{\hat{v}}\left(\theta_{(2)}\right) < \hat{v}$ proved in Proposition 5.

[j] The condition $|\Theta| = 2$ is not important for this proposition. It is easy to show that the martingale property (29) holds for any number of shocks.

Now suppose that the utility function is bounded below by $\underline{v} > -\infty$. In this case $-K'(v_t(\theta^t))$ is a (possibly unbounded) submartingale. Note that the point $\underline{v}$ is absorbing, ie, $u_{\underline{v}}(\theta) = \underline{u}$ and $w_{\underline{v}}(\theta) = \underline{v}$ for all $\theta \in \Theta$. Consider an invariant distribution Ψ of continuation utilities on $\mathbb{V}$, and let $\mathrm{Supp}(\Psi) \subset \mathbb{V}$ denote its support. Let $\mathcal{M}_v$ denote the Markov chain characterizing the law of motion of continuation utilities, starting at v. Define the set $\mathrm{S}_1 \subset \mathrm{Supp}(\Psi)$ consisting of all the continuation utility values v for which $\mathcal{M}_v$ reaches $\underline{v}$ in a finite number of steps with positive probability, and the set $\mathrm{S}_2 = \mathrm{Supp}(\Psi) \setminus \mathrm{S}_1$. By construction of S_1, every state $v \in \mathrm{S}_1 \setminus \{\underline{v}\}$ is transient, so that such a v cannot be in the support of the invariant distribution. Now, for $v \in \mathrm{S}_2$, the Markov chain $\mathcal{M}_v$ defines a sequence $\{v_t(\theta^t)\}_{t,\theta^t}$. By construction of S_2, the process $K'(v_t(\theta^t))$ is a martingale and the previous arguments show that v cannot be in the support of the invariant distribution. Therefore $\mathrm{Supp}(\Psi) = \{\underline{v}\}$. □

The result of Proposition 6 is often referred to as the *immiseration* result. It shows that a feature of the optimal contract is that agents' consumption, c_t^*, goes to 0 with probability 1 as $t \to \infty$. When the utility function is unbounded below, this implies that agents' utility diverges to $-\infty$; otherwise the only invariant distribution is degenerate and assigns probability 1 to $\dfrac{U(0)}{1-\beta}$. Note that the fact that $c_t^* \to 0$ with probability 1 does *not* mean that everyone's consumption converges to zero. As we saw in Proposition 5, an agent with shock realization $\theta_{(1)}$ always gets a strictly higher promised utility (and hence consumption) in the future. Thus there are always some agents (whose measure goes to zero as $t \to \infty$) with strictly positive consumption. In Section 3.1.2 we shut down the intertemporal transfer of resources and show that the immiseration result still holds in this setting. This will imply that in order to provide incentives for agents to reveal their private information, the planner needs to increase inequality without bounds over time. As time goes to infinity this inequality grows until a measure 0 of agents consume the entire endowment of the economy.

The intuition for this result is as follows. To provide agents with incentives to reveal information in the current period, the principal needs to commit to increasing inequality (in promised utilities and therefore consumption) in the future. When the interest rate is equal to the discount factor, as we assumed in this section, there are no offsetting forces and inequality under the optimal contract grows over time. In the infinite period economy it approaches an extreme level as $t \to \infty$, where only a measure zero of agents have positive consumption. We revisit this result in subsequent sections, especially in Sections 2.4.3 and 3.2.

2.4.3 Existence of a Nondegenerate Invariant Distribution

In this section we show in a simple example that there can exist a nondegenerate invariant distribution of utilities if additional constraints are imposed. We study the case where the planner is required to promise future utilities in a compact set $[\underline{w}, \bar{w}]$, with $\underline{v} < \underline{w} < \bar{w} < \bar{v}$. In Section 3.2 we show how similar constraints can emerge from more sophisticated political economy arguments, but for now we simply impose[k]

$$w(\theta) \in [\underline{w}, \bar{w}], \forall \theta \in \Theta \tag{37}$$

in problem (23). It is easy to see that the solution to the modified Bellman equation continues to satisfy the results of Lemma 1 except for the fact that $K'(v)$ is now strictly negative and finite on the set $[\underline{w}, \bar{w}]$.

Lemma 2 *Suppose that all the assumptions of Proposition 5 are satisfied and in addition the constraint (37) is imposed. Then there are no absorbing points:* $w_v(\theta_{(1)}) > w_v(\theta_{(2)})$ *for all* $v \in [\underline{w}, \bar{w}]$.

Proof Suppose $w_v(\theta_{(1)}) = w_v(\theta_{(2)}) = v \in [\underline{w}, \bar{w})$ for some v, implying (from (19) and (27)) that $u_v(\theta_{(1)}) = u_v(\theta_{(2)}) = (1-\beta)v$. Thus $K(v) = \underline{K}(v) \equiv -\frac{1}{1-\beta}C((1-\beta)v)$, where $\underline{K}(v)$ was defined in the proof of Lemma 1. Subtracting Eqs. (30) from (32) (written as an inequality because of the condition $w_v(\theta_{(1)}) \geq \underline{w}$), we obtain

$$K'(v) \leq -\frac{1}{\theta_{(1)}} C'((1-\beta)v) = \frac{1}{\theta_{(1)}} \underline{K}'(v).$$

But $\theta_{(1)} < 1$ contradicts the fact that $K(v') \geq \underline{K}(v')$ for all $v' > v$. For $v = \bar{w}$, a similar reasoning with Eqs. (31) and (33) (written as an inequality because of the condition $w_v(\theta_{(2)}) \leq \bar{w}$) implies that $K'(w_v(\theta_{(2)})) \geq \frac{1}{\theta_{(2)}} \underline{K}'(w_v(\theta_{(2)})) + \frac{\xi_v}{\pi(\theta_{(2)})}\left(1 - \frac{\theta_{(1)}}{\theta_{(2)}}\right)$ and leads to the same conclusion. □

We show next that a nondegenerate long-run distribution of utilities and consumption exists.

Proposition 7 *Suppose that all the assumptions of Proposition 5 are satisfied and in addition the constraint (37) is imposed. Then there exists a unique invariant and nondegenerate distribution* Ψ *of utilities, and for any initial measure* Ψ_0 *on the state space,* $\mathcal{T}^n(\Psi_0)$ *converges to* Ψ^* *as* $n \to \infty$ *at a geometric rate that is uniform in* Ψ_0.

Proof The result follows from Theorem 11.12 in Stokey et al. (1989), which holds if condition M (p. 348 in Stokey et al., 1989) is satisfied. To show this condition, it is

[k] Our arguments here adapt Atkeson and Lucas (1995) and Phelan (1995). See also Farhi and Werning (2007) and Hosseini et al. (2013).

sufficient to prove that there exists $\underline{\varepsilon} > 0$ and an integer $N < \infty$ such that, for all $v \in [\underline{w}, \bar{w}]$, $P^N(v, \underline{w}) \geq \underline{\varepsilon}$, where P denotes the transition matrix of the Markov chain $\mathcal{M}$ that characterizes the law of motion of continuation utilities; that is, the probability of reaching $\underline{w}$ in N steps starting from any v is at least as large than $\underline{\varepsilon}$.

To show this we proceed in two steps. First, we prove that if the continuation utility v in the current period is close enough to $\underline{w}$, receiving a high taste shock $\theta_{(2)}$ implies that the promised utility in the next period is $\underline{w}$. That is, there exists $\varepsilon > 0$ such that, for all $v \leq \underline{w} + \varepsilon$, we have $w_v(\theta_{(2)}) = \underline{w}$. Suppose by contradiction that this is not the case, and consider a sequence $v_n > \underline{w}$ with $\lim_{n \to \infty} v_n = \underline{w}$, such that $w_{v_n}(\theta_{(2)}) > \underline{w}$ for all n. The martingale property (29) then writes

$$K'(v_n) = \pi(\theta_{(1)}) K'(w_{v_n}(\theta_{(1)})) + \pi(\theta_{(2)}) K'(w_{v_n}(\theta_{(2)})).$$

Letting $n \to \infty$ in this equation imposes $w_{\underline{w}}(\theta_{(1)}) = w_{\underline{w}}(\theta_{(2)}) = \underline{w}$, which contradicts Lemma 2.

Second, we prove that there exists $\delta > 0$ such that, for any $v > \underline{w} + \varepsilon$, receiving a high taste shock $\theta_{(2)}$ implies that the promised utility in the next period, $w_v(\theta_{(2)})$, is smaller than $v - \delta$. To show this, note that since $w_v(\theta_{(2)})$ is continuous in v, it is either bounded away from the 45 degree line for $v \geq \underline{w} + \varepsilon$, or $w_v(\theta_{(2)}) = v$ for some $v \in [\underline{w} + \varepsilon, \bar{w}]$. By the martingale property, the latter implies that $w_v(\theta_{(1)}) = v$, contradicting Lemma 2.

These results imply that there exists $N < \infty$ such that, for any $v \in [\underline{w}, \bar{w}]$, the promised utility after a sequence of N high taste shocks $\theta_{(2)}$, starting from v, is $\underline{w}$. This implies that $\underline{\varepsilon} < \pi(\theta_{(2)})^N$ is a uniform lower bound on the probability of being at $\underline{w}$ in N steps. Thus condition M is satisfied in Stokey et al. (1989), which concludes the proof. □

The immiseration result does not hold in the case where expected discounted utilities are constrained by (37), because the lower bound $\underline{w} > \underline{v}$ acts as a reflective (rather than absorbing) barrier (Lemma 2), creating a form of mean reversion that leads to a nondegenerate invariant distribution.

2.4.4 A Simple Example

In this section we address one remaining issue of our analysis. Proposition 4 showed that the allocation generated by the policy functions of our Bellman equation is the solution to the original problem (9) as long as it satisfies the limiting conditions (25) and (26). These conditions are trivially satisfied if the utility function is bounded, but many convenient functional forms assume an unbounded utility. In this section we analyze a simple example with unbounded utility in which we can easily verify conditions (25) and (26). This example also leads to a characterization of the solution to the Bellman equation "almost" in closed form.

We assume a logarithmic utility function $U(c) = \ln c$. Similar arguments can be applied to CRRA and CARA preferences. Note that $(\mathbf{u}, \mathbf{v}) \in \Gamma(v_0)$ if and only if

$(\mathbf{u}-(1-\beta)v_0, \mathbf{v}-v_0) \in \Gamma(0)$, where $\Gamma(\cdot)$ is defined in (16). We can thus rewrite the dual planner's problem (21) as

$$K(v_0) = \max_{(\mathbf{u},\mathbf{v})\in\Gamma(v_0)} \mathbb{E}_0\left[-\sum_{t=1}^{\infty}\beta^{t-1}\exp(u_t)\right]$$
$$= \max_{(\tilde{\mathbf{u}},\tilde{\mathbf{v}})\in\Gamma(0)} \mathbb{E}_0\left[-\sum_{t=1}^{\infty}\beta^{t-1}\exp(\tilde{u}_t+(1-\beta)v_0)\right] = \exp((1-\beta)v_0)K(0).$$

This implies that if $\{u_0(\theta), w_0(\theta)\}_{\theta\in\Theta}$ is the solution to the Bellman equation (23) for $v=0$, then $\{u_0(\theta)+(1-\beta)v, w_0(\theta)+v\}_{\theta\in\Theta}$ is the solution to (23) for any v. This property allows us to establish bounds on the left hand sides of (25) and (26). If we start with some initial v_0 and generate $(\mathbf{u},\mathbf{v})$ using the policy functions $(u_v(\theta), w_v(\theta))$ of the Bellman equation (23) as described in Section 2.3, we have

$$v_t(\theta^t) = w_{v_{t-1}(\theta^{t-1})}(\theta_t) = v_{t-1}\left(\theta^{t-1}\right) + w_0(\theta_t)$$
$$= w_{v_{t-2}(\theta^{t-2})}(\theta_{t-1}) + w_0(\theta_t) = v_{t-2}\left(\theta^{t-2}\right) + w_0(\theta_{t-1}) + w_0(\theta_t)$$
$$= \cdots = v_1\left(\theta^1\right) + w_0(\theta_2) + \cdots + w_0(\theta_t) = v_0 + \sum_{s=1}^{t} w_0(\theta_s).$$

Let $\underline{A} \equiv \min_\Theta\{w_0(\theta)\}$ and $\bar{A} \equiv \max_\Theta\{w_0(\theta)\}$, so that $\underline{A} \le w_0(\theta) \le \bar{A}$ for all $\theta \in \Theta$. Then $\beta^t(v_0+t\underline{A}) \le \beta^t v_t(\theta^t) \le \beta^t(v_0+t\bar{A})$ for all t, θ^t. Since $\lim_{t\to\infty}\beta^t t = 0$, this implies that $\lim_{t\to\infty}\beta^t v_t(\theta^t) = 0$ for all $\theta^\infty \in \Theta^\infty$, which implies both (25) and (26).

Since the value function K is homogeneous, it is easy to find it "almost" in closed form. Our arguments established that $K(v) = a\exp((1-\beta)v)$ for some $a < 0$. The parameter a can then be found as a fixed point of the equation

$$a = \max_{(\vec{u},\vec{w})\in B(0)} \sum_{\theta\in\Theta} \pi(\theta)[-\exp(u(\theta)) + \beta a \exp((1-\beta)w(\theta))].$$

The arguments used in this example can be extended to utility functions in the CRRA or CARA classes by observing that if $(\mathbf{u},\mathbf{v}) \in \Gamma(v_0)$ then $\left(\frac{1}{|v_0|}\mathbf{u}, \frac{1}{|v_0|}\mathbf{v}\right) \in \Gamma\left(\frac{v_0}{|v_0|}\right)$.

2.5 Autocorrelated Shocks

2.5.1 General Approach

We now address the case where the taste shocks θ follow a first-order Markov process. The goal of this section is to derive a recursive formulation for the planner's dual problem. We assume that the probabilities of the first period types $\theta_1 \in \Theta$ are given by $\pi\left(\theta_1 \middle| \theta_{(1)}\right)$, ie, as if the type realization in period 0 was the seed value

$\theta_{(1)}$. This assumption carries no loss of generality and simplifies the exposition. Fernandes and Phelan (2000) show how to write a recursive formulation of the planner's problem in this environment.

We define the analogue of the temporary incentive-compatibility constraint (10) in the case where shocks are first-order Markov as follows. For all $\theta^{t-1},\theta,\hat{\theta}$,

$$\theta U\left(c_t\left(\theta^{t-1},\theta\right)\right) + \beta \sum_{s=1}^{T-t} \sum_{\theta^{t+s}\in\Theta^{t+s}} \beta^{s-1}\pi_{t+s}\left(\theta^{t+s}\middle|\theta^{t-1},\theta\right)\theta_{t+s}U\left(c_{t+s}\left(\theta^{t-1},\theta,\theta_{t+1}^{t+s}\right)\right)$$

$$\geq \ \theta U\left(c_t\left(\theta^{t-1},\hat{\theta}\right)\right) + \beta \sum_{s=1}^{T-t} \sum_{\theta^{t+s}\in\Theta^{t+s}} \beta^{s-1}\pi_{t+s}\left(\theta^{t+s}\middle|\theta^{t-1},\theta\right)\theta_{t+s}U\left(c_{t+s}\left(\theta^{t-1},\hat{\theta},\theta_{t+1}^{t+s}\right)\right). \tag{38}$$

The one-shot-deviation result of Proposition 2 extends to the problem with persistent shocks:

Lemma 3 *Suppose that either T is finite, or U is bounded. Suppose moreover that the shocks θ follow a first-order Markov process. An allocation $\mathbf{c}$ satisfies (8) if and only if it satisfies (38).*

Proof Suppose that (8) is violated for some strategy $\boldsymbol{\sigma}'$ but (38) holds. If $\boldsymbol{\sigma}'$ involves misreporting at finitely many nodes, the arguments of Proposition 2 apply directly. If T is infinite and $\boldsymbol{\sigma}'$ recommends lying at infinitely many nodes, we have, by the previous result,

$$\sum_{t=1}^{\infty}\sum_{\theta^t\in\Theta^t}\beta^{t-1}\pi_t(\theta^t)\theta_t U(c_t(\theta^t)) \geq \sum_{t=1}^{\infty}\sum_{\theta^t\in\Theta^t}\beta^{t-1}\pi_t(\theta^t)\theta_t U(c_t(\sigma'^t(\theta^t)))$$

$$\ldots - \beta^T\left[\sum_{s=1}^{\infty}\sum_{\theta^{T+s}\in\Theta^{T+s}}\beta^{s-1}\pi_{T+s}\left(\theta^{T+s}\right)\theta_{T+s}\left(U\left(c_{T+s}\left(\sigma'^{T+s}\left(\theta^{T+s}\right)\right)\right) - U\left(c_{T+s}\left(\theta^{T+s}\right)\right)\right)\right].$$

Since the utility is bounded, the second line converges to zero as $T \to \infty$, which establishes that if $\mathbf{c}$ satisfies (38) then it satisfies (8). □

We follow Section 2.3 and redefine our maximization problem with respect to $u_t(\theta^t)$ rather than $c_t(\theta^t)$. We now emphasize the main differences that persistent shocks introduce. As in Section 2.3, we start by assuming that T is finite.

Finite-period economy

In this section we consider the case $T < \infty$. For any history $\theta^t \in \Theta^t$ and any $\theta' \in \Theta$, define $v_t(\theta^t|\theta')$ as

$$v_t(\theta^t|\theta') \equiv \sum_{s=1}^{T-t} \sum_{\theta^{t+s} \in \Theta^{t+s}} \beta^{s-1} \pi_{t+s}\left(\theta_{t+1}^{t+s}|\theta'\right)\theta_{t+s} u_{t+s}\left(\theta^t, \theta_{t+1}^{t+s}\right), \tag{39}$$

where $\left(\theta^t, \theta_{t+1}^{t+s}\right)$ denote the histories θ^{t+s} whose first t elements are θ^t. This allows us to write (38) as

$$\theta u_t\left(\theta^{t-1}, \theta\right) + \beta v_t\left(\theta^{t-1}, \theta|\theta\right) \geq \theta u_t\left(\theta^{t-1}, \hat{\theta}\right) + \beta v_t\left(\theta^{t-1}, \hat{\theta}|\theta\right), \forall \theta^{t-1}, \theta, \hat{\theta}. \tag{40}$$

Unlike the case of i.i.d. shocks, considered in Section 2.3.1, the continuation utility of an agent who reports θ^t depends not only on the history of reports but also on the true period-t shock θ'_t of the agent. The economic intuition for this result is that when shocks are autocorrelated, the realization of the shock θ'_t is informative about the realization of future shocks from period $t+1$ onward. Repeated substitution allows us to rewrite v_t as

$$v_t(\theta^t|\theta') = \sum_{\theta \in \Theta} \pi(\theta|\theta')[\theta u_{t+1}(\theta^t, \theta) + \beta v_{t+1}(\theta^t, \theta|\theta)], \forall \theta^t, \theta', \tag{41}$$

with the convention that $v_T\left(\theta^T|\theta'\right) = 0$ if T is finite. The initial utility v_0 is given by

$$v_0 = \sum_{\theta \in \Theta} \pi\left(\theta|\theta_{(1)}\right)[\theta u_1(\theta) + \beta v_1(\theta|\theta)]. \tag{42}$$

Let $\mathbf{v} = \left\{v_t\left(\theta^t|\theta'_t\right)\right\}_{t \geq 1, \theta^t \in \Theta^t, \theta'_t \in \Theta}$. The set $\Gamma(v_0)$ is now defined as the set of allocations $(\mathbf{u}, \mathbf{v})$ that satisfy (40)–(42). The direct extension of the arguments of Section 2.3 implies that the optimal incentive-compatible allocation (ie, the solution to the primal problem (9)) exists and is a solution to the dual maximization problem

$$\widetilde{K}_0(v_0) \equiv \max_{(\mathbf{u}, \mathbf{v}) \in \Gamma(v_0)} \left[-\sum_{t=1}^{T} \sum_{\theta^t \in \Theta^t} \beta^{t-1} \pi_t\left(\theta^t|\theta_{(1)}\right) C(u_t(\theta^t))\right]. \tag{43}$$

This problem can be written recursively following the same ideas as we used to obtain the Bellman equation (18), with two differences: (i) the state space is larger when the shocks are autocorrelated, and (ii) the space of feasible values for the state variables is now more difficult to characterize. We show both of these differences using backward induction arguments.

The need for the larger state space can be seen already from the incentive constraints. Each history of reports θ^t has an associated $|\Theta|$-dimensional *vector* of "promised utilities" $v_t(\theta^t|\cdot) = \left\{v_t\left(\theta^t|\theta_{(j)}\right)\right\}_{j=1}^{|\Theta|}$, where for each j, $v_t\left(\theta^t|\theta_{(j)}\right)$ is the promised utility allocated to the agent who reported history θ^t and whose true type in period t was actually $\theta_{(j)}$. Moreover, the expectation over the future realizations of shocks in period $t+1$ depends on the period-t realized shock θ'_t. Therefore the state space has dimensionality $|\Theta| + 1$. We now describe the recursive construction of the value function $K_{t-1}\left(v\left(\theta_{(1)}\right), \ldots, v\left(\theta_{(|\Theta|)}\right), \theta_-\right)$

and its domain $\mathcal{V}_{t-1} \times \Theta$. Let $\mathcal{V}_{T-1}$ be the space of all vectors $v(\cdot) \in \mathbb{R}^{|\Theta|}$ with the property that there exists some $u \in \mathbb{U}$ such that $v(\theta_{(i)}) = u \sum_{\theta \in \Theta} \pi(\theta | \theta_{(i)}) \theta$ for all $i \in \Theta$. Let $K_{T-1}(v(\cdot), \theta) = -C(u)$ for all such $v(\cdot) \in \mathcal{V}_{T-1}$. This definition simply captures the fact that in the last period the principal cannot provide any insurance against the period-T shocks (by incentive compatibility (40)), and K_{T-1} is then (minus) the cost of the feasible promises that the principal can make in period $T - 1$. For $0 \leq t \leq T-2$ define K_t recursively as

$$K_t(v(\cdot), \theta_-) = \max_{\{u(\theta), w(\theta|\cdot)\}_{\theta \in \Theta}} \sum_{\theta \in \Theta} \pi(\theta|\theta_-)[-C(u(\theta)) + \beta K_{t+1}(w(\theta|\cdot), \theta)] \tag{44}$$

subject to the promise-keeping constraints

$$v(\theta_{(j)}) = \sum_{\theta \in \Theta} \pi(\theta | \theta_{(j)})[\theta u(\theta) + \beta w(\theta|\theta)], \forall j \in \{1, \ldots, |\Theta|\}, \tag{45}$$

the incentive-compatibility constraints

$$\theta u(\theta) + \beta w(\theta|\theta) \geq \theta u(\hat{\theta}) + \beta w(\hat{\theta}|\theta), \forall \theta, \hat{\theta} \in \Theta, \tag{46}$$

and

$$u(\theta) \in \mathbb{U}, \quad w(\theta|\cdot) \in \mathcal{V}_{t+1}, \quad \forall \theta \in \Theta. \tag{47}$$

The domain of K_t is $\mathcal{V}_t \times \Theta$, where $\mathcal{V}_t$ is defined as the set of all $v(\cdot) \in \mathbb{R}^{|\Theta|}$ with the property that there exist $\{u(\theta), w(\theta|\cdot)\}_{\theta \in \Theta}$ such that the constraints (45), (46), and (47) are satisfied.

So far we defined K_t from purely mathematical considerations by observing that the solution to the maximization problem (43) after any history θ^t could be found independently of any other history $\hat{\theta}^t$, as long as we keep track of the vector $v(\cdot)$ and the realization of the period-t shock θ'_t. It is useful to describe the economic intuition behind these equations. Eq. (46) is simply the incentive constraint, familiar from Section 2.3. Eq. (45) for $\theta_{(j)} = \theta_-$ summarizes the expected utility that an agent with period-$(t-1)$ shock θ_- receives in period t. This equation is the analogue of the promise-keeping constraint (19) in the i.i.d. case. Eq. (45) for $\theta_{(j)} \neq \theta_-$ are auxiliary "threat-keeping" constraints, which allow us to keep track of the incentives provided in the previous period. Since allocations are incentive compatible, no agent misrepresents his type along the equilibrium path and hence no agent actually obtains utility $v(\theta_{(j)})$ for $\theta_{(j)} \neq \theta_-$. One can think of those $v(\theta_{(j)})$ as threats that the principal chooses in period $t-1$ to ensure that agents do not misrepresent their type. Eq. (45) in period t ensures that the principal's subsequent choices are consistent with that threat. The principal chooses a common allocation for all the agents that report θ_-. This common allocation simultaneously delivers utility $v(\theta_-)$ to the agents with true type θ_- (ie, when expected values are computed using the probabilities $\pi(\theta|\theta_-)$), and $v(\theta_{(j)})$ to the agents with true types $\theta_{(j)}$ (ie, when expected values are computed using the probabilities $\pi(\theta|\theta_{(j)})$) for each $j \in \{1, \ldots, |\Theta|\}$.

The relationship between the function $K_0(v(\cdot),\theta_-)$ defined in (44) and the function $\widetilde{K}_0(v_0)$ defined in (43) is as follows. Observe that there are no auxiliary threat-keeping constraints in the set $\Gamma(v_0)$. It is mathematically equivalent to saying that those constraints are slack. Thus, given our assumption that shocks in period 1 are drawn from $\pi\left(\cdot|\theta_{(1)}\right)$, the relationship between $K_0(v(\cdot),\theta_-)$ and $\widetilde{K}_0(v_0)$ is simply

$$\widetilde{K}_0(v_0) = \max_{v(\cdot)\in\mathcal{V}_0,\, v\left(\theta_{(1)}\right)=v_0} K_0\left(v(\cdot),\theta_{(1)}\right). \tag{48}$$

This gives a simple way to find the solution $(\mathbf{u}^*,\mathbf{v}^*)$ for the primal problem (9). The value of this problem should be such that the feasibility constraint holds with equality, which can be found as a solution to $\widetilde{K}_0(v_0) = -\dfrac{1-\beta^T}{1-\beta}e$. Then from the maximization problem (48) we generate the vector $v_0(\cdot)$. Finally, we use the policy functions to the Bellman equation (44) to generate the solution $(\mathbf{u}^*,\mathbf{v}^*)$, analogous to the i.i.d. case.

Infinite-period economy

We now turn to the recursive formulation in the infinite-period economy, $T=\infty$. Assume for simplicity that the utility function is bounded, ie, $\mathbb{U}=[\underline{u},\bar{u}]$. Let $\mathcal{V}$ be the set of promised utility vectors $v(\cdot)$ for which there exists an allocation $\mathbf{u}$ such that

$$v\left(\theta_{(j)}\right) = \sum_{t=1}^{\infty}\sum_{\theta^t\in\Theta^t}\beta^{t-1}\pi_t\left(\theta^t\middle|\theta_{(j)}\right)\theta_t u_t(\theta^t),\ \forall j\in\{1,\ldots,|\Theta|\}, \tag{49}$$

and for all $t\geq 1$, for all $\theta^{t-1},\theta,\hat{\theta}$,

$$\begin{aligned}&\theta u_t\left(\theta^{t-1},\theta\right)+\beta\left\{\sum_{s=1}^{\infty}\sum_{\theta^{t+s}\in\Theta^{t+s}}\beta^{s-1}\pi_{t+s}\left(\theta^{t+s}\middle|\theta^{t-1},\theta\right)\theta_{t+s}u_{t+s}\left(\theta^{t-1},\theta,\theta_{t+1}^{t+s}\right)\right\}\\ \geq\ &\theta u_t\left(\theta^{t-1},\hat{\theta}\right)+\beta\left\{\sum_{s=1}^{\infty}\sum_{\theta^{t+s}\in\Theta^{t+s}}\beta^{s-1}\pi_{t+s}\left(\theta^{t+s}\middle|\theta^{t-1},\theta\right)\theta_{t+s}u_{t+s}\left(\theta^{t-1},\hat{\theta},\theta_{t+1}^{t+s}\right)\right\}.\end{aligned} \tag{50}$$

For any $\theta_-\in\Theta$ and $v(\cdot)\in\mathcal{V}$, the Bellman equation writes

$$K(v(\cdot),\theta_-) = \max_{\{u(\theta),w(\theta|\cdot)\}_{\theta\in\Theta}}\sum_{\theta\in\Theta}\pi(\theta|\theta_-)[-C(u(\theta))+\beta K(w(\theta|\cdot),\theta)] \tag{51}$$

subject to (45), (46), and $u(\theta)\in\mathbb{U}$, $w(\theta|\cdot)\in\mathcal{V}$ for all θ.

This Bellman equation is a direct extension of the Bellman equation (23) in the i.i.d. case. The need to keep track of a larger number of state variables in the case of general Markov shocks follows from our discussion in the finite period economy. One additional consideration that (51) introduces is that it is defined over a set $\mathcal{V}$, which needs to be found. The work of Abreu et al. (1990) provides a method of finding this set and characterizing its properties.

Proposition 8 *The set $\mathcal{V}$ is nonempty, compact, and convex. It is the largest bounded fixed point of the operator $\mathscr{A}$ defined for an arbitrary compact set $\tilde{\mathcal{V}} \subset \mathbb{R}^{|\Theta|}$ as*

$$\mathscr{A}\tilde{\mathcal{V}} = \left\{ v(\cdot) \text{ s.t. } \exists \{u(\theta), w(\theta|\cdot)\}_{\theta \in \Theta} : (45), (46) \text{ hold and } (u(\theta),\ w(\theta|\cdot)) \in \mathbb{U} \times \tilde{\mathcal{V}},\ \forall \theta \right\}.$$

It is the limit of the monotonically decreasing sequence of compact sets $\{\mathcal{V}_n\}_{n=0,1,\ldots}$ defined as $\mathcal{V}_0 = \left[\frac{\theta_{(1)}}{1-\beta}\underline{u}, \frac{\theta_{(|\Theta|)}}{1-\beta}\bar{u}\right]^{|\Theta|}$ and $\mathcal{V}_n = \mathscr{A}\mathcal{V}_{n-1}$ for $n \geq 1$, so that $\mathcal{V} = \lim_{n\to\infty}\mathcal{V}_n = \bigcap_{n=1}^{\infty}\mathcal{V}_n$.

Proof The set $\mathcal{V}$ is nonempty because any allocation that is independent of the report is incentive compatible. $\mathcal{V}$ is convex since $v_{\mathbf{u}}\left(\theta_{(j)}\right)$ is affine in $\mathbf{u}$ for all $j \in \{1, \ldots, |\Theta|\}$, where $v_{\mathbf{u}}\left(\theta_{(j)}\right)$ is defined by the right hand side of (49). The construction of $\mathcal{V}$ as the largest compact fixed point of the operator $\mathscr{A}$ follows from the results of Abreu et al. (1990). Here we give a simple proof that $\mathcal{V}$ is compact and is a fixed point of $\mathscr{A}$.

Let $\mathcal{U}$ denote the space of allocations $\mathbf{u} = \{u_t(\theta^t)\}_{t\geq 1, \theta^t \in \Theta^t}$, with $u_t(\theta^t) \in [\underline{u}, \bar{u}]$ for all $t \geq 1$, $\theta^t \in \Theta^t$. Since $|\Theta^t| < \infty$ for all $t \geq 1$, $\mathcal{U}$ is the countable product of the compact metric spaces $[\underline{u}, \bar{u}]$. Embedding $\mathcal{U}$ with the product topology, we obtain that $\mathcal{U}$ is a compact metric space (the compactness is a standard result that follows from a diagonalization argument). A sequence $\mathbf{u}^{(n)}$ in $\mathcal{U}$ converges as $n \to \infty$ if and only if all of its projections $u_t^{(n)}(\theta^t)$ converge in $[\underline{u}, \bar{u}]$ as $n \to \infty$.

We now show that $\mathcal{V}$ is compact. Since the utility function is bounded, $v_{\mathbf{u}}$ is bounded and hence $\mathcal{V}$ is bounded. To prove that $\mathcal{V}$ is closed, let $\left\{\vec{v}^{(n)}\right\}_{n=1}^{\infty}$ be a Cauchy sequence in $\mathcal{V}$, and let $\vec{v}^{(\infty)} = \left\{v^{(\infty)}\left(\theta_{(j)}\right)\right\}_{j=1,\ldots,J}$ its limit. Let $\left\{\mathbf{u}^{(n)}\right\}_{n=1}^{\infty}$ be a sequence of allocations such that $v^{(n)}\left(\theta_{(j)}\right) = v_{\mathbf{u}^{(n)}}\left(\theta_{(j)}\right)$ for all $j \in \{1, \ldots, |\Theta|\}$ and all $n \geq 1$. Since $\mathcal{U}$ is compact, $\left\{\mathbf{u}^{(n)}\right\}_{n=1}^{\infty}$ contains a convergent subsequence $\left\{\mathbf{u}^{(\varphi(n))}\right\}_{n=1}^{\infty}$; denote by $\mathbf{u}^{(\infty)}$ its limit. We have $\mathbf{u}^{(\infty)} \in \mathcal{U}$. Since $v_{\mathbf{u}}\left(\theta_{(j)}\right)$ is continuous in $\mathbf{u}$ we get, for all $j \in \{1, \ldots, |\Theta|\}$,

$$v^{(\infty)}\left(\theta_{(j)}\right) = \lim_{n\to\infty} v^{(\varphi(n))}\left(\theta_{(j)}\right) = \lim_{n\to\infty} v_{\mathbf{u}^{(\varphi(n))}}\left(\theta_{(j)}\right) = v_{\mathbf{u}^{(\infty)}}\left(\theta_{(j)}\right).$$

Finally, since $\mathbf{u}^{(n)}$ satisfies the incentive constraints (50), by continuity we obtain that $\mathbf{u}^{(\infty)}$ satisfies (50) as well. Thus $\vec{v}^{(\infty)} \in \mathcal{V}$ and hence $\mathcal{V}$ is closed. Since $\mathcal{V} \subset \mathbb{R}^{|\Theta|}$, we obtain that $\mathcal{V}$ is compact.

Next we show that $\mathcal{V}$ is a fixed point of $\mathscr{A}$, that is $\mathscr{A}\mathcal{V} = \mathcal{V}$. First let $\vec{v} \in \mathcal{V}$. There exists $\mathbf{u} = \{u_t(\theta^t)\}_{t,\theta^t} \in \mathcal{U}$ that satisfies the incentive constraints (50) and delivers $v\left(\theta_{(j)}\right) = v_{\mathbf{u}}\left(\theta_{(j)}\right)$ for all $j \in \{1, \ldots, |\Theta|\}$. Define the allocation rule $\{u(\theta), w(\theta|\cdot)\}_{\theta \in \Theta}$

by $u(\theta) = u_1(\theta)$ and $w\left(\theta\middle|\theta_{(j')}\right) = v_{\mathbf{u}_2^\infty(\theta)}\left(\theta_{(j')}\right)$, where $\mathbf{u}_2^\infty(\theta)$ is the continuation of the allocation $\mathbf{u}$ from period 2 onward given the history $\theta^1 = \theta$. We have $w(\theta|\cdot) \in \mathcal{V}$ for all $\theta \in \Theta$ because the allocation $\mathbf{u}_2^\infty(\theta)$ satisfies the incentive-compatibility condition (50) after all histories. Moreover, we have

$$\begin{aligned} \theta u(\theta) + \beta w(\theta|\theta) &= \theta u_1(\theta) + \beta v_{\mathbf{u}_2^\infty(\theta)}(\theta) \\ \geq \theta u_1\left(\hat{\theta}\right) + \beta v_{\mathbf{u}_2^\infty\left(\hat{\theta}\right)}(\theta) &= \theta u\left(\hat{\theta}\right) + \beta w\left(\hat{\theta}|\theta\right), \end{aligned}$$

where the inequality follows from (50). Hence $\{u(\theta), w(\theta|\cdot)\}_{\theta\in\Theta}$ satisfies (46). Finally, by construction $\{u(\theta), w(\theta|\cdot)\}_{\theta\in\Theta}$ satisfies (45). Thus, $\vec{v} \in \mathscr{A}\mathcal{V}$ and hence $\mathcal{V} \subset \mathscr{A}\mathcal{V}$. For the converse, suppose that $\vec{v} \in \mathscr{A}\mathcal{V}$. Then there exists some allocation rule $\{u(\theta), w(\theta|\cdot)\}_{\theta\in\Theta}$ such that the promise-keeping constraint (45) and the incentive constraints (46) hold and $w(\theta|\cdot) \in \mathcal{V}$ for all θ. Define an allocation $\mathbf{u}$ as follows. Let $u_1(\theta) = u(\theta)$. For each $\theta \in \Theta$, since $w(\theta|\cdot) \in \mathcal{V}$ there exists some allocation $\tilde{\mathbf{u}}\ (\theta)$ such that $v_{\tilde{\mathbf{u}}(\theta)}\left(\theta_{(j)}\right) = w\left(\theta\middle|\theta_{(j)}\right)$ for all $j \in \{1, \ldots, |\Theta|\}$. Define $\mathbf{u}_2^\infty(\theta) = \tilde{\mathbf{u}}\ (\theta)$. The allocation $\mathbf{u}$ constructed in this way is in $\mathcal{U}$, satisfies the incentive constraints (50), and delivers $v_{\mathbf{u}}\left(\theta_{(j)}\right) = v\left(\theta_{(j)}\right)$. Thus, $\vec{v} \in \mathcal{V}$ and hence $\mathscr{A}\mathcal{V} \subset \mathcal{V}$. □

2.5.2 Continuum of Shocks and the First-Order Approach

The previous section provides a general way to characterize recursively the solution to the optimal insurance problem when shocks are Markovian. One practical difficulty in using the Bellman equation (51) in applications is that the dimensionality of the state space grows with the number of shocks. As the number of shocks becomes large, solving problem (51) becomes intractable. To keep the problem manageable, it is useful to have a method that keeps the number of state variables small.

One approach is to guess that only some of the incentive constraints (38) bind at the optimum. In this case all the nonbinding constraints can be dropped, which also eliminates the need to keep track of the corresponding state variables. The natural candidate for binding constraints are the local constraints that ensure that a type θ does not want to mimic the types closest to his. In this section we describe how to construct this relaxed problem and provide sufficient conditions that can be verified ex post to make sure that the dropped incentive constraints are satisfied.[1]

This analysis can be done with a discrete number of shocks, but it becomes particularly simple if instead we allow for a continuum of shocks. In this case applying the envelope theorem to the incentive-compatibility condition gives a simple and

[1] It is important to keep in mind that there is a large class of incentive problems with persistent shocks in which nonlocal incentive constraints bind (see, eg, Battaglini and Lamba, 2015) and thus the relaxed problem may not satisfy the sufficient conditions.

tractable way to derive the Bellman equation. This problem has been analyzed by Kapička (2013) and Pavan et al. (2014); here we follow the exposition of the former.

Let the taste shocks θ_t in each period belong to an interval $\Theta = \left(\underline{\theta}, \overline{\theta}\right) \subset \mathbb{R}_{+}^{*}$, with $\overline{\theta} < \infty$. We assume that the stochastic process for the shocks θ_t is Markov with continuous density $\pi(\theta_t|\theta_{t-1})$. We use $\pi_s(\cdot|\theta_t)$ to denote the p.d.f. of histories θ_{t+1}^{t+s} given that the shock θ_t occurred in period t, that is,

$$\pi_s\left(\theta_{t+1}^{t+s}|\theta_t\right) = \pi(\theta_{t+s}|\theta_{t+s-1}) \times \cdots \times \pi(\theta_{t+1}|\theta_t).$$

Assume as in the previous section that these probabilities are generated from the seed value $\theta_0 = \theta_{(1)}$. We make the following assumptions:

Assumption 4 Assume that the density $\pi(\theta|\cdot)$ is uniformly Lipschitz continuous for all θ, and that the derivatives $\hat{\pi}(\theta|\theta_-) \equiv \dfrac{\partial \pi(\theta|\theta_-)}{\partial \theta_-}$ exist and are uniformly bounded.

These assumptions can be substantially relaxed (see Kapička, 2013; Pavan et al., 2014 for more general treatments of the problem), but they considerably simplify our analysis. To simplify the integrability conditions, we further assume in this section that the utility function is bounded.

Having a continuum of shocks does not change the arguments leading to the recursive characterization of the constraints (40), (41), with the only difference that the sum over a finite number of shock realizations in Eq. (41) is now replaced by an integral. Constraints (41) and (40) can be written as

$$v_t\left(\theta^{t-1}, \hat{\theta}_t|\theta_t\right) = \int_{\Theta}\left[\theta' u_{t+1}\left(\theta^{t-1}, \hat{\theta}_t, \theta'\right) + \beta v_{t+1}\left(\theta^{t-1}, \hat{\theta}_t, \theta'|\theta'\right)\right]\pi(\theta'|\theta_t)d\theta', \forall \theta^t, \hat{\theta}_t, \tag{52}$$

and

$$\theta_t u_t\left(\theta^{t-1}, \theta_t\right) + \beta v_t\left(\theta^{t-1}, \theta_t|\theta_t\right) = \max_{\hat{\theta}\in\Theta}\left\{\theta_t u_t\left(\theta^{t-1}, \hat{\theta}\right) + \beta v_t\left(\theta^{t-1}, \hat{\theta}|\theta_t\right)\right\}, \forall \theta^{t-1}, \theta_t. \tag{53}$$

Lemma 4 *Suppose Assumption 4 is satisfied and the utility function is bounded. Then the function $v_t\left(\theta^{t-1}, \theta_t|\cdot\right)$ is differentiable with respect to the realized period-t type θ for each history of reports $\theta^t = \left(\theta^{t-1}, \theta_t\right)$ and its derivative evaluated at θ_t is given by*

$$\hat{v}_t(\theta^t) = \int_{\Theta}[\theta' u_{t+1}(\theta^t, \theta') + \beta v_{t+1}(\theta^t, \theta'|\theta')]\hat{\pi}(\theta'|\theta_t)d\theta'. \tag{54}$$

Moreover, if an allocation is incentive compatible, then for all $t \geq 1$, $\theta^t \in \Theta^t$,

$$\theta_t u_t\left(\theta^{t-1},\theta_t\right) + \beta v_t\left(\theta^{t-1},\theta_t|\theta_t\right) = \int_{\underline{\theta}}^{\theta_t} \left\{u_t\left(\theta^{t-1},\theta\right) + \beta \hat{v}_t\left(\theta^{t-1},\theta\right)\right\} d\theta + \nu\left(\theta^{t-1}\right), \tag{55}$$

where $\nu\left(\theta^{t-1}\right) = \lim_{\theta \to \underline{\theta}}\left\{\theta u_t\left(\theta^{t-1},\theta\right) + \beta v_t\left(\theta^{t-1},\theta|\theta\right)\right\}$.

Proof Let $S_{t+1}(\theta^t,\theta') = \theta' u_{t+1}(\theta^t,\theta') + \beta v_{t+1}(\theta^t,\theta'|\theta')$. Then

$$\begin{aligned}\frac{v_t\left(\theta^{t-1},\theta_t|\theta+\Delta\theta\right) - v_t\left(\theta^{t-1},\theta_t|\theta\right)}{\Delta\theta} &= \int_\Theta S_{t+1}(\theta^t,\theta')\frac{\pi(\theta'|\theta+\Delta\theta) - \pi(\theta'|\theta)}{\Delta\theta}d\theta' \\ &\xrightarrow[\Delta\theta \to 0]{} \int_\Theta S_{t+1}(\theta^t,\theta')\hat{\pi}(\theta'|\theta)d\theta',\end{aligned}$$

where the last step follows from the Dominated Convergence Theorem, noting that $S_{t+1}(\theta^t,\theta')$ and $\left|\frac{\pi(\theta'|\theta+\Delta\theta) - \pi(\theta'|\theta)}{\Delta\theta}\right|$ are bounded by the uniform Lipschitz continuity of $\pi(\theta'|\cdot)$. Hence $v_t\left(\theta^{t-1},\theta_t|\cdot\right)$ is differentiable, and it is Lipschitz continuous on $\left[\eta,\overline{\theta}\right)$ for all $\eta > \underline{\theta}$ since $\hat{\pi}(\theta'|\theta)$ is uniformly bounded.

Let $\hat{S}_t\left(\theta^{t-1},\hat{\theta}_t|\theta_t\right) = \theta_t u_t\left(\theta^{t-1},\hat{\theta}_t\right) + \beta v_t\left(\theta^{t-1},\hat{\theta}_t|\theta_t\right)$. Then $\hat{S}_t$ is differentiable in θ_t on $\left(\eta,\overline{\theta}\right)$ (denote by $\hat{S}_{\theta,t}$, its derivative) and Lipschitz in θ_t on $\left[\eta,\overline{\theta}\right)$. Hence it is absolutely continuous and has a bounded derivative with respect to θ_t on $\left(\eta,\overline{\theta}\right)$. Since the allocation is incentive compatible, $\hat{S}_t\left(\theta^{t-1},\hat{\theta}_t|\theta_t\right)$ is maximized at $\hat{\theta}_t = \theta_t$. By Theorem 2 in Milgrom and Segal (2002), $\hat{S}_t\left(\theta^{t-1},\theta_t|\theta_t\right)$ can be represented as an integral of its derivative:

$$\begin{aligned}\hat{S}_t\left(\theta^{t-1},\theta_t|\theta_t\right) &= \int_\eta^{\theta_t} \hat{S}_{\theta,t}\left(\theta^{t-1},\theta,\theta\right)d\theta + \hat{S}_t\left(\theta^{t-1},\eta|\eta\right) \\ &= \int_\eta^{\theta_t}\left\{u_t\left(\theta^{t-1},\theta\right) + \beta\hat{v}_t\left(\theta^{t-1},\theta|\theta\right)\right\}d\theta + \hat{S}_t\left(\theta^{t-1},\eta|\eta\right).\end{aligned}$$

Take the limit as $\eta \to \underline{\theta}$ to get expression (55). □

We can now define a relaxed problem by replacing the temporarily incentive-compatibility constraints (53) by the envelope condition (55) for all histories. This substantially simplifies the analysis, as the latter constraint only depends on the lifetime utility and marginal lifetime utility of the truthteller, rather than the continuation utility of all possible types as in Section 2.5. In the recursive formulation of the planner's problem, the choice variables are the current utility $u(\theta)$, the continuation utility of the truthtelling agent $w(\theta)$, and the marginal change in the continuation utility of the truthtelling agent

$\hat{w}(\theta)$. The state variables are the reported taste shock realization θ_- in the previous period, the promised utility v of an agent who truthfully announced θ_- last period, and the marginal promised utility $\hat{v}$ of an agent who truthfully announced θ_- last period.

We now show the recursive formulation of the relaxed problem in an infinite-period economy. Let $\hat{\mathcal{V}}(\theta_-)$ denote the set of lifetime utility and marginal lifetime utility pairs $(v,\hat{v}) \in \mathbb{R}^2$ for which there exist values $\{u(\theta), w(\theta), \hat{w}(\theta)\}_{\theta \in \Theta}$ such that the following conditions hold:

(i) the envelope condition:

$$\theta u(\theta) + \beta w(\theta) = \int_{\underline{\theta}}^{\theta} \{u(\theta') + \beta \hat{w}(\theta')\} d\theta' + \lim_{\theta' \to \underline{\theta}} \{\theta' u(\theta') + \beta w(\theta')\}, \forall \theta \in \Theta, \tag{56}$$

(ii) the promise-keeping constraint:

$$v = \int_{\Theta} \{\theta' u(\theta') + \beta w(\theta')\} \pi(\theta' | \theta_-) d\theta', \tag{57}$$

(iii) the marginal promise-keeping constraint:

$$\hat{v} = \int_{\Theta} \{\theta' u(\theta') + \beta w(\theta')\} \hat{\pi}(\theta' | \theta_-) d\theta', \tag{58}$$

(iv) $(w(\theta), \hat{w}(\theta)) \in \hat{\mathcal{V}}(\theta)$ for all $\theta \in \Theta$.

Note that in general $\hat{\mathcal{V}}(\theta)$ depends on the realized value of θ. It can be characterized along the lines of Proposition 8.

For any $\theta_- \in \Theta$ and pair $(v,\hat{v}) \in \hat{\mathcal{V}}(\theta_-)$, the Bellman equation writes

$$K(v,\hat{v},\theta_-) = \sup_{(\vec{u},\vec{w},\vec{\hat{w}})} \int_{\Theta} \{-C(u(\theta)) + \beta K(w(\theta), \hat{w}(\theta), \theta)\} \pi(\theta | \theta_-) d\theta \tag{59}$$

subject to (56)–(58) and $u(\theta) \in \mathbb{U}$, $(w(\theta), \hat{w}(\theta)) \in \hat{\mathcal{V}}(\theta)$ for all $\theta \in \Theta$.

We finally discuss when the relaxed problem gives the solution to the original problem. The envelope condition (55) is necessary but not sufficient for an allocation to be temporarily incentive compatible. A sufficient condition is given in Proposition 9:

Proposition 9 *Suppose that an allocation* $\mathbf{u}$ *satisfies the envelope condition (55) and, in addition,*

$$u_t(\theta^{t-1}, \hat{\theta}_t) + \beta \hat{v}_t(\theta^{t-1}, \hat{\theta}_t | \theta_t) \tag{60}$$

is increasing in $\hat{\theta}_t$ *for all* t, θ^{t-1} *and almost all* θ_t, *where* $\hat{v}_t(\theta^{t-1}, \hat{\theta}_t | \theta_t) \equiv \frac{\partial}{\partial \theta} v_t(\theta^{t-1}, \hat{\theta}_t | \theta_t)$. *Then* $\mathbf{u}$ *is incentive compatible.*

Proof Fix t, θ^{t-1}, and let $S_t(\theta^{t-1}, \theta_t) \equiv \hat{S}_t(\theta^{t-1}, \theta_t | \theta_t)$. An allocation is temporarily incentive compatible if $S_t(\theta^{t-1}, \theta_t) \geq \hat{S}_t(\theta^{t-1}, \hat{\theta}_t | \theta_t)$ for all $\theta^{t-1}, \theta_t, \hat{\theta}_t$. Eq. (55) shows that $S_t(\theta^{t-1}, \cdot)$ is differentiable for almost all $\theta \in \Theta$ with

$$\frac{\partial}{\partial \theta} S_t(\theta^{t-1}, \theta) = u_t(\theta^{t-1}, \theta) + \beta \hat{v}_t(\theta^{t-1}, \theta).$$

We thus have

$$\begin{aligned} & S_t(\theta^{t-1}, \theta_t) - S_t(\theta^{t-1}, \hat{\theta}_t) \\ & = \int_{\hat{\theta}_t}^{\theta_t} \frac{\partial}{\partial \theta'} S_t(\theta^{t-1}, \theta') \, d\theta' = \int_{\hat{\theta}_t}^{\theta_t} \left\{ u_t(\theta^{t-1}, \theta') + \beta \hat{v}_t(\theta^{t-1}, \theta' | \theta') \right\} d\theta' \\ & \geq \int_{\hat{\theta}_t}^{\theta_t} \left\{ u_t(\theta^{t-1}, \hat{\theta}_t) + \beta \hat{v}_t(\theta^{t-1}, \hat{\theta}_t | \theta') \right\} d\theta' = \hat{S}_t(\theta^{t-1}, \hat{\theta}_t | \theta_t) - S_t(\theta^{t-1}, \hat{\theta}_t), \end{aligned}$$

where the inequality follows from the monotonicity of (60), and the last equality follows from the differentiability of $\hat{S}_t(\theta^{t-1}, \hat{\theta}_t | \theta_t)$, with $\frac{\partial}{\partial \theta} \hat{S}_t(\theta^{t-1}, \hat{\theta}_t | \theta) = u_t(\theta^{t-1}, \hat{\theta}_t) + \beta \hat{v}_t(\theta^{t-1}, \hat{\theta}_t | \theta)$. We obtain that $\mathbf{u}$ is temporarily incentive compatible. □

If the shocks are i.i.d., the second term in expression (60) drops out and the proposition is equivalent to a simple requirement that $u_t(\theta^t)$ is increasing in θ_t, and one can show that this requirement is necessary as well. In the static setting, u satisfies the Spence–Mirrlees condition and this sufficient condition reduces to the familiar necessary and sufficient condition that allocations are monotonic (see Myerson, 1981). Unfortunately, in the dynamic model with persistent shocks, the monotonicity condition on (60) is not necessary, and moreover there is no one-to-one mapping between marginal lifetime utilities and allocations. Moreover, in practice condition (60) is difficult to verify directly, and we have to either try to derive weaker sufficient conditions, or check ex post (possibly numerically) in specific applications whether the solution to the relaxed problem is indeed an optimal allocation.

2.6 Hidden Storage

We now suppose that agents have access to a storage technology, a problem analyzed by Allen (1985) and Cole and Kocherlakota (2001).[m] The model is the same as in Sections 2.3 and 2.4 (with i.i.d. and discrete types), except that individuals can now store nonnegative amounts of goods at rate R. The planner cannot observe these private savings.

[m] See also Werning (2002), Golosov and Tsyvinski (2007), Farhi et al. (2009), and Ales and Maziero (2009).

The planner is still able to both borrow and lend at the same rate R as the agents.[n] We show that allowing for hidden private storage dramatically changes the optimal social insurance contract: in this environment, no social insurance can be provided.

To understand the argument, suppose first (following Allen, 1985) that agents can both borrow *and* lend at rate R. In this case, agents can always perfectly smooth across time their consumption. Hence they always report the shocks that yield the highest net present value of transfers, regardless of their true history. Consequently, incentive compatibility requires that the planner gives all individuals the same present value of transfers, which must then be equal to the present value of the endowment, $e + e/R$. Therefore, the planner simply gives the economy's endowment to the agents, who self-insure from then on. In particular, there is no transfer of resources *across* households, ie, no risk sharing is possible.

Now suppose (following Cole and Kocherlakota, 2001) that the agent can only privately save, but *not* borrow, at the interest rate R. Assume for simplicity that the horizon lasts two periods (see Cole and Kocherlakota, 2001 for a generalization to $T \leq \infty$ periods).[o] We still denote by $\mathbf{c} = \{c_1(\theta_1), c_2(\theta_1, \theta_2)\}$ the agent's consumption, but now the transfers from the planner to the agent may be different and are denoted by $\boldsymbol{\tau} = \{\tau_1(\theta_1), \tau_2(\theta_1, \theta_2)\}$. Denote by $k(\theta_1)$ the agent's private storage, and by K the public saving or borrowing. An efficient allocation is defined as a tuple $\{\mathbf{c}, \boldsymbol{\tau}, k, K\}$ that solves:

$$\max_{\{\mathbf{c},\boldsymbol{\tau},k,K\}} \sum_{\theta_1\in\Theta} \pi(\theta_1)\left\{\theta_1 U(c_1(\theta_1)) + \sum_{\theta_2\in\Theta} \beta\pi(\theta_2)\theta_2 U(c_2(\theta_1,\theta_2))\right\} \tag{61}$$

subject to the planner's feasibility constraints: $\forall \theta_1, \theta_2 \in \Theta$,

$$\begin{aligned} &\sum_{\theta_1\in\Theta} \pi(\theta_1)\tau_1(\theta_1) + K = e, \\ &\sum_{(\theta_1,\theta_2)\in\Theta^2} \pi(\theta_1)\pi(\theta_2)\tau_2(\theta_1,\theta_2) = e + RK, \end{aligned} \tag{62}$$

the agent's resource constraints: $\forall \theta_1, \theta_2 \in \Theta$,

$$\begin{aligned} &c_1(\theta_1) + k(\theta_1) = \tau_1(\theta_1), \\ &c_2(\theta_1,\theta_2) = \tau_2(\theta_1,\theta_2) + Rk(\theta_1), \\ &k(\theta_1) \geq 0, \end{aligned} \tag{63}$$

and the incentive-compatibility constraints: $\forall \hat{\theta}_1, \hat{\theta}_2 \in \Theta$, $\forall \hat{k} \geq 0$,

[n] This definition of feasibility is that of Ljungqvist and Sargent (2012) rather than that of Cole and Kocherlakota (2001), who assume that the planner cannot borrow.

[o] The results of this section extend to finite horizons if the utility function has non-increasing absolute risk aversion, and to the infinite horizon if the utility function is bounded.

$$\sum_{\theta_1\in\Theta}\pi(\theta_1)\left\{\theta_1 U(\tau_1(\theta_1)-k(\theta_1))+\sum_{\theta_2\in\Theta}\beta\pi(\theta_2)\theta_2 U(\tau_2(\theta_1,\theta_2)+Rk(\theta_1))\right\}$$
$$\geq\sum_{\theta_1\in\Theta}\pi(\theta_1)\left\{\theta_1 U\left(\tau_1\left(\hat{\theta}_1\right)-\hat{k}(\theta_1)\right)+\sum_{\theta_2\in\Theta}\beta\pi(\theta_2)\theta_2 U\left(\tau_2\left(\hat{\theta}_1,\hat{\theta}_2\right)+R\hat{k}(\theta_1)\right)\right\}. \tag{64}$$

We first note that there is no loss to having the planner do all the (public plus private) saving publicly, since the agent and the planner have the same rate of return R.

Lemma 5 *Given any incentive-compatible and feasible allocation* $\{\mathbf{c},\boldsymbol{\tau},k,K\}$, *there exists another incentive-compatible and feasible allocation* $\{\mathbf{c},\boldsymbol{\tau}^0,0,K^0\}$.

Proof Define the transfers $\tau_1^0(\theta_1)=\tau_1(\theta_1)-k(\theta_1)$, $\tau_2^0(\theta_1,\theta_2)=\tau_2(\theta_1,\theta_2)+Rk(\theta_1)$, and the public saving $K^0=K+\sum_{\theta_1\in\Theta}\pi(\theta_1)k(\theta_1)$. The allocation $\{\mathbf{c},\boldsymbol{\tau}^0,0,K^0\}$ with $\mathbf{c}=\boldsymbol{\tau}^0$ clearly satisfies the planner's and the households budget constraints, and hence is feasible. We now show that it is incentive compatible. Indeed, suppose that there exists $\left(\hat{\theta}_1,\hat{\theta}_2,\hat{k}\right)$ such that

$$\sum_{\theta_1\in\Theta}\pi(\theta_1)\left\{\theta_1 U\left(\tau_1^0\left(\hat{\theta}_1\right)-\hat{k}(\theta_1)\right)+\sum_{\theta_2\in\Theta}\beta\pi(\theta_2)\theta_2 U\left(\tau_2^0\left(\hat{\theta}_1,\hat{\theta}_2\right)+R\hat{k}(\theta_1)\right)\right\}$$
$$>\sum_{\theta_1\in\Theta}\pi(\theta_1)\left\{\theta_1 U\left(\tau_1^0(\theta_1)\right)+\sum_{\theta_2\in\Theta}\beta\pi(\theta_2)\theta_2 U\left(\tau_2^0(\theta_1,\theta_2)\right)\right\}.$$

Then the strategy $\left(\hat{\theta}_1,\hat{\theta}_2,\left\{k(\theta_1)+\hat{k}(\theta_1)\right\}\right)$ dominates $(\theta_1,\theta_2,k(\theta_1))$, so that $\{\mathbf{c},\boldsymbol{\tau},k,K\}$ is not incentive compatible. □

Next, note that the incentive constraint in period 2 imposes that the second-period transfers are independent of the report $\hat{\theta}_2$ (otherwise the agent would always report the type that yields the highest transfer regardless of his true type). Thus we can rewrite the transfers from the planner to the agent as $\tau_1(\theta_1),\tau_2(\theta_1)$.

The possibility of hidden storage in the incentive constraints (64) makes the planner's problem (61)–(64) difficult to solve directly. In a first step, we thus consider a simpler planner's problem with a larger constraint set: we suppose that the agent can only lie upward by one notch. Thus we analyze the following relaxed problem:

$$\max_{\{\tau_1(\theta_1),\tau_2(\theta_1)\}}\quad\sum_{\theta_1\in\Theta}\pi(\theta_1)\left\{\theta_1 U(\tau_1(\theta_1))+\sum_{\theta_2\in\Theta}\beta\pi(\theta_2)\theta_2 U(\tau_2(\theta_1))\right\} \tag{65}$$

subject to

$$\sum_{\theta_1\in\Theta}\pi(\theta_1)\tau_1(\theta_1)+K=e,$$
$$\sum_{\theta_1\in\Theta}\pi(\theta_1)\tau_2(\theta_1)=e+RK, \tag{66}$$

and, for all $\hat{k}\geq 0$ and σ such that $\sigma\left(\theta_{(j)}\right)\in\left\{\theta_{(j)},\theta_{(j+1)}\right\}$ for all $j\in\{1,\ldots,|\Theta|-1\}$ and $\sigma\left(\theta_{(|\Theta|)}\right)=\theta_{(|\Theta|)}$,

$$\sum_{\theta_1\in\Theta}\pi(\theta_1)\left\{\theta_1U(\tau_1(\theta_1))+\sum_{\theta_2\in\Theta}\beta\pi(\theta_2)\theta_2U(\tau_2(\theta_1))\right\}$$
$$\geq\sum_{\theta_1\in\Theta}\pi(\theta_1)\left\{\theta_1U\left(\tau_1(\sigma(\theta_1))-\hat{k}(\theta_1)\right)+\sum_{\theta_2\in\Theta}\beta\pi(\theta_2)\theta_2U\left(\tau_2(\sigma(\theta_1))+R\hat{k}(\theta_1)\right)\right\}. \tag{67}$$

We start by analyzing the relaxed problem (65). We do this in two steps. First, we show:

Lemma 6 *Consider any allocation that solves (65), say* $\{\mathbf{c},\boldsymbol{\tau},0,K\}$. *It must satisfy*

$$\theta_{(j)}U'\left(c_1\left(\theta_{(j)}\right)\right)=\beta R\sum_{\theta_{(j')}\in\Theta}\pi\left(\theta_{(j')}\right)\theta_{(j')}U'\left(c_2\left(\theta_{(j)},\theta_{(j')}\right)\right), \tag{68}$$

for all $j\in\{1,\ldots,|\Theta|\}$.

Proof Suppose first by contradiction that there exists $i\in\{1,\ldots,|\Theta|\}$ such that

$$\theta_{(i)}U'\left(c_1\left(\theta_{(i)}\right)\right)<\beta R\sum_{\theta_{(j')}\in\Theta}\pi\left(\theta_{(j')}\right)\theta_{(j')}U'\left(c_2\left(\theta_{(i)},\theta_{(j')}\right)\right).$$

Then, by saving $\hat{k}\left(\theta_{(i)}\right)>0$, agent $\theta_{(i)}$ raises his ex ante discounted utility, which contradicts the incentive constraint. Thus, because of the availability of private saving, individuals can only be borrowing constrained and not saving constrained.

Next, suppose that there exists $i\in\{1,\ldots,|\Theta|\}$ such that

$$\theta_{(i)}U'\left(c_1\left(\theta_{(i)}\right)\right)>\beta R\sum_{\theta_{(j')}\in\Theta}\pi\left(\theta_{(j')}\right)\theta_{(j')}U'\left(c_2\left(\theta_{(i)},\theta_{(j')}\right)\right). \tag{69}$$

We then construct an alternative incentive-compatible and feasible allocation $\left\{\tilde{\mathbf{c}},\tilde{\boldsymbol{\tau}},0,\tilde{K}\right\}$ that yields strictly higher ex ante utility than $\{\mathbf{c},\boldsymbol{\tau},0,K\}$. Specifically, let

$$\tilde{\tau}_1\left(\theta_{(i)}\right) = \tau_1\left(\theta_{(i)}\right) + \varepsilon_1,$$
$$\tilde{\tau}_2\left(\theta_{(i)}\right) = \tau_2\left(\theta_{(i)}\right) - \varepsilon_2,$$
$$\tilde{K} = K - \pi\left(\theta_{(i)}\right)\varepsilon_1,$$

where $(\varepsilon_1, \varepsilon_2)$ are chosen such that

$$\theta_{(i)} U\left(\tilde{\tau}_1\left(\theta_{(i)}\right)\right) + \beta \sum_j \pi\left(\theta_{(j)}\right)\theta_{(j)} U\left(\tilde{\tau}_2\left(\theta_{(i)}\right)\right)$$
$$= \theta_{(i)} U\left(\tau_1\left(\theta_{(i)}\right)\right) + \beta \sum_j \pi\left(\theta_{(j)}\right)\theta_{(j)} U\left(\tau_2\left(\theta_{(i)}\right)\right),$$

and

$$\theta_{(i)} U'\left(\tilde{\tau}_1\left(\theta_{(i)}\right)\right) \geq \beta R \sum_{\theta_{(j')} \in \Theta} \pi\left(\theta_{(j')}\right)\theta_{(j')} U'\left(\tilde{\tau}_2\left(\theta_{(i)}\right)\right). \tag{70}$$

That is, the alternative allocation slightly raises the transfer to agent $\theta_{(i)}$ in period 1 and slightly lowers it in period 2, in a way that makes him indifferent between the initial and the perturbed allocation, and by an amount small enough that he is still (weakly) borrowing constrained.

Since (69) holds, by the envelope condition we have $\varepsilon_2 > R\varepsilon_1$. Therefore this alternative allocation frees up resources, ie,

$$\sum_{\theta_1 \in \Theta} \pi(\theta_1)\tilde{\tau}_2(\theta_1) < e + R\tilde{K}.$$

These resources can be used to raise agents' ex ante utility in the following way: we can give them in period 2 to the household that reports the lowest taste shock θ_1. This does not violate any incentive constraints, since by assumption agents can only lie upward, and this does not lead to any private storage since the additional consumption is given in the second period.

We finally show that the alternative allocation $\left\{\tilde{\mathbf{c}}, \tilde{\boldsymbol{\tau}}, 0, \tilde{K}\right\}$ is incentive compatible. First, the incentive compatibility is satisfied for individual $\theta_{(i)}$, since his payoffs from truthtelling and from lying are unchanged by construction, and (70) ensures that he still finds it optimal to not privately store ($\hat{k} = 0$).

Thus it remains to prove that agent $\theta_{(i-1)}$ does not want to lie upward. Intuitively, the perturbation is constructed so that the planner borrows (ie, reduces public saving K) at rate R, and then offers a loan ε_1 to the borrowing constrained individual at his shadow interest rate $\varepsilon_2/\varepsilon_1 > R$ (which generates extra resources). Now, the individuals who lie have a lower actual taste shock than their report (ie, $\theta_{(i-1)} < \hat{\theta}_{(i)}$), and hence a lower shadow interest rate than that of the thruthteller $\theta_{(i)}$: they are less desperate to consume a bit more today in exchange for a larger consumption loss ε_2 tomorrow. They are thus made strictly worse off by the planner's loan if they lie.

To show this formally, define, for any $\theta \in \mathbb{R}_+$,

$$Z(\theta) = \max_{k \geq 0} \left\{ \theta U\left(\tau_1\left(\theta_{(i)}\right) - k\right) + \beta \sum_j \pi\left(\theta_{(j)}\right)\theta_{(j)} U\left(\tau_2\left(\theta_{(i)}\right) + Rk\right) \right\},$$

$$W(\theta) = \max_{k \geq 0} \left\{ \theta U\left(\tau_1\left(\theta_{(i)}\right) + \varepsilon_1 - k\right) + \beta \sum_j \pi\left(\theta_{(j)}\right)\theta_{(j)} U\left(\tau_2\left(\theta_{(i)}\right) - \varepsilon_2 + Rk\right) \right\}.$$

By construction of the perturbed allocation, we have $Z\left(\theta_{(i)}\right) = W\left(\theta_{(i)}\right)$. We want to show that $Z\left(\theta_{(i-1)}\right) > W\left(\theta_{(i-1)}\right)$ (so that agent $\theta_{(i-1)}$ finds it even worse to lie than he did before the planner perturbed the allocation). Suppose by contradiction that $W\left(\theta_{(i-1)}\right) \geq Z\left(\theta_{(i-1)}\right)$. Then by the mean value theorem, we have $W'(\theta) \leq Z'(\theta)$ for some $\theta \in \left(\theta_{(i-1)}, \theta_{(i)}\right)$. This can be written as

$$U\left(\tau_1\left(\theta_{(i)}\right) - k_W(\theta) + \varepsilon_1\right) \leq U\left(\tau_1\left(\theta_{(i)}\right) - k_Z(\theta)\right),$$

where $k_W(\theta)$ and $k_Z(\theta)$ denote the argmax of $W(\theta)$ and $Z(\theta)$, respectively. This equation leads to $k_W(\theta) - \varepsilon_1 \geq k_Z(\theta) \geq 0$, which in turn implies $k_W\left(\theta_{(i-1)}\right) \geq k_W(\theta) \geq \varepsilon_1$, as we can easily show by differentiating the relevant first-order condition that $k_W(\cdot)$ is weakly monotonic. Therefore, we have

$$\begin{aligned} W\left(\theta_{(i-1)}\right) &= \theta_{(i-1)} U\left(\tau_1\left(\theta_{(i)}\right) - \left\{k_W\left(\theta_{(i-1)}\right) - \varepsilon_1\right\}\right) \\ &\quad + \beta \sum_j \pi\left(\theta_{(j)}\right)\theta_{(j)} U\left(\tau_2\left(\theta_{(i)}\right) + Rk_W\left(\theta_{(i-1)}\right) - \varepsilon_2\right) \\ &< \theta_{(i-1)} U\left(\tau_1\left(\theta_{(i)}\right) - \left\{k_W\left(\theta_{(i-1)}\right) - \varepsilon_1\right\}\right) \\ &\quad + \beta \sum_j \pi\left(\theta_{(j)}\right)\theta_{(j)} U\left(\tau_2\left(\theta_{(i)}\right) + R\left\{k_W\left(\theta_{(i-1)}\right) - \varepsilon_1\right\}\right) \\ &\leq Z\left(\theta_{(i-1)}\right), \end{aligned}$$

where the first inequality uses the fact that $\varepsilon_2 > R\varepsilon_1$, and the second inequality invokes the fact (shown above) that $k_W\left(\theta_{(i-1)}\right) - \varepsilon_1 \geq 0$. Therefore, we have proved by contradiction that the agent $\theta_{(i-1)}$ does not want to lie upward and take the planner's loan at the implied rate $\varepsilon_2/\varepsilon_1 > R$ at which agent $\theta_{(i)}$ is indifferent. □

The second step consists in showing that all agents receive the same present value of transfers:

Lemma 7 *For all $\theta_1 \in \Theta$,*

$$\tau_1(\theta_1) + \frac{1}{R}\tau_2(\theta_1) = \left(1 + \frac{1}{R}\right)e. \tag{71}$$

Proof The planner's intertemporal budget constraint writes

$$\sum \pi(\theta_1)\left(\tau_1(\theta_1)+\frac{1}{R}\tau_2(\theta_1)\right)=\left(1+\frac{1}{R}\right)e.$$

Thus, to prove the result, it is sufficient to show that for all $j \in \{1,\ldots,|\Theta|-1\}$, we have $\psi_j = \psi_{j+1}$, where we denote

$$\psi_j \equiv \tau_1\left(\theta_{(j)}\right)+\frac{1}{R}\tau_2\left(\theta_{(j)}\right).$$

Suppose first by contradiction that there exists $i \in \{1,\ldots,|\Theta|-1\}$ such that $\psi_i < \psi_{i+1}$. Define, for any (θ,ψ),

$$\tilde{Z}(\theta,\psi)=\max_{k\in\mathbb{R}}\left\{\theta U(\psi-k)+\beta\sum_j \pi\left(\theta_{(j)}\right)\theta_{(j)}U(Rk)\right\},$$

If agent $\theta_{(i)}$ reports his true type $\theta_{(i)}$, he reaches utility $\tilde{Z}\left(\theta_{(i)},\psi_i\right)$, since we know from the previous lemma that his consumption is optimally smoothed across periods. If instead he lies and reports $\theta_{(i+1)}$, he reaches utility $\tilde{Z}\left(\theta_{(i)},\psi_{i+1}\right)$ (and in particular will still be able to perfectly smooth his consumption), because his constraint $k \geq 0$ does not bind (since it does not bind for individuals with the higher taste shock). Thus agent $\theta_{(i)}$ is strictly better off lying upward, which contradicts incentive compatibility.

Suppose next that there exists $i \in \{1,\ldots,|\Theta|-1\}$ such that $\psi_i > \psi_{i+1}$. We then construct an alternative incentive-compatible and feasible allocation that yields strictly higher ex ante utility. Specifically, define the "certainty equivalent" $\overline{\psi}$ by

$$\begin{aligned}&\pi\left(\theta_{(i)}\right)\tilde{Z}\left(\theta_{(i)},\overline{\psi}\right)+\pi\left(\theta_{(i+1)}\right)\tilde{Z}\left(\theta_{(i+1)},\overline{\psi}\right)\\ =\;&\pi\left(\theta_{(i)}\right)\tilde{Z}\left(\theta_{(i)},\psi_i\right)+\pi\left(\theta_{(i+1)}\right)\tilde{Z}\left(\theta_{(i+1)},\psi_{i+1}\right).\end{aligned}$$

Since the utility function U is concave, this alternative allocation frees up resources that can be used to raise ex ante utility, as we already described above. Moreover, it is easy to see that all the incentive constraints remain satisfied: agent $\theta_{(i+1)}$ is now strictly better off when reporting truthfully; agent $\theta_{(i)}$ is now indifferent between reporting truthfully and lying upward, since he gets the same present value of resources for both reports, and his consumption is optimally smoothed when he reports the truth; and agent $\theta_{(i-1)}$ is now strictly worse off if he lies, since his present value of resources at $\theta_{(i)}$ is lower. □

Lemmas 6 and 7 together imply that the relaxed problem (65)–(67) has a unique solution $\{\mathbf{c}^*,\boldsymbol{\tau}^*,0,K^*\}$, with $\boldsymbol{\tau}^* = \mathbf{c}^*$ and $K^* = e - \sum \pi(\theta_1)\tau_1(\theta_1)$, and where $\mathbf{c}^*$ is given by the solution to the problem

$$\max_{\{c_1(\theta_1),c_2(\theta_1,\theta_2)\}} \sum_{\theta_1\in\Theta}\pi(\theta_1)\left\{\theta_1 U(c_1(\theta_1)) + \sum_{\theta_2\in\Theta}\beta\pi(\theta_2)\theta_2 U(c_2(\theta_1,\theta_2))\right\} \tag{72}$$

subject

$$c_1(\theta_1)+\frac{1}{R}c_2(\theta_1,\theta_2)=\left(1+\frac{1}{R}\right)e, \quad \forall(\theta_1,\theta_2)\in\Theta^2. \tag{73}$$

This is because Eqs. (68) and (71) characterize the unique solution to (72)–(73). The solution to the latter problem is the allocation in an economy where each household can borrow *and* lend at the risk-free gross interest rate R, subject to the natural debt limit, with a present value of income equal to the endowment $\left(1+\frac{1}{R}\right)e$.

We finally prove that the solution to the original planner's problem (61)–(64) is the same as the solution to (65)–(67).

Proposition 10 *Any allocation* $\{\mathbf{c},\boldsymbol{\tau},k,K\}$ *is efficient, ie, solves (61)–(64), if and only if* $\mathbf{c}=\mathbf{c}^*$*, where* $\mathbf{c}^*$ *is the solution to problem (72)–(73).*

Proof In the solution to the problem (65)–(67), the agents receive the same net present value of transfers regardless of what taste shock they report. Moreover, telling the truth and not storing is weakly optimal, because the planner already optimally smoothes the consumption of a truthtelling agent, so that lying would not increase the present value of transfers nor improve their allocation over time. Therefore any solution to (65)–(67) is fully incentive compatible in the sense of (64), ie, with respect to the unrestricted set of possible deviations $(\hat{\theta},\hat{k})$.

The conclusion of this section is that in an environment with hidden storage, the optimal transfers that the planner chooses effectively relax the nonnegativity constraint on household storage. However, the optimal transfers offer *no insurance across agents*, as the present value of transfers must equal the economy's endowment *for all* histories $(\theta_1,\theta_2)\in\Theta^2$ (Eq. (73)). As a result, the allocation replicates a self-insurance economy; Cole and Kocherlakota (2001) propose a decentralization of this allocation that can be interpreted as an explicit microfoundation for the models with exogenously incomplete markets, eg, Aiyagari (1994).

2.7 Other Models

The techniques that we introduced in the previous sections in the context of the taste shock model can be easily applied to many more environments. First, Green (1987) and Thomas and Worrall (1990) study a model closely related to the one we analyzed above, in which the agent receives privately observed i.i.d. or persistent endowment (or income) shocks $\theta_t\in\Theta$: in each period $t\geq 1$, the agent observes his income shock θ_t and reports its realization to the planner, who then provides a transfer $\tau_t(\theta^t)$ to the agent. Second, Spear and Srivastava (1987) and Phelan and Townsend (1991) study a moral hazard model in which agents exert a privately observed effort level $\theta_t\in\Theta$ in each period. The output produced from that effort is stochastic and observable to the planner. The case where current effort affects only current output corresponds to the i.i.d. assumption 3 in the taste shock model, while the case where current effort also affects future output corresponds to the taste shock model with persistent types. Third, Thomas and Worrall (1988), Kocherlakota (1996), and Ligon et al. (2002) show that models of limited commitment, in which there is no asymmetry of information but one or both parties are free to walk away from the insurance contract, can be analyzed using similar recursive techniques using promised utilities as state variables; we discuss examples of these models in Section 4.

Here we describe briefly how to apply our recursive techniques to a model of repeated moral hazard. Agents exert an effort level $\theta_t \in \Theta = [0, \infty)$ in each period. The planner does not observe the agent's effort, but only the (random) output produced from that effort, $y_t \in Y = \{y_{(1)}, y_{(2)}\}$, with $0 = y_{(1)} < y_{(2)}$. The flow utility at time t is $U(c_t) - h(\theta_t)$, where the utility from consumption $U(\cdot): \mathbb{R}_+ \to \mathbb{R}$ is differentiable, strictly increasing, and strictly concave, and the disutility of effort $h(\cdot): \mathbb{R}_+ \to \mathbb{R}$ is differentiable, strictly increasing, and strictly convex with $h(0) = 0$ and $h'(0) \geq 0$.

We assume that the probability of output $y_t \in Y$ in period t depends only on the effort $\theta_t \in \Theta$ exerted by the agent in the *current* period.[p] We denote it by $\pi(y_t|\theta_t)$, and we suppose that $0 = \pi(y_{(2)}|0) < \pi(y_{(2)}|\theta) < 1$ for all $\theta > 0$, and $\pi(y_{(2)}|\cdot)$ is twice differentiable with $\pi_\theta(y_{(2)}|\cdot) > 0$. An allocation in this model consists of a sequence $\boldsymbol{\theta} = \{\theta_t(y^{t-1})\}_{t\geq 1}$ (with $y^0 = \varnothing$) describing the effort recommended by the planner to the agent given the observed history of output at the beginning of each period t, and a sequence of utility payments $\mathbf{u} = \{u_t(y^t)\}_{t\geq 1}$ given the observed history of output at the end of each period t. The planner chooses the incentive-compatible allocation $\{\mathbf{c}, \boldsymbol{\theta}\}$ that minimizes the cost of delivering lifetime utility v_0, that is, letting $C \equiv U^{-1}$,

$$\begin{aligned} K(v_0) \equiv \max_{\boldsymbol{\theta}, \mathbf{u}} \quad & \mathbb{E}^{\boldsymbol{\theta}}\left[\sum_{t=1}^{\infty} \beta^{t-1}\{y_t - C(u_t(y^t))\}\right] \\ \text{subject to} \quad & \mathbb{E}^{\boldsymbol{\theta}}\left[\sum_{t=1}^{\infty} \beta^{t-1}\left\{u_t(y^t) - h\left(\theta_t\left(y^{t-1}\right)\right)\right\}\right] = v_0, \\ & \mathbb{E}^{\hat{\boldsymbol{\theta}}}\left[\sum_{t=1}^{\infty} \beta^{t-1}\left\{u_t(y^t) - h\left(\hat{\theta}_t\left(y^{t-1}\right)\right)\right\}\right] \leq v_0, \quad \forall \hat{\boldsymbol{\theta}}, \end{aligned} \tag{74}$$

[p] The analysis of the case where current effort also affects future output is slightly more involved than that of Section 2.5. This is because there is a form of nonseparability of the agent's lifetime utility (incentives in a given period depend no longer only on his current true type and past reports, but also on his past true types) which implies that truthful revelation does not necessarily hold after the agent has deviated from the recommended action in the past; see Example of Section S.5 in Pavan et al. (2014). Thus, after a deviation an agent may prefer to engage in a strategy of infinite deviations, so that one generally cannot restrict attention to one shot deviations in such settings. Fernandes and Phelan (2000) nevertheless show how to modify the arguments of Section 2.5 to write a recursive formulation of this problem.

where the superscripts over expectations $\mathbb{E}^{\boldsymbol{\theta}}$ and $\mathbb{E}^{\hat{\boldsymbol{\theta}}}$ indicate that the probability distributions over the paths of output $\{y^t\}_{t\geq 1}$ depend on the agent's strategies $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ (respectively), that is, for any t and random variable $X_t(y^t)$ we let $\mathbb{E}^{\boldsymbol{\theta}}[X_t] \equiv \sum_{y^t \in Y^t} \pi_t(y^t|\theta^t) X_t(y^t)$. Thus, each expectation in the incentive constraint depends on the agent's effort directly through the cost of effort $h(\theta_t)$, and indirectly through its effect on the probability distribution $\pi_t(y^t|\boldsymbol{\theta})$ over the paths of y^t.

Defining the continuation utility of an obedient (truthtelling) agent up to and after date t as in (11), we can rewrite this problem recursively:

$$
\begin{aligned}
K(v) = \max_{\theta, \vec{u}, \vec{w}} \; & \sum_{y \in Y} \pi(y|\theta)[y - C(u(y)) + \beta K(w(y))] \\
\text{s.t.} \quad & v = \sum_{y \in Y} \pi(y|\theta)[u(y) - h(\theta) + \beta w(y)], \\
& \pi_\theta\left(y_{(2)}|\theta\right)\left[\left(u\left(y_{(2)}\right) - u\left(y_{(1)}\right)\right) + \beta\left(w\left(y_{(2)}\right) - w\left(y_{(1)}\right)\right)\right] - h'(\theta) \leq 0 \\
& \qquad \text{with equality if } \theta > 0,
\end{aligned}
\tag{75}
$$

where the incentive-compatibility constraint is replaced by a first-order condition, assuming for simplicity that this condition is sufficient.

Following the steps leading to Proposition 5, we can obtain a characterization of the solution to the planner's problem. For any interior v, the optimal contract $\left(\theta_v, \vec{u}_v, \vec{w}_v\right)$ satisfies the following martingale property (with respect to the probability measure associated with the optimum effort strategy $\mathbb{P}^{\boldsymbol{\theta}}$):

$$
K'(v) = \mathbb{E}^{\theta_v}[-C'(u_v)] = \mathbb{E}^{\theta_v}[K'(w_v)]. \tag{76}
$$

The first-order conditions of the problem imply moreover that $K'\left(w_v\left(y_{(j)}\right)\right) = -C'\left(u_v\left(y_{(j)}\right)\right)$, so that this property can be rewritten as:

$$
\frac{1}{u'(c_t(y^t))} = \sum_{y_{t+1} \in Y} \pi(y_{t+1}|\theta_{t+1}(y^t)) \frac{1}{u'(c_{t+1}(y^t, y_{t+1}))}. \tag{77}
$$

This equation is known in the literature as the *Inverse Euler Equation* (see Diamond and Mirrlees, 1978; Rogerson, 1985; Spear and Srivastava, 1987; Golosov et al., 2003). We derive implications of this equation in Section 4.1 and show by comparing it to the individual's Euler equation in a decentralized economy that agents' savings must be constrained in the optimal insurance arrangement.

We can further analyze problem (75) along the lines of the proof of Proposition 5. A utility-effort pair (v, θ_v) is absorbing if and only if $\theta_v = 0$ and $(u_v(y), w_v(y)) = ((1-\beta)v, v)$. The recommended effort θ_v is strictly positive as long as the promised utility is small enough, $v < \bar{v}$. If $h'(0) = 0$, we find that $\bar{v} = \infty$, so that the recommended effort is always positive, and the Martingale Convergence Theorem implies that immiseration occurs: $v_t(\theta^t) \to \underline{v}$ as $t \to \infty$ with probability 1. If instead $h'(0) > 0$, the principal will eventually "retire" the agent (ie, recommend effort $\theta_t(y^t) = 0$ and provide constant consumption $c_t(y^t) = c$) when $v_t(\theta^t) \geq \bar{v}$, as for a large enough promised utility the benefit of inducing him to work outweighs the cost of providing the necessary incentives and compensating him for the higher effort. We leave the formal proof and derivation of the value of $\bar{v}$ to the reader.

3. ADVANCED TOPICS

In this section we discuss three additional topics that significantly expand the applicability of the recursive contract theory. In Section 3.1 we overview the theory of Lagrange multipliers and show how it can help solve many dynamic incentive problems recursively even if they do not fit into the canonical setup described in Section 2. Section 3.2 shows how to extend the analysis to settings in which the ability of the principal to commit is imperfect. Finally, in Section 3.3, we describe the analysis of dynamic contracting problems in continuous time using martingale methods. Throughout this section we do not aim at the same level of rigor as in Section 2; we omit several technical details and refer to the relevant papers for the complete proofs.

3.1 Lagrange Multipliers

The key feature that allowed us to analyze the dynamic contracting problem (17) is that we could write the incentive constraints in a simple recursive form. In many applications, however, the optimal contracting problem often has additional constraints that cannot easily be written recursively. For example, if we replaced the present value budget constraint (2) with a requirement that the total consumption of all agents should be equal to the total endowment in *each* period, the previous method could not be applied directly. In this section we describe a simple approach that allows us to extend our analysis to such problems. The main idea behind this approach is to assign Lagrange multipliers to all the constraints that do not have a straightforward recursive representation, and to apply the techniques developed in the previous sections to the resulting Lagrangian.

We start in Section 3.1.1 by giving a general theoretical background about the properties of Lagrange multipliers in infinite dimensional spaces. Infinite dimensional spaces are common in macroeconomic applications but the Lagrangian techniques are more subtle in such spaces than in finite dimensions. The main results of this section are, first, Theorems 3 and 4, which provide conditions under which the Lagrangian exists and characterize the solution to the constrained optimization problem, and second, Theorem 5, which provides sufficient conditions that ensure that the Lagrangian can be written as an infinite sum, allowing us to apply the standard techniques familiar from finite-dimensional optimization theory. Sections 3.1.2–3.1.4 give several examples of applications of these techniques. The reader only interested in practical applications can skip Section 3.1.1 in the first reading.

3.1.1 Main Theoretical Results

The classical reference about using Lagrange multipliers to solve optimization problems is Luenberger (1969). Here we state two main results from this book, adapting them to our setting. To use this approach, we need to set our problem in abstract linear spaces.[q] Before starting our analysis we introduce the notions of convex cones and mappings, dual spaces, and l_p spaces.

First, let P be a convex cone in a vector space $\mathcal{V}$, that is, P satisfies $\alpha x + \beta y \in P$ for all $x, y \in P$ and $\alpha, \beta > 0$. This convex cone defines a partial order $\leq$ on $\mathcal{V}$, such that $x \geq y$ if $x - y \in P$. By definition, P is the positive cone with respect to this partial order, ie, the subset $\mathcal{V}^+ = \{x \in \mathcal{V} : x \geq 0\}$. We write $x > 0$ if x is an interior point of the positive cone P. By introducing a cone defining the positive vectors in the vector space $\mathcal{V}$, we thus define an ordering relation $\leq$ and make it possible to consider inequality problems in the abstract vector space $\mathcal{V}$. (Often the positive cones of the vector spaces we consider are constructed naturally, eg, the positive orthant of $\mathbb{R}^n$ or the nonnegative continuous functions of $\mathcal{C}([a,b])$.) A mapping $G : \mathcal{V}_1 \to \mathcal{V}_2$ from a vector space $\mathcal{V}_1$ to a vector space $\mathcal{V}_2$ having a cone P defined as the positive cone is said to be *convex* if the domain Ω of G is a convex set and if $G(\alpha x_1 + (1-\alpha)x_2) \leq \alpha G(x_1) + (1-\alpha)G(x_2)$ for all $x_1, x_2 \in \Omega$ and all $\alpha \in (0,1)$.

Second, the dual $\mathcal{V}^*$ of a normed vector space $\mathcal{V}$ is the space of all bounded linear functionals on $\mathcal{V}$, ie, $f : \mathcal{V} \to \mathbb{R}$. The norm of an element $f \in \mathcal{V}^*$ is $\|f\| = \sup_{\|x\| \leq 1} |f(x)|$. The value of the linear functional $x^* \in \mathcal{V}^*$ at the point $x \in \mathcal{V}$, that is $x^*(x)$, is denoted by $\langle x, x^* \rangle$. For $1 \leq p < \infty$, the space l_p consists of all sequences of scalars $\{u_1, u_2, \ldots\}$ for which $\sum_{n=1}^{\infty} |u_n|^p < \infty$, and the space l_∞ consists of the bounded sequences. The norm of an element $u = \{u_n\}_{n \geq 1} \in l_p$ is defined as $\|u\|_p = \left(\sum_{i=1}^{\infty} |u_n|^p\right)^{1/p}$ for $p < \infty$, and as $\|u\|_p = \sup_n |u_n|$ for $p = \infty$. Then for every $p \in [1, \infty)$, the dual space of l_p is l_q, where $q = (1 - p^{-1})^{-1}$. This is because every bounded linear functional f on l_p, $1 \leq p < \infty$, can be represented uniquely in the form $f(u) = \sum_{n=1}^{\infty} v_n u_n$, where $v = \{v_n\}_{n \in \mathbb{N}^*}$ is an element of l_q; specifically, for all $n \geq 1$, $v_n \equiv f(e_n)$, where $e_n \in l_p$ is the sequence that is identically zero except for a 1 in the n^{th} component. The dual of l_∞, however, *strictly contains* l_1. Finally, given a normed space $\mathcal{V}$ together with a positive convex cone $P \subset \mathcal{V}$, it is natural to define a corresponding positive convex cone P^* in the dual space $\mathcal{V}^*$ by $P^* = \{x^* \in \mathcal{V}^* : \forall x \in P,\ \langle x, x^* \rangle \geq 0\}$.

[q] For a review of basic functional analysis, see Luenberger (1969), or Chapters 3 and 15 in Stokey et al. (1989).

We can now introduce the theory of Lagrange multipliers. Consider a problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & \varphi(\mathbf{x}) \\ \text{subject to} \quad & \Phi(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in \Gamma, \end{aligned} \tag{78}$$

where Γ is a convex subset of a vector space X, $\varphi : \Gamma \rightarrow \mathbb{R}$ is a convex functional, and $\Phi : \Gamma \rightarrow Z$ is a convex mapping to a normed vector space Z that has positive cone P. Let Z^* be the dual space of Z and Z_+^* be its positive orthant (ie, all $\mathbf{z}^* \in Z^*$ such that $\mathbf{z}^* \geq \mathbf{0}$). We assume throughout this section that the minimum of problem (78) is attained. This assumption is not necessary but it simplifies the statement of the theorems, and we will see in our context (Proposition 11) that it can often be verified directly. Theorem 1, p. 217, and Corollary 1, p. 219 in Luenberger (1969), give the main results for solving the minimization problem (78) using Lagrange multipliers.

Theorem 3 *Assume that the minimum in (78) is achieved at $\hat{\mathbf{x}}$. Suppose that P contains an interior point, and that there exists $\mathbf{x}' \in \Gamma$ such that $\Phi(\mathbf{x}') < 0$. Then there is $\hat{\mathbf{z}}^* \in Z_+^*$ such that the Lagrangian*

$$\mathcal{L}(\mathbf{x}, \mathbf{z}^*) = \varphi(\mathbf{x}) + \langle \Phi(\mathbf{x}), \mathbf{z}^* \rangle$$

has a saddle point at $(\hat{\mathbf{x}}, \hat{\mathbf{z}}^*)$, ie,

$$\mathcal{L}(\hat{\mathbf{x}}, \mathbf{z}^*) \leq \mathcal{L}(\hat{\mathbf{x}}, \hat{\mathbf{z}}^*) \leq \mathcal{L}(\mathbf{x}, \hat{\mathbf{z}}^*), \forall \mathbf{x} \in \Gamma, \mathbf{z}^* \in Z_+^*. \tag{79}$$

Moreover,

$$\langle \Phi(\hat{\mathbf{x}}), \hat{\mathbf{z}}^* \rangle = 0.$$

Theorem 3 establishes that for convex problems there generally exists a Lagrangian such that the solution to the original constrained minimization problem is also a solution to the minimization of the unconstrained Lagrangian. The next result (Theorem 2, p. 221 in Luenberger, 1969) ensures the sufficiency:

Theorem 4 *Let $X, Z, \Gamma, P, \varphi, \Phi$ be as above and assume that the positive cone $P \subset Z$ is closed. Suppose that there exist $\hat{\mathbf{z}}^* \in Z_+^*$ and an $\hat{\mathbf{x}} \in \Gamma$ such that the Lagrangian $\mathcal{L}(\mathbf{x}, \mathbf{z}^*)$ has a saddle point at $(\hat{\mathbf{x}}, \hat{\mathbf{z}}^*)$. Then $\hat{\mathbf{x}}$ is a solution to (78).*

Thus, if φ and Φ are convex, the positive cone $P \subset Z$ is closed and has nonempty interior, and the regularity condition $\Phi(\mathbf{x}') < 0$ is satisfied, then the saddle point condition is necessary and sufficient for the optimality of $\hat{\mathbf{x}}$.

One way to find a saddle point of $\mathcal{L}$ is to use the following result (see Bertsekas et al., 2003).

Corollary 1 $(\hat{\mathbf{x}},\hat{\mathbf{z}}^*)$ is a saddle point of $\mathcal{L}$ if and only if the equality

$$\inf_{\mathbf{x}\in\Gamma}\sup_{\mathbf{z}^*\in Z_+^*}\mathcal{L}(\mathbf{x},\mathbf{z}^*)=\sup_{\mathbf{z}^*\in Z_+^*}\inf_{\mathbf{x}\in\Gamma}\mathcal{L}(\mathbf{x},\mathbf{z}^*) \tag{80}$$

is satisfied, and

$$\begin{aligned}\hat{\mathbf{x}}= &\quad \arg\min_{\mathbf{x}\in\Gamma}\sup_{\mathbf{z}^*\in Z_+^*}\mathcal{L}(\mathbf{x},\mathbf{z}^*),\\ \hat{\mathbf{z}}^*= &\quad \arg\max_{\mathbf{z}^*\in Z_+^*}\inf_{\mathbf{x}\in\Gamma}\mathcal{L}(\mathbf{x},\mathbf{z}^*).\end{aligned} \tag{81}$$

In particular, suppose that the conditions of Theorem 3 hold, so that $\mathcal{L}(\mathbf{x},\mathbf{z}^*)$ has a saddle point at $(\hat{\mathbf{x}},\hat{\mathbf{z}}^*)$. Suppose moreover that $\arg\min_{\mathbf{x}\in\Gamma}\mathcal{L}(\mathbf{x},\mathbf{z}^*)$ exists for each $\mathbf{z}^*\in Z_+^*$ and is unique for $\mathbf{z}^*=\hat{\mathbf{z}}^*$. Then $(\hat{\mathbf{x}},\hat{\mathbf{z}}^*)$ is the solution to $\max_{\mathbf{z}^*\in Z_+^*}\min_{\mathbf{x}\in\Gamma}\mathcal{L}(\mathbf{x},\mathbf{z}^*)$.

Proof Suppose $(\hat{\mathbf{x}},\hat{\mathbf{z}}^*)$ is a saddle point. Then

$$\inf_{\mathbf{x}\in\Gamma}\sup_{\mathbf{z}^*\in Z_+^*}\mathcal{L}(\mathbf{x},\mathbf{z}^*)\leq\sup_{\mathbf{z}^*\in Z_+^*}\mathcal{L}(\hat{\mathbf{x}},\mathbf{z}^*)=\mathcal{L}(\hat{\mathbf{x}},\hat{\mathbf{z}}^*)=\inf_{\mathbf{x}\in\Gamma}\mathcal{L}(\mathbf{x},\hat{\mathbf{z}}^*)\leq\sup_{\mathbf{z}^*\in Z_+^*}\inf_{\mathbf{x}\in\Gamma}\mathcal{L}(\mathbf{x},\mathbf{z}^*).$$

By the max–min inequality, $\inf_{\mathbf{x}\in\Gamma}\sup_{\mathbf{z}^*\in Z_+^*}\mathcal{L}(\mathbf{x},\mathbf{z}^*)\geq\sup_{\mathbf{z}^*\in Z_+^*}\inf_{\mathbf{x}\in\Gamma}\mathcal{L}(\mathbf{x},\mathbf{z}^*)$, establishing that all these inequalities hold with equality, and hence (80) and (81) are satisfied.

Conversely, suppose that (80) and (81) hold. Then

$$\sup_{\mathbf{z}^*\in Z_+^*}\inf_{\mathbf{x}\in\Gamma}\mathcal{L}(\mathbf{x},\mathbf{z}^*)=\inf_{\mathbf{x}\in\Gamma}\mathcal{L}(\mathbf{x},\hat{\mathbf{z}}^*)\leq\mathcal{L}(\hat{\mathbf{x}},\hat{\mathbf{z}}^*)\leq\sup_{\mathbf{z}^*\in Z_+^*}\mathcal{L}(\hat{\mathbf{x}},\mathbf{z}^*)=\inf_{\mathbf{x}\in\Gamma}\sup_{\mathbf{z}^*\in Z_+^*}\mathcal{L}(\mathbf{x},\mathbf{z}^*).$$

Eq. (80) implies that $(\hat{\mathbf{x}},\hat{\mathbf{z}}^*)$ is a saddle point.

Finally suppose that the conditions of Theorem 3 are satisfied, so that $\mathcal{L}(\mathbf{x},\mathbf{z}^*)$ has a saddle point at $(\hat{\mathbf{x}},\hat{\mathbf{z}}^*)$, and that $\mathbf{x}(\mathbf{z}^*)\equiv\arg\min_{\mathbf{x}\in\Gamma}\mathcal{L}(\mathbf{x},\mathbf{z}^*)$ exists for each $\mathbf{z}^*\in Z_+^*$ and is unique for $\mathbf{z}^*=\hat{\mathbf{z}}^*$. Then, by (81) we have $\hat{\mathbf{z}}^*=\arg\max_{\mathbf{z}^*\in Z_+^*}\mathcal{L}(\mathbf{x}(\mathbf{z}^*),\mathbf{z}^*)$. By the uniqueness assumption we have $\mathcal{L}(\mathbf{x}(\hat{\mathbf{z}}^*),\hat{\mathbf{z}}^*)<\mathcal{L}(\mathbf{x},\hat{\mathbf{z}}^*)$ for all $\mathbf{x}\neq\mathbf{x}(\hat{\mathbf{z}}^*)$, so that the saddle point (79) can only be achieved at $(\mathbf{x}(\hat{\mathbf{z}}^*),\hat{\mathbf{z}}^*)$, establishing that $\hat{\mathbf{x}}=\mathbf{x}(\hat{\mathbf{z}}^*)$. We obtain that the solution to $\max_{\mathbf{z}^*\in Z_+^*}\min_{\mathbf{x}\in\Gamma}\mathcal{L}(\mathbf{x},\mathbf{z}^*)$ is $(\hat{\mathbf{x}},\hat{\mathbf{z}}^*)$. □

The max-min problem in Corollary 1 provides a simple way to find the solution to the minimization problem together with the corresponding Lagrangian. The uniqueness qualifier is important for that result; without it there may exist solutions to the $\max\min$ problem that are not saddle points, ie, that are not a solution to the original optimization problem (see, eg, Messner and Pavoni, 2016).

In economic applications, $\mathbf{\Phi}$ often represents per-period constraints and it can be written as $\mathbf{\Phi}=\{\Phi_1,\Phi_2,\ldots\}$. The most natural vector space to choose in such situations is the space of bounded sequences, l_∞. In this case we define the positive cone P of l_∞ as the positive orthant, ie, the subset of nonnegative sequences of l_∞. Exercise 15.7 in

Stokey et al. (1989) shows that l_∞ is the only l_p space that has a positive orthant with a nonempty interior, which is a requirement needed to apply the theorems above.

A limitation of the space l_∞ is that its dual is complicated. It contains the space of summable sequences l_1, but it also includes other sequences which are not summable. This makes the analysis difficult because the linear operator $\langle \mathbf{\Phi}(\mathbf{x}), \hat{\mathbf{z}}^* \rangle$ may take a complicated form. The analysis simplifies if it can be ensured that the mappings φ and $\mathbf{\Phi}$ are not affected by how $\mathbf{x}$ behaves "at infinity," in which case we can provide an l_1 representation of the Lagrange multipliers and each constraint $\mathbf{\Phi}_n(\mathbf{x})$ will have a scalar multiplier λ_n associated with it. For any $\mathbf{x}, \mathbf{y} \in l_\infty$, define an operator $x^T(\mathbf{x}, \mathbf{y})$ as $x^T(\mathbf{x}, \mathbf{y}) = x_t$ if $t \leq T$ and $x^T(\mathbf{x}, \mathbf{y}) = y_t$ if $t > T$. We use the notation $x_t^T(\mathbf{x}, \mathbf{y})$ to denote the t-th element of this operator.

Assumption 5 Let $X, Z = l_\infty$, $\mathbf{\Psi} = \{\mathbf{x} \in \Gamma : \varphi(\mathbf{x}) < \infty\}$. Suppose that:

(i) If $(\mathbf{x}, \mathbf{y}) \in \mathbf{\Psi} \times l_\infty$ satisfy $x^T(\mathbf{x}, \mathbf{y}) \in \mathbf{\Psi}$ for all T large enough, then $\varphi\left(x^T(\mathbf{x}, \mathbf{y})\right) \to \varphi(\mathbf{x})$ as $T \to \infty$.

(ii) If $\mathbf{x}$,$\mathbf{y} \in \Gamma$ and $x^T(\mathbf{x}, \mathbf{y}) \in \Gamma$ for all T large enough, then:

$$\begin{aligned} &(a) \quad \forall t, \ \lim_{T \to \infty} \mathbf{\Phi}_t\left(x^T(\mathbf{x}, \mathbf{y})\right) = \mathbf{\Phi}_t(\mathbf{x}), \\ &(b) \quad \exists M \text{ s.t. } \forall T \text{ large enough}, \left\| \mathbf{\Phi}\left(x^T(\mathbf{x}, \mathbf{y})\right) \right\| \leq M, \\ &(c) \quad \forall T \text{ large enough}, \ \lim_{t \to \infty} \left[\mathbf{\Phi}_t\left(x^T(\mathbf{x}, \mathbf{y})\right) - \mathbf{\Phi}_t(\mathbf{y})\right] = 0. \end{aligned}$$

Le Van and Saglam (2004) prove that under these assumptions the Lagrangian can be written as an infinite sum:[r]

Theorem 5 *Let $\hat{\mathbf{x}}$ be a solution to (78). Suppose that for all $\mathbf{x} \in \Gamma$, we have $\mathbf{\Phi}(\mathbf{x}) \in l_\infty$. Assume that there exists $\mathbf{x}' \in \Gamma$ such that $\mathbf{\Phi}(\mathbf{x}') < 0$, that is, $\sup_t \mathbf{\Phi}_t(\mathbf{x}') < 0$ (Slater condition). Assume finally that Assumption 5 is satisfied and that $x^T(\hat{\mathbf{x}}, \mathbf{x}') \in \Gamma \cap \mathbf{\Psi}$ for all T large enough. Then there exists $\hat{\mathbf{z}}^* \in l_1$ with $\hat{\mathbf{z}}^* \geq \mathbf{0}$ such that*

$$\sum_{t=1}^{\infty} \hat{z}_t^* \mathbf{\Phi}_t(\hat{\mathbf{x}}) = 0,$$

and

$$\varphi(\mathbf{x}) + \sum_{t=1}^{\infty} \hat{z}_t^* \mathbf{\Phi}_t(\mathbf{x}) \geq \varphi(\hat{\mathbf{x}}) + \sum_{t=1}^{\infty} \hat{z}_t^* \mathbf{\Phi}_t(\hat{\mathbf{x}}), \forall \mathbf{x} \in \Gamma.$$

In the next sections, we apply this theory to dynamic contracting problems.

[r] See also Rustichini (1998) who provides an alternative set of sufficient conditions ensuring the summability of the Lagrange multipliers.

3.1.2 Application: Recursive Contracts in General Equilibrium

Consider a simple modification of the setup in Section 2.3, in which the planner can no longer freely borrow and lend at an exogenous interest rate. Instead we require the economy-wide feasibility constraint to hold period by period, ie,

$$\sum_{\theta^t \in \Theta^t} \pi_t(\theta^t) C(u_t(\theta^t)) \leq e, \forall t \geq 1. \tag{82}$$

This problem is analyzed by Atkeson and Lucas (1992). For simplicity we assume that $|\Theta| = 2$ and that shocks are i.i.d. to parallel our discussion in Sections 2.3 and 2.4. Thus we study the problem

$$\max_{\mathbf{u}} \mathbb{E}_0 \left[\sum_{t=1}^{\infty} \beta^{t-1} \theta_t u_t(\theta^t) \right] \tag{83}$$

subject to

$$\mathbb{E}_0[C(u_t(\theta^t))] \leq e, \forall t \geq 1, \tag{84}$$

and

$$\mathbb{E}_0 \left[\sum_{t=1}^{\infty} \beta^{t-1} \theta_t \{ u_t(\theta^t) - u_t(\sigma^t(\theta^t)) \} \right] \geq 0, \forall \boldsymbol{\sigma}. \tag{85}$$

Assume for now that the maximum in the problem (83) is attained for all $e > 0$; we will show this formally below.

Let Γ be the set of sequences $\mathbf{u} = \{u_t(\theta^t)\}_{t \geq 1, \theta^t \in \Theta^t}$, indexed by (t, θ^t), such that $\mathbf{u}$ satisfies the period-0 incentive constraint (85) and the sequence $\{\mathbb{E}_0[C(u_t)] - e\}_{t=1}^{\infty}$ is bounded in sup-norm. The set Γ is convex and has an interior point, eg, $u_t(\theta^t) = \varepsilon$ for all t, θ^t and $\varepsilon > 0$ sufficiently small.

We start with the sufficient conditions first. Let X be the space of all infinite sequences, $Z = l_\infty$, and $\mathbf{\Phi} = \{\Phi_1, \Phi_2, \ldots\}$, where $\mathbf{\Phi}_t : \Gamma \to \mathbb{R}$ is defined by $\mathbf{\Phi}_t(\mathbf{u}) = \mathbb{E}_0[C(u_t)] - e$. Suppose that we can find a nonnegative sequence $\boldsymbol{\lambda} = \{\lambda_t\}_{t=1}^{\infty}$ such that the problem[s]

$$\max_{\mathbf{u} \in \Gamma} \mathbb{E}_0 \left[\sum_{t=1}^{\infty} \beta^{t-1} \theta_t u_t \right] - \sum_{t=1}^{\infty} \lambda_t \{ \mathbb{E}_0[C(u_t)] - e \} \tag{86}$$

has a maximum $\hat{\mathbf{u}}$ and $\mathbb{E}_0[C(\hat{u}_t)] = e$ for all t. To verify that $(\hat{\mathbf{u}}, \boldsymbol{\lambda})$ is a saddle point, observe that for any $\mathbf{z}^* \in Z_+^*$, we have $\langle \mathbf{\Phi}(\hat{\mathbf{u}}), \mathbf{z}^* \rangle \leq 0 = \langle \mathbf{\Phi}(\hat{\mathbf{u}}), \boldsymbol{\lambda} \rangle$. Moreover, the regularity condition $\mathbf{\Phi}(\mathbf{u}') < 0$ holds for some $\mathbf{u}' \in \Gamma$, ie, $\mathbf{u}'$ satisfies incentive compatibility

[s] To be consistent with discussion in Section 3.1.1, we use the fact that minimizing φ is equivalent to maximizing $-\varphi$.

(take $u_t'(\theta^t) = \varepsilon$). Therefore $\hat{\mathbf{u}}$ is a solution to the original problem (83) by Theorem 4. Note that we impose no boundedness assumption on the utility function.

To illustrate an application of this result, consider an example with logarithmic preferences. We argue that the Lagrange multiplier $\boldsymbol{\lambda}$ has the form $\lambda_t = \lambda_1 \beta^t$ for some λ_1. Following the same steps as in Section 2.3, replace the period-0 incentive constraints with a sequence of one-shot constraints. Moreover, we consider an auxiliary planner's problem that has a recursive structure, by augmenting the set of constraints with the promise-keeping condition

$$\mathbb{E}_0\left[\sum_{t=1}^{\infty}\beta^{t-1}\theta_t u_t(\theta^t)\right] = v_0. \tag{87}$$

The constraint set is then the set $\Gamma(v_0)$ defined in (16). We can rewrite the problem (86)–(87) as

$$\max_{(\mathbf{u},\mathbf{v})\in\Gamma(v_0)} v_0 - \sum_{t=1}^{\infty}\lambda_t\{\mathbb{E}_0[C(u_t)] - e\}.$$

The solution to this problem coincides, given our guess $\lambda_t = \lambda_1\beta^t$, with the solution to the problem

$$\max_{(\mathbf{u},\mathbf{v})\in\Gamma(v_0)} -\sum_{t=1}^{\infty}\beta^{t-1}\mathbb{E}_0[C(u_t)],$$

which is, of course, the same problem as the one we analyzed in Section 2.3. Therefore, if we can show that the solution to that problem satisfies the feasibility constraint (82) for each period t, we found the solution to our new problem. We recover the solution to the original problem by maximizing the auxiliary problem over v_0.

We now check that this is the case. Let $(\mathbf{u}, \mathbf{v})$ be the allocation generated by the policy functions to the Bellman equation (23) for some v_0. The optimality conditions (29) imply

$$K'(v_0) = \mathbb{E}_0[-C'(u_1)] = \mathbb{E}_0[K'(v_1)] = \mathbb{E}_0[\mathbb{E}_1[-C'(u_2)]] = \mathbb{E}_0[-C'(u_2)].$$

When preferences are logarithmic, $C = C' = \exp$, thus forward induction implies $\mathbb{E}_0[C(u_1)] = \mathbb{E}_0[C(u_t)]$ for all t. Since v_0 must satisfy $K(v_0) = -\frac{1}{1-\beta}e$, this implies that $\mathbb{E}_0[C(u_t)] = e$ for all t, establishing our result (and justifying our guess for λ_t).

When we set up the maximization problem (86) we assumed the existence of a summable sequence $\boldsymbol{\lambda}$ such that the feasibility constraints are satisfied with equality in all periods at the optimum. We subsequently showed how to explicitly construct such a sequence of multipliers in an example with logarithmic preferences. We now conclude this section by discussing sufficient conditions ensuring the existence of a summable sequence of Lagrange multipliers. Note that without any further assumptions, the maximization problem

$$\max_{\mathbf{u}\in\Gamma,\,\Phi(\mathbf{u})\leq\mathbf{0}} \mathbb{E}_0\left[\sum_{t=1}^{\infty}\beta^{t-1}\theta_t u_t\right]$$

satisfies all the conditions of Theorem 3, so that a Lagrangian exists. To show that it is a summable sequence we verify conditions of Theorem 5. It is the easiest to do in the case of bounded utility.[t] In this case any sequence $\mathbf{u}$ lies in l_∞. Assumption 5.i holds following the arguments we use below in the proof of Proposition 11. Since the constraint (82) holds for each t, we immediately have $\mathbb{E}_0[C(u_t)] = \mathbb{E}_0\left[C\left(x_t^T(\mathbf{u},\mathbf{v})\right)\right]$ for T sufficiently large holding t fixed, and $\mathbb{E}_0\left[C\left(x_t^T(\mathbf{u},\mathbf{v})\right)\right] = \mathbb{E}_0[C(v_t)]$ for t sufficiently large holding T fixed, which verifies Assumptions 5.ii.a and 5.ii.c. Assumption 5.ii.b holds by definition of Γ. Therefore Theorem 5 establishes that the Lagrange multipliers form a summable sequence.

Existence of a Maximum

We finally show the existence of the maximum in problem (83). Note that we already showed the existence in Section 2.3.2 using the (finite-dimensional) Bellman formulation of the problem. Here we do so directly, using techniques that can be applied to other contexts where the previous approach is not readily available.

It is not obvious a priori that the maximum in this problem exists. In finite dimensional spaces, the continuity of the objective function and the compactness of the constraint sets are easily obtained, implying directly the existence of a maximum. These properties are more difficult to obtain in infinite period economies. The next proposition guarantees that the infinite-horizon planner's problem is a well-defined maximization problem, ie, there exist feasible $(\mathbf{u}^*, \mathbf{v}^*)$ for which the supremum is achieved. The reader interested mostly in the applications can skip this section.

Proposition 11 *The maximum in the problem (83) is attained for all $e > 0$.*

Proof One of the easiest ways to show the existence of the maximum in the planner's problem is to truncate the economy at any finite period T, show the existence of the solution for this truncated economy, and finally show that the limit of this solution achieves the supremum of the original problem as $T \rightarrow \infty$. To show these we adapt the arguments of Ekeland and Scheinkman (1986).

We first restrict allocations in each period to compact sets as follows. Fix $e > 0$. For any $t \geq 1$ and $\theta^t \in \Theta^t$, define $\bar{\bar{u}}_t(\theta^t) \in (\underline{u}, \bar{u})$ by

$$\bar{\bar{u}}_t(\theta^t) = C^{-1}\left(\frac{e}{\pi_t(\theta^t)}\right).$$

[t] See Rustichini (1998) for existence arguments when the utility is not bounded.

If $u_t(\theta^t) > \bar{\bar{u}}_t(\theta^t)$ for any history $\theta^t \in \Theta^t$, then $\mathbb{E}_0[C(u_t)] > e$, and the allocation is not feasible. This gives us an upper bound $u_t(\theta^t) \le \bar{\bar{u}}_t(\theta^t)$ for all t, θ^t. Let

$$\bar{\bar{v}}_t = \mathbb{E}_t\left[\sum_{s=1}^{\infty}\beta^{s-1}\theta_{t+s}\bar{\bar{u}}_{t+s}(\theta^{t+s})\right].$$

If $\bar{u} < \infty$, we have $\bar{\bar{v}}_t < \dfrac{\bar{u}}{1-\beta} = \bar{v}$. If $\bar{u} = \bar{v} = \infty$, then we can write

$$\bar{\bar{v}}_t \le \max_{\Theta}\theta \times \sum_{s=1}^{\infty}\beta^{s-1}\left\{\sum_{\theta^{t+s}}\pi_{t+s}(\theta^{t+s})U\left(\frac{e}{\pi_{t+s}(\theta^{t+s})}\right)\right\} \le \theta_{(|\Theta|)}\sum_{s=1}^{\infty}\beta^{s-1}U(|\Theta|e) < \infty,$$

where the second inequality follows from the concavity of U. Therefore we have $v_t(\theta^t) \le \bar{\bar{v}}_t < \bar{v}$, for all t, θ^t. Next, if $\underline{u} > -\infty$, let $\underline{\underline{u}}_t(\theta^t) = \underline{u}$ for all t, θ^t. Now suppose instead that $\underline{u} = -\infty$. We have $\mathbb{E}_0\left[\sum_{t=1}^{\infty}\beta^{t-1}\theta_t u_t(\theta^t)\right]$ diverges toward $-\infty$ when $\beta^{s-1}\theta_s u_s(\theta^s) \to -\infty$ for some (s, θ^s), because

$$\begin{aligned}\sum_{t=1}^{\infty}\beta^{t-1}\sum_{\theta^t \in \Theta^t \setminus \{\theta^s\}}\pi_t(\theta^t)\theta_t u_t(\theta^t) &\le \sum_{t=1}^{\infty}\beta^{t-1}\sum_{\theta^t \in \Theta^t \setminus \{\theta^s\}}\pi_t(\theta^t)\theta_t \bar{\bar{u}}_t(\theta^t)\\ &= \bar{\bar{v}}_0 - \beta^{s-1}\pi_s(\theta^s)\theta_s\bar{\bar{u}}_s(\theta^s) < \infty.\end{aligned}$$

Thus, if $u_s(\theta^s)$ is small enough, the allocation is dominated by $\tilde{u}_t(\theta^t) = C^{-1}(e)$ for all (t, θ^t). Hence for each (t, θ^t) we have a lower bound $u_t(\theta^t) \ge \underline{\underline{u}}_t(\theta^t) > -\infty$. Similarly we have $v_t(\theta^t) \ge \underline{\underline{v}}_t > -\infty$, where

$$\underline{\underline{v}}_t = \mathbb{E}_t\left[\sum_{s=1}^{\infty}\beta^{s-1}\theta_{t+s}C^{-1}(e)\right] = \frac{C^{-1}(e)}{1-\beta}.$$

Therefore, defining $\underline{\underline{u}}_t \equiv \min_{\Theta^t}\underline{\underline{u}}_t(\theta^t)$, $\underline{\underline{v}}_t \equiv \min_{\Theta^t}\underline{\underline{v}}_t(\theta^t)$, $\bar{\bar{u}}_t \equiv \max_{\Theta^t}\bar{\bar{u}}_t(\theta^t)$ and $\bar{\bar{v}}_t \equiv \max_{\Theta^t}\bar{\bar{v}}_t(\theta^t)$, we have shown that we can impose the additional constraints

$$\begin{aligned}\underline{\underline{u}}_t &\le u_t(\theta^t) \le \bar{\bar{u}}_t,\\ \underline{\underline{v}}_t &\le v_t(\theta^t) \le \bar{\bar{v}}_t,\end{aligned}$$

for all t, θ^t.

Next, we truncate the economy to $T < \infty$ periods and allow the planner to provide incentives in the last period "for free." That is, we define

$$V^T(e) = \sup_{\substack{u_t(\theta^t) \in \left[\underline{\underline{u}}_t, \bar{\bar{u}}_t\right] \\ v_t(\theta^t) \in \left[\underline{\underline{v}}_t, \bar{\bar{v}}_t\right]}} \mathbb{E}_0\left[\sum_{t=1}^{T}\beta^{t-1}\theta_t u_t(\theta^t)\right]$$

subject to the promise-keeping constraints

$$v_t(\theta^t) = \sum_{\theta \in \Theta} \pi(\theta)[\theta u_{t+1}(\theta^t, \theta) + \beta v_{t+1}(\theta^t, \theta)], \forall t \leq T-1,$$

the incentive-compatibility constraints

$$\theta u_t\left(\theta^{t-1}, \theta\right) + \beta v_t\left(\theta^{t-1}, \theta\right) \geq \theta u_t\left(\theta^{t-1}, \hat{\theta}\right) + \beta v_t\left(\theta^{t-1}, \hat{\theta}\right), \forall t \leq T,$$

and the feasibility constraints

$$\mathbb{E}_0[C(u_t)] \leq e, \quad \forall t \leq T.$$

Note that the last incentive constraint is:

$$\theta u_T\left(\theta^{T-1}, \theta\right) + \beta v_T\left(\theta^{T-1}, \theta\right) \geq \theta u_T\left(\theta^{T-1}, \hat{\theta}\right) + \beta v_T\left(\theta^{T-1}, \hat{\theta}\right) \text{ for all } \theta^{T-1}, \hat{\theta},$$

and the last two promise-keeping constraints are:

$$v_{T-1}\left(\theta^{T-1}\right) = \sum_{\theta \in \Theta} \pi(\theta)\left[\theta u_T\left(\theta^{T-1}, \theta\right) + \beta v_T\left(\theta^{T-1}, \theta\right)\right] \quad \text{and} \quad \underline{\underline{v}}_T \leq v_T\left(\theta^T\right) \leq \bar{\bar{v}}_T,$$

that is, the promise in period T has no resource cost. In the truncated problem we maximize a continuous function over a compact set, namely $\prod_{\substack{1 \leq t \leq T \\ \theta^t \in \Theta^t}} \left[\underline{\underline{u}}_t, \bar{\bar{u}}_t\right] \times \left[\underline{\underline{v}}_t, \bar{\bar{v}}_t\right]$, so a maximum exists (which is, in fact, unique, since the objective is strictly convex). Call this maximum $\left(\mathbf{u}^T, \mathbf{v}^T\right) = \left\{u_t^T(\theta^t); v_t^T(\theta^t)\right\}_{t, \theta^t}$.

We now show that $\lim_{T \to \infty}\left(\mathbf{u}^T, \mathbf{v}^T\right)$ achieves the maximum of the original problem. By definition of a supremum, for any $\varepsilon > 0$ we can find an incentive-compatible and feasible allocation $\left(\tilde{\mathbf{u}}, \tilde{\mathbf{v}}\right)$ for the original problem such that

$$\mathbb{E}_0\left[\sum_{t=1}^{\infty} \beta^{t-1} \theta_t \tilde{u}_t(\theta^t)\right] > V(e) - \varepsilon.$$

(Note that the r.h.s. is finite.) The truncation at T periods satisfies all the constraints of the truncated economy, so

$$V^T(e) \geq \mathbb{E}_0\left[\sum_{t=1}^{T} \beta^{t-1} \theta_t \tilde{u}_t(\theta^t)\right], \forall T \geq 1.$$

Hence

$$\lim \inf_{T \to \infty} V^T(e) \geq \mathbb{E}_0\left[\sum_{t=1}^{\infty} \beta^{t-1} \theta_t \tilde{u}_t(\theta^t)\right].$$

Since ε is arbitrary,

$$\lim_{T\to\infty} \inf V^T(e) \geq V(e).$$

To show the reverse inequality, fix $t \geq 1$. For all $T \geq t$, $\left(u_t^T(\theta^t), v_t^T(\theta^t)\right) \in \left[\underline{\underline{u}}_t, \bar{\bar{u}}_t\right] \times \left[\underline{\underline{v}}_t, \bar{\bar{v}}_t\right]$. Thus, the sequences $\left\{u_t^T(\theta^t)\right\}_{T\geq t}$ and $\left\{v_t^T(\theta^t)\right\}_{T\geq t}$ must have convergent subsequences as $T \to \infty$. We can then use a diagonal procedure to obtain an incentive-compatible and feasible allocation $\left\{\left(u_t^\infty(\theta^t), v_t^\infty(\theta^t)\right)\right\}_{t\geq 1, \theta^t \in \Theta^t}$, as follows. Arrange states as

$$\mathcal{R} = \left\{\theta_{(1)}, \ldots, \theta_{(|\Theta|)}, \left(\theta_{(1)}, \theta_{(1)}\right), \ldots, \left(\theta_{(1)}, \theta_{(|\Theta|)}\right), \ldots\right\}.$$

Choose a subsequence of u^T, v^T so that the first element converges, ie,

$$\lim_{T\to\infty} \left(u_1^T\left(\theta_{(1)}\right), v_1^T\left(\theta_{(1)}\right)\right) = \left(u_1^\infty\left(\theta_{(1)}\right), v_1^\infty\left(\theta_{(1)}\right)\right).$$

From that subsequence choose another subsequence so that the second element converges, ie,

$$\lim_{T\to\infty} \left(u_1^T\left(\theta_{(2)}\right), v_1^T\left(\theta_{(2)}\right)\right) = \left(u_1^\infty\left(\theta_{(2)}\right), v_1^\infty\left(\theta_{(2)}\right)\right).$$

Repeat the procedure to get (u^∞, v^∞), and call the final subsequence $\{T_n\}_{n\geq 0}$. Since for each $t \leq T, \theta^t \in \Theta^t$, $\left(u_t^T(\theta^t), v_t^T(\theta^t)\right)$ lie in a closed set defined by the incentive constraints, $\left(u_t^\infty(\theta^t), v_t^\infty(\theta^t)\right)$ also lie in the same set, ie, they are incentive compatible. Since $C(u)$ is continuous on $\left[\underline{\underline{u}}_t, \bar{\bar{u}}_t\right]$, $C^T(\theta^t) \equiv C(u_t^T(\theta^t))$ (and $C^T(\theta^t) = 0$ for $t \geq T$) converges pointwise,

$$\lim_{T_n\to\infty} C^{T_n}(\theta^t) = C^\infty(\theta^t) \in \left[C\left(\underline{\underline{u}}_t\right), C(\bar{\bar{u}}_t)\right].$$

Now we can think of $\left\{\pi_1\left(\theta_{(1)}\right), \ldots, \pi_1\left(\theta_{(|\Theta|)}\right), \beta\pi_2\left(\theta_{(1)}, \theta_{(1)}\right), ..\right\}$ as a measure on $\mathcal{R}$. For all $t \geq 1$, $\left\{\theta_t u_t^{T_n}(\theta^t)\right\}_{n\geq 1}$ is a sequence of positive measurable functions on that space that converges pointwise to $\theta_t u_t^\infty(\theta^t)$ as $n \to \infty$. By Fatou's lemma (Lemma 7.9 in Stokey et al., 1989) $\theta_t u_t^\infty(\theta^t)$ is also measurable, and

$$\lim_{n\to\infty} \sup \sum_{t=1}^{T_n} \sum_{\theta^t \in \Theta^t} \beta^{t-1} \pi_t(\theta_t) \theta_t u_t^{T_n}(\theta^t) \leq \sum_{t=1}^{\infty} \sum_{\theta^t \in \Theta^t} \beta^{t-1} \pi_t(\theta_t) \theta_t u_t^\infty(\theta^t) \leq V(e),$$

where the last inequality follows from the fact that $\left\{u_t^\infty(\theta^t)\right\}$ satisfies the constraints of problem (9), but may not maximize the objective. Therefore, we obtain

$$\lim_{n\to\infty} \sup V^{T_n}(e) \leq V(e).$$

We therefore showed that $\lim_{n\to\infty} V^{T_n}(e)$ exists and

$$V(e) = \lim_{n\to\infty} V^{T_n}(e).$$

Moreover, we showed that a maximum of $V(e)$ is achieved by the limit of the sequence $\left(\mathbf{u}^{T_n}, \mathbf{v}^{T_n}\right)$. This concludes the proof. □

3.1.3 Application: Sustainability Constraints

Suppose that in addition to constraint (82) we further impose a constraint that social welfare in any period cannot drop below a threshold $\underline{U}$,

$$\mathbb{E}_0\left[\sum_{s=1}^{\infty}\beta^{s-1}\theta_{t+s}u_{t+s}\right] \geq \underline{U}, \forall t \geq 1. \tag{88}$$

Such constraints naturally arise in various settings with imperfect commitment, participation constraints, etc. We discuss an example of those in Section 4.4 in the context of an international finance model, where they capture the need to provide incentives for the agents to stick to the contract rather than defaulting and reverting to their outside option (in that case, the value of autarky).[u] We add this constraint (88) to problem (83) and assume that the utility function is bounded.

As before, we have for all t,

$$\lim_{T\to\infty}\mathbb{E}_0\left[\sum_{s=1}^{\infty}\beta^{s-1}\theta_{t+s}x_t^T(\mathbf{u},\mathbf{u}')\right] = \mathbb{E}_0\left[\sum_{s=1}^{\infty}\beta^{s-1}\theta_{t+s}u_{t+s}\right],$$

since the utility is bounded; and for t sufficiently large holding T fixed, we have

$$\mathbb{E}_0\left[\sum_{s=1}^{\infty}\beta^{s-1}\theta_{t+s}x_t^T(\mathbf{u},\mathbf{u}')\right] = \mathbb{E}_0\left[\sum_{s=1}^{\infty}\beta^{s-1}\theta_{t+s}u'_{t+s}\right],$$

which verifies Assumptions 5.ii.a and 5.ii.c for the constraints (88). The other parts of Assumption 5 are verified as before. As long as $\underline{U}$ is not too high, we can find an interior point $\mathbf{x}'$ that satisfies $\mathbf{\Phi}(\mathbf{x}') < 0$. Theorem 5 thus establishes that there exists a nonnegative summable sequence of Lagrange multipliers $\{\mu_t\}_{t=1}^{\infty}$, such that the solution to our problem is also a solution to

$$\max_{\mathbf{u}\in\Gamma}\mathbb{E}_0\left[\sum_{t=1}^{\infty}\beta^{t-1}\theta_t u_t\right] + \sum_{t=1}^{\infty}\mu_t\mathbb{E}_0\left[\sum_{s=1}^{\infty}\beta^{s-1}\theta_{t+s}u_{t+s}\right] - \sum_{t=1}^{\infty}\tilde{\lambda}_t\mathbb{E}_0[C(u_t)].$$

Since $\{\mu_t\}_{t=1}^{\infty}$ is summable, we can rewrite the equation above as

$$\mathbb{E}_0\left[\sum_{t=1}^{\infty}\beta^{t-1}\theta_t u_t + \sum_{t=1}^{\infty}\sum_{s=1}^{\infty}\beta^{s-1}\mu_t\theta_{t+s}u_{t+s} - \sum_{t=1}^{\infty}\tilde{\lambda}_t C(u_t)\right] = \mathbb{E}_0\left[\sum_{t=1}^{\infty}\overline{\beta}_t\{\theta_t u_t - \lambda_t C(u_t)\}\right], \tag{89}$$

[u] Such constraints would also appear in models of political economy in which a government is tempted to reoptimize (see, eg, Acemoglu et al., 2008; Sleet and Yeltekin, 2008; Farhi et al., 2012).

where $\overline{\beta}_t = \beta^{t-1} + \mu_1\beta^{t-2} + \cdots + \mu_{t-2}\beta + \mu_{t-1}$, letting $\mu_0 = 0$, with $\sum_{t=1}^{\infty}\overline{\beta}_t < \infty$ and $\lambda_t = \widetilde{\lambda}_t/\overline{\beta}_t$. This problem can be solved using our usual techniques. Augmenting the problem with a promise-keeping constraint, we can replace $\mathbf{u} \in \Gamma$ with $(\mathbf{u},\mathbf{v}) \in \Gamma(v_0)$ and observe that the problem can be written recursively, letting $\hat{\beta}_{t+1} = \overline{\beta}_{t+1}/\overline{\beta}_t$, as

$$\begin{aligned} k_t(v) = & \max_{\{u(\theta),w(\theta)\}_{\theta\in\Theta}} & \mathbb{E}\left[\theta u - \lambda_t C(u) + \hat{\beta}_{t+1}k_{t+1}(w)\right] \\ & \text{subject to} & (19),(20). \end{aligned}$$

We can extend Lemma 1 directly to k_t, with the only exception that k_t is not strictly decreasing but rather inversely U-shaped. (We can easily show that $v \mapsto k_t(v)$ is concave, continuous, and satisfies $\lim_{v\to\bar{v}} k_t(v) = -\infty$ and $\lim_{v\to\underline{v}} k_t(v) = \underline{v}$.) Much of the analysis in Section 2.4 continues to hold but the condition (29) now becomes

$$k_t'(v) = \frac{\hat{\beta}_{t+1}}{\beta}\mathbb{E}\left[k_{t+1}'(w_v)\right]. \tag{90}$$

Observe that $\hat{\beta}_{t+1} \geq \beta$ with strict inequality if $\mu_t > 0$. Therefore the marginal cost k_t' is no longer a martingale if constraint (88) binds, which implies a form of mean reversion. To see this, observe that since k is inversely U-shaped, k_t' is positive for low v and negative for high v. Therefore Eq. (90) shows that the marginal cost decreases in expectation if v is low (because then $\mathbb{E}\left[k_{t+1}'(w_v)\right] = \frac{\beta}{\hat{\beta}_{t+1}}k_t'(v) \leq k_t'(v)$) and increases if v is high (because then $\mathbb{E}\left[k_{t+1}'(w_v)\right] = \frac{\beta}{\hat{\beta}_{t+1}}k_t'(v) \geq k_t'(v)$).

3.1.4 Using Lagrange Multipliers Instead of Promised Utilities

In our discussion so far we have used the following technique to solve dynamic incentive problems: we formed a Lagrangian using all the constraints except the incentive constraints, and then introduced the promised utilities in order to write the incentive constraints recursively. In principle, there is nothing special about the incentive constraints per se: we could extend the Lagrangian to those constraints also, eliminating the need to use promised utilities at all. Here we describe how this can be done, using a version of our benchmark partial equilibrium model of Section 2.3.2 as an example.

Consider the maximization problem

$$\begin{aligned} K(v_0) \equiv \max_{\mathbf{u}} \mathbb{E}\left[-\sum_{t=1}^{\infty}\beta^{t-1}C(u_t)\right] \\ \text{subject to } (10)\ \forall t \geq 1, \text{ and } (14), \end{aligned} \tag{91}$$

that we analyzed in Section 2.3.2. Since the objective function is strictly concave and the constraint set is convex, its solution $\hat{\mathbf{u}}$ is unique. Define $W : \mathbb{R} \to \mathbb{R}$ as

$$W(\alpha) \equiv \max_{\mathbf{u}} \mathbb{E}\left[\sum_{t=1}^{\infty} \beta^{t-1}(\alpha\theta_t u_t - C(u_t))\right] \tag{92}$$
$$\text{subject to } (10)\ \forall t \geq 1.$$

If $\hat{\alpha}$ is the Lagrange multiplier on constraint (14), then $W(\hat{\alpha})$ is simply the Lagrangian associated with problem (91), whose unique maximum (the objective is strictly concave and the constraints are linear) is attained at $\hat{\mathbf{u}}$ by Theorem 4.[v]

We now show how applying Corollary 1 leads to a recursive characterization of this problem, using different techniques than those described in Section 2.3.2. We then discuss the strengths and weaknesses of these two alternative approaches. For simplicity we assume that $|\Theta| = 2$; the analysis extends straightforwardly to any number of shocks.

When $|\Theta| = 2$ there are two incentive constraints (10) in period 1 corresponding to shocks $\theta_{(1)}$ and $\theta_{(2)}$. Adapting the arguments of Proposition 5 shows that the constraint of type $\theta_{(2)}$ is slack. Let $\hat{\xi}(\theta_{(1)})$ be the Lagrange multiplier on the first-period incentive constraint of type $\theta_{(1)}$ in problem (92). By Corollary 1 (written for a maximization rather than minimization problem),[w] $\hat{\xi}(\theta_{(1)})$ and the solution to (92) are also the solution to

$$\begin{aligned}
W(\alpha) \equiv \min_{\xi \geq 0} \ \max_{\mathbf{u}} \ & \mathbb{E}\left[\sum_{t=1}^{\infty} \beta^{t-1}(\alpha\theta_t u_t - C(u_t))\right] \\
& + \xi\left[\ \left\{\theta_{(1)} u_1(\theta_{(1)}) + \mathbb{E}\left[\sum_{t=2}^{\infty} \beta^{t-1}\theta_t u_t \middle| \theta_1 = \theta_{(1)}\right]\right\}\right. \\
& \left. - \left\{\theta_{(1)} u_1(\theta_{(2)}) + \mathbb{E}\left[\sum_{t=2}^{\infty} \beta^{t-1}\theta_t u_t \middle| \theta_1 = \theta_{(2)}\right]\right\}\right] \\
& \text{subject to } (10)\ \forall t \geq 2.
\end{aligned}$$

Rearrange these terms and use the definition of W to obtain

$$\begin{aligned}
W(\alpha) \equiv \min_{\xi \geq 0} \ \max_{u(\theta_{(1)}),\, u(\theta_{(2)})} \\
\pi(\theta_{(1)}) & \left[\left(\alpha\theta_{(1)} + \frac{\xi\theta_{(1)}}{\pi(\theta_{(1)})}\right) u_1(\theta_{(1)}) - C(u_1(\theta_{(1)})) + \beta W\left(\alpha + \frac{\xi}{\pi(\theta_{(1)})}\right)\right] \\
+ \pi(\theta_{(2)}) & \left[\left(\alpha\theta_{(2)} - \frac{\xi\theta_{(1)}}{\pi(\theta_{(2)})}\right) u_1(\theta_{(2)}) - C(u_1(\theta_{(2)})) + \beta W\left(\alpha - \frac{\xi}{\pi(\theta_{(2)})}\right)\right].
\end{aligned} \tag{93}$$

[v] $\hat{\alpha}$ can be found from a max min problem as in Corollary 1 by observing that by monotonicity we can replace the equality constraint (14) in problem (91) with the inequality constraint $\mathbb{E}\left[\sum_{t=1}^{\infty} \beta^{t-1}\theta_t u_t(\theta^t)\right] \geq v_0$.

[w] We can verify the sufficient conditions allowing us to apply Corollary 1 using the same steps as in Section 3.1.2.

Problem (93) is an alternative way to characterize the solution to the maximization problem (91). The function W can be found using standard contraction mapping techniques (see Marcet and Marimon, 2015 for proofs). The policy functions to this Bellman equation can then be used to generate the solution $\hat{\mathbf{u}}$ the same way we did in Section 2.3.2.

We conclude this section by comparing the two alternative recursive formulations (23) and (93). On the one hand, the max operator in (23) is simpler to handle than the min max operator in (93).[x] This makes (23) easier to use in many simple applications. On the other hand, the function W is defined over an a priori known domain, $\mathbb{R}$, while the domain of K is endogenous. We could easily characterize the latter in the setup of Section 2.3.2 (see Footnote g). In more general settings, however (with Markov shocks, additional constraints, etc.), characterizing the state space is more difficult and requires using the techniques of Abreu et al. (1990) (see Proposition 8), so that using the tools described in this section can be simpler. An in-depth discussion of this approach is outside the scope of this chapter and we refer the interested reader to the papers that describe it in more detail. The pioneering work that first developed this approach is Marcet and Marimon (2015). The more recent applications are Messner et al. (2012, 2014), Cole and Kubler (2012), and Espino et al. (2013).

3.2 Mechanism Design Without Commitment

In our discussion so far we assumed that the principal, which provides insurance to agents, has perfect commitment: it implicitly promises a menu of allocations for infinitely many periods and never entertains the possibility of reneging on those promises as time goes by. This assumption was critical in proving the Revelation Principle in Theorem 1. However, this assumption is not innocuous. For example, we saw in Proposition 6 that long-run immiseration is a common feature of the optimal insurance contracts. While such a contract is optimal ex ante in period 0, it provides the worst possible allocation in the long run. Any benevolent principal would like to reoptimize at that point. Thus the assumption that the principal has perfect commitment is very strong in many applications.

In this section we discuss several approaches to analyze dynamic contracting problems in environments where the principal cannot commit. We start with the set up of Section 2.2 with two modifications. First, we assume that insurance in that economy

[x] Many insights can be obtained from problem (93) without considering the min part. Since $\hat{\xi}\left(\theta_{(1)}\right) \geq 0$ (in fact, with a strict inequality from the discussion in Proposition 5), problem (93) immediately shows that the weight $\alpha\left(\theta_{(1)}\right) \equiv \hat{\alpha} + \dfrac{\hat{\xi}\left(\theta_{(1)}\right)}{\pi\left(\theta_{(1)}\right)}$ increases for the agent who reports $\theta_{(1)}$, ie, $\alpha\left(\theta_{(1)}\right) \geq \hat{\alpha}$, while the weight $\alpha\left(\theta_{(2)}\right) \equiv \hat{\alpha} - \dfrac{\hat{\xi}\left(\theta_{(1)}\right)}{\pi\left(\theta_{(2)}\right)}$ decreases, ie, $\alpha\left(\theta_{(2)}\right) \leq \hat{\alpha}$. To see the implication of this fact, observe that the solution to (92), $\hat{\mathbf{u}}^{\alpha}$, has the property that $\mathbb{E}\left[\sum_{t=1}^{\infty} \beta^{t-1} \theta_t u_t^{\alpha}\right]$ is increasing in α, so that higher weights correspond to higher lifetime utilities. Therefore we showed, without explicitly considering the min operator, that the expected lifetime utility starting from next period increases for the agent who reports $\theta_{(1)}$, and decreases for the agent who reports $\theta_{(2)}$. See Acemoglu et al. (2011) for another application of this technique.

is provided by a benevolent principal, which we call "the government," that cannot commit ex ante, in period 0, to its future actions. Second, in order to focus on information revelation and various generalizations of Theorem 1 we abstract from borrowing and lending and assume that the total consumption of all agents should be equal to the total endowment e in each period, as in Section 3.1.2.

Since the government cannot commit, we formally describe the environment as an infinitely repeated game between one large player (the government) and a continuum of atomistic agents.[y] Each period of the game is divided into two stages. In the first stage agents report information about the realization of their idiosyncratic shock to the government, and in the second stage the government chooses allocations.[z] As in Section 2.2, it is helpful to start by describing communication between the agent and the principal using a general message space M.

Agents' reporting strategies in period t are maps $\tilde{\sigma}_t : M^{t-1} \times \Theta^t \times H^t \rightarrow \Delta(M)$, and the government's strategy is a map $\tilde{c}_t : M^t \times \breve{H}^t \rightarrow \Delta(\mathbb{R}_+)$, where M^{t-1} and Θ^t are the histories of reports and the realizations of shocks for each agent, and H^t and $\breve{H}^t$ (described below) are the aggregate histories of the game. To avoid complicating our discussion with measure-theoretic apparatus, we assume that $\Delta(M)$ and $\Delta(\mathbb{R}_+)$ only randomize between finitely many elements. The assumption of a continuum of agents simplifies the analysis. By the law of large numbers, $\tilde{\boldsymbol{\sigma}}$ generates the aggregate distribution of reports that the government receives from the agents, and $\tilde{\boldsymbol{c}}$ generates the distribution of consumption allocations provided by the government. Moreover, these distributions are not affected if an individual agent (who is of measure zero) deviates from his equilibrium strategy. Assuming that these aggregate distributions are observable history H^t consists of the aggregate distributions generated by $\tilde{\boldsymbol{\sigma}}$ and $\tilde{\boldsymbol{c}}$ up to period $t-1$, while $\breve{H}^t$ consists of H^t and the distribution of aggregate reports generated by $\tilde{\sigma}_t$.

We describe how to characterize the perfect Bayesian equilibrium (PBE) of this game that delivers the highest ex ante utility to agents. Well-known arguments (see Chari and Kehoe, 1990 or a textbook treatment in Chapter 23 of Ljungqvist and Sargent, 2012 for details) imply that to characterize such an equilibrium it is sufficient to focus only on a subset of the histories of the game. Namely, it is sufficient to characterize the reporting done by the agents and allocations provided by the government "on the equilibrium path." If the government ever deviates from the equilibrium path distribution of allocations (up to a measure zero) then in subsequent histories agents and the government switch to the worst PBE. With a slight abuse of notation, we use $\left(\tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{c}}\right)$ to describe

[y] See Chari and Kehoe (1990, 1993) for classic references.

[z] Although this set up appears a bit stylized, many of its features emerge naturally in more sophisticated models of political economy. For example, models in which policies are chosen via probabilistic voting à la Lindbeck and Weibull (1987) in each period often reduce to our set up with a benevolent government that cannot commit to its future actions. See Farhi et al. (2012), Scheuer and Wolitzky (2014), or Dovis et al. (2015) for applications.

the behavior of agents and the government "on the equilibrium path," ie, the mappings $\tilde{\sigma}_t : M^{t-1} \times \Theta^t \rightarrow \Delta(M)$ and $\tilde{c}_t : M^t \rightarrow \Delta(\mathbb{R}_+)$ which no longer have the aggregate histories $H^t, \breve{H}^t$ as arguments. We use $\tilde{\sigma}_t\left(m \middle| m^{t-1}, \theta^t\right)$ to denote the probability that an agent with history $\left(m^{t-1}, \theta^t\right)$ reports the message m in period t.

The pair $\left(\tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{c}}\right)$ must satisfy three constraints. First, in equilibrium each individual agent finds it optimal to stick to his reporting strategy $\tilde{\boldsymbol{\sigma}}$ rather than to deviate to any other reporting strategy $\tilde{\boldsymbol{\sigma}}'$, so that the constraint (4) is satisfied. Note that to write this constraint we implicitly used the assumption of a continuum of agents. If an individual agent chooses $\tilde{\boldsymbol{\sigma}}'$ rather than $\tilde{\boldsymbol{\sigma}}$, the aggregate distribution of reports to the government remains unchanged and therefore the equilibrium allocations remain the same. Thus the same $\tilde{\boldsymbol{c}}$ appears on both sides of the incentive constraint. Second, any allocation that the government chooses must also be feasible, ie, satisfy

$$\mathbb{E}^{\tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{\sigma}}}[c_t] \leq e, \quad \forall t. \tag{94}$$

Third, the government should not find it optimal to deviate from its equilibrium play at any point of time. This constraint can be written as

$$\mathbb{E}_t^{\tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{\sigma}}}\left[\sum_{s=t}^{\infty} \beta^{s-t} \theta_s U(c_s)\right] \geq \tilde{W}_t\left(\left\{\tilde{\sigma}_s\right\}_{s=1}^t\right) + \frac{\beta}{1-\beta} U(e), \quad \forall t. \tag{95}$$

The left hand side of this constraint is the government's payoff from continuing to play its equilibrium strategy in period t. The right hand side consists of two parts: the value of the best one-time deviation $\tilde{W}_t$ (to be defined below) followed by the value of the worst PBE starting from the next period. Since we assumed that shocks are i.i.d., it is easy to show that the worst PBE is such that agents reveal no information to the government and receive forever the same per capita allocation e independently of the shock. The expected value of this allocation is $\frac{1}{1-\beta} U(e)$.

We now derive the value of deviation $\tilde{W}_t$. Let $\mu_t(m^t)$ denote the measure of agents who report history m^t. It is defined recursively as $\mu_{-1} = 1$ and

$$\mu_t(m^t) = \mu_{t-1}\left(m^{t-1}\right) \sum_{\theta^t \in \Theta^t} \pi(\theta^t) \tilde{\sigma}_t\left(m_t \middle| m^{t-1}, \theta^t\right).$$

The measure μ_t depends on the entire history of reports up to period $t, \left\{\tilde{\sigma}_s\right\}_{s=1}^t$. We use $\mathbb{E}^{\tilde{\boldsymbol{\sigma}}}[\theta | m^t]$ to denote the government's posterior expectation of an agent's type being θ, conditional on the history of reports m^t. The best deviation solves

$$\tilde{W}_t\left(\left\{\tilde{\sigma}_s\right\}_{s=1}^t\right) = \max_{\left\{c^w(m^t)\right\}_{m^t \in M^t}} \sum_{m^t \in M^t} \mu_t(m^t)\left[\mathbb{E}^{\tilde{\boldsymbol{\sigma}}}[\theta | m^t] U(c^w(m^t))\right] \tag{96}$$

subject to the feasibility constraint

$$\sum_{m^t \in M^t} \mu_t(m^t) c^w(m^t) \leq e. \tag{97}$$

At this stage of our discussion it is useful to compare our set up to that with commitment in Section 2.2. Relative to the environment in that section, we have one additional constraint, (95). The important feature of this constraint is that posterior beliefs appear on both sides of this constraint. This changes the analysis. In particular, note that the proof of Theorem 1 does not need to go through when constraint (95) is imposed. If we replace $\widetilde{\Gamma} = \left(M, \widetilde{\boldsymbol{c}} \circ \widetilde{\boldsymbol{\sigma}}\right)$ with a direct truthful mechanism $\left(\Theta, \boldsymbol{c} \circ \boldsymbol{\sigma}^{truth}\right)$, we still obtain feasible and incentive-compatible allocations for all agents as in the proof of Theorem 1. However we have $\widetilde{W}_t\left(\left\{\sigma_s^{truth}\right\}_{s=1}^t\right) \geq \widetilde{W}_t\left(\left\{\widetilde{\sigma}_s\right\}_{s=1}^t\right)$, generally with a strict inequality, since a direct mechanism reveals more precise information to the government and increases its incentives to deviate. Since by construction we have $\mathbb{E}^{\widetilde{\boldsymbol{c}} \circ \widetilde{\boldsymbol{\sigma}}}\left[\sum_{s=t}^{\infty} \beta^{s-t} \theta_s U(c_s)\right] = \mathbb{E}^{\boldsymbol{c} \circ \boldsymbol{\sigma}^{truth}}\left[\sum_{s=t}^{\infty} \beta^{s-t} \theta_s U(c_s)\right]$, the direct truthtelling mechanism tightens the sustainability constraint of the government. Intuitively, this mechanism always reveals more information to the government than any other communication mode, increasing the gains for the government from ex post reoptimization and lowering ex ante welfare.

The discussion in the previous paragraph implies that it is generally *not* without loss of generality to restrict attention to mechanisms in which agents report their type directly to the government, as we did in Section 2.2, and that one needs to work with more general message spaces to characterize the optimal insurance in this setting. Here we outline how it can be done. Our discussion is based on Golosov and Iovino (2014); for more detailed discussion and proofs we refer the reader to that paper.[aa]

To find the optimal insurance without commitment, the best PBE solves

$$\max_{\widetilde{\boldsymbol{c}}, \widetilde{\boldsymbol{\sigma}}} \mathbb{E}^{\widetilde{\boldsymbol{c}} \circ \widetilde{\boldsymbol{\sigma}}}\left[\sum_{t=1}^{\infty} \beta^{t-1} \theta_t U(c_t)\right] \tag{98}$$

subject to (4), (94), and (95). Under some technical conditions this problem can be significantly simplified. In particular with i.i.d. shocks the history of past realizations of shocks is irrelevant and we can simply restrict attention to reporting strategies of the form $\widetilde{\sigma}_t : M^{t-1} \times \Theta_t \rightarrow \Delta(M)$. Similarly, one can also show the analogue of Proposition 1 that

[aa] Formally, Golosov and Iovino (2014) study a slightly more general game that allows agents' and government's strategies to depend on the realization of payoff-irrelevant variables. This convexifies the set of equilibrium payoffs and ensures that some technical conditions simplifying the analysis hold. To streamline the exposition we simply assume that those conditions are satisfied.

stochastic allocations of consumption are suboptimal, so that we can assume without loss of generality that $\tilde{c}_t : M^t \to \mathbb{R}_+$. Finally without loss of generality we can restrict M to a finite set.[ab]

We now show how to write this problem recursively. As in Section 2, it is more convenient to change variables and optimize with respect to $u_t = U(\tilde{c}_t)$, and constraint (4) simplifies if we use a one-shot deviation principle. Using the same arguments as those leading to Eq. (13), we can rewrite (4) as: for all $m^t \in M^t$,

$$v_t(m^t) = \sum_{(\theta,m)\in\Theta\times M} \pi(\theta)\tilde{\sigma}_{t+1}(m|m^t,\theta)[\theta u_{t+1}(m^t,m) + \beta v_{t+1}(m^t,m)], \tag{99}$$

and for all $(m^t,\theta) \in M^t \times \Theta$, for all $m \in M$, and some $m_\theta \in M$,

$$\theta u_{t+1}(m^t,m_\theta) + \beta v_{t+1}(m^t,m_\theta) \geq \theta u_{t+1}(m^t,m) + \beta v_{t+1}(m^t,m) \tag{100}$$

with for all $m \in M$,

$$\tilde{\sigma}_{t+1}(m|m^t,\theta)[\{\theta u_{t+1}(m^t,m_\theta) + \beta v_{t+1}(m^t,m_\theta)\} - \{\theta u_{t+1}(m^t,m) + \beta v_{t+1}(m^t,m)\}] = 0. \tag{101}$$

Eq. (99) is simply a generalization of (12) to the setting in which agents reveal noisy information to the government. The next two equations form the incentive-compatibility conditions. Eq. (100) says that there must be some message m_θ that agent θ prefers to all others, given the past history of messages m^t and shock realization θ. Eq. (101) says that if an agent with current shock realization θ reports any message m other than m_θ with positive probability $\tilde{\sigma}_{t+1}(m|m^t,\theta)$, then he must be indifferent between reporting m and m_θ, since any report he sends must give him the highest utility. Eqs. (100) and (101) are a generalization of (13) and have a recursive structure, with $v_t(m^t)$ playing the role of the state variable.

We now show how to write the problem of maximizing (98) subject to (94), (95), and (99)–(101) recursively using the Lagrangian techniques introduced in Section 3.1.2. Let $\tilde{\boldsymbol{\lambda}} = \left\{\tilde{\lambda}_t\right\}_{t=1}^{\infty}$ and $\tilde{\boldsymbol{\chi}} = \{\tilde{\chi}_t\}_{t=1}^{\infty}$ be sequences of multipliers on the constraints (94) and (95), respectively. Assuming that these sequences are summable (see Section 3.1.1), we can write the Lagrangian, using Abel's formula, as[ac]

$$\max_{\mathbf{u},\tilde{\sigma}} \mathbb{E}^{\tilde{\sigma}} \sum_{t=1}^{\infty} \overline{\beta}_t \left[\theta_t u_t - \lambda_t C(u_t) - \chi_t \tilde{W}_t\right] \tag{102}$$

subject to (99)–(101), where $\overline{\beta}_t = \beta^{t-1} + \sum_{s=1}^{t} \beta^{t-s}\tilde{\chi}_s$, $\lambda_t = \tilde{\lambda}_t / \overline{\beta}_t$, and $\chi_t = \tilde{\chi}_t / \overline{\beta}_t$. Note that this problem is very similar to the problem considered in Section 3.1.3, except that

[ab] Specifically, the cardinality of M can be taken to be $2|\Theta| - 1$.

[ac] Since consumption allocations are deterministic, we write $\mathbb{E}^{\tilde{\sigma}*}$ rather than $\mathbb{E}^{\mathbf{u}\circ\tilde{\sigma}*}$.

now we choose the optimal amount of information that is revealed to the government, $\tilde{\boldsymbol{\sigma}}$, and the costs of information revelation are captured by the terms $\chi_t \widetilde{W}_t$.

This problem still does *not* have a natural recursive form. Our recursive characterization in Section 3.1.2 relied on the fact that the linearity of the objective function allowed us to separately solve for the optimal allocations after any history θ^t (ie, in the setting without commitment, after any history of reports m^t) without paying attention to the other histories. The key difficulty now is that $\widetilde{W}_t$ depends on the distribution of reports that are sent by all agents. We show here how to write a recursive formulation under the assumption that preferences are logarithmic. Golosov and Iovino (2014) use the techniques of Section 3.1.1 to obtain the same characterization for arbitrary concave utility functions.

When preferences are logarithmic, $\widetilde{W}_t$ is easy to simplify. The first-order conditions of problem (102) give

$$\lambda_t^w C'\left(u_t^w(m^t)\right) = \mathbb{E}^{\tilde{\boldsymbol{\sigma}}}[\theta|m^t] = \frac{\tilde{\sigma}_t\,(m_t|m^{t-1},\theta)}{\sum\limits_{\theta'\in\Theta}\pi(\theta')\tilde{\sigma}_t\,\left(m_t|m^{t-1},\theta'\right)},$$

where $u_t^w \equiv U\left(u_t^w\right)$ and λ_t^w is the Lagrange multiplier on constraint (97). With logarithmic preferences, $C' = C = \exp$. Using this fact together with (97) we can easily find that $\lambda_t^w = 1/e$. The key property is that this multiplier does not depend on particular values of $\left\{\tilde{\sigma}_t\right\}_{m^t,\theta^t}$, and therefore $\widetilde{W}_t$ can be written as

$$\widetilde{W}_t\left(\left\{\tilde{\sigma}_s\right\}_{s=1}^t\right) = \sum_{m^{t-1}}\mu_{t-1}\left(m^{t-1}\right)W_t\left(\left\{\tilde{\sigma}_t\left(m|m^{t-1},\theta\right)\right\}_{(m,\theta)\in M\times\Theta}\right),$$

where

$$W_t\left(\left\{\tilde{\sigma}(m|\cdot,\theta)\right\}_{(m,\theta)\in M\times\Theta}\right) = \max_{\{u^w(m)\}_{m\in M}}\ \sum_{(m,\theta)\in M\times\Theta}\pi(\theta)\,\tilde{\sigma}(m|\cdot,\theta)\left[\theta u^w(m) - \lambda_t^w C(u^w(m))\right].$$

If we substitute this equation into (102), we can easily write the problem recursively, letting $\hat{\beta}_{t+1} \equiv \overline{\beta}_{t+1}/\overline{\beta}_t$, as

$$k_t(v) = \max_{\substack{\{u(m),w(m),\sigma(m|\theta)\}_{(m,\theta)\in M\times\Theta} \\ \sigma(\cdot|\theta)\in\Delta(M)}} \mathbb{E}^{\sigma}\left[\theta u - \lambda_t C(u) + \hat{\beta}_{t+1}k_{t+1}(w)\right] - \chi_t W_t\left(\left\{\sigma(m|\theta)\right\}_{m,\theta}\right) \tag{103}$$

subject to: for all θ,

$$v = \sum_{(\theta,m)\in\Theta\times M}\pi(\theta)\sigma(m|\theta)\left[\theta u(m) + \beta w(m)\right],$$

for all m and some m_θ,

$$\theta u(m_\theta)+\beta w(m_\theta)\geq \theta u(m)+\beta w(m),$$

and for all m,

$$\sigma(m|\theta)[\{\theta u(m_\theta)+\beta w(m_\theta)\}-\{\theta u(m)+\beta w(m)\}]=0.$$

Note that problem (103) is very similar to problem (18) in Section 2, with two modifications. First, the objective function has an additional term $-\chi_t W_t$ which captures the additional cost of information revelation off the equilibrium path. Second, agents generally play mixed strategies over the message space M rather than a pure reporting strategy over the set Θ.

Golosov and Iovino (2014) analyze this problem and show that the optimal amount of information that each agent reveals depends on the promised utility v. The key insight of their paper is that the agents who should reveal more information to the government are those for whom such revelation saves the most resources to the government *on the equilibrium path*. In particular, in the set up discussed above, the government loses relatively little resources if it delivers a low value of v without knowing the realization of θ, while information revelation by agents with higher v leads the government to save more resources. Golosov and Iovino (2014) show that for all v sufficiently small agents reveal no information to the government and play the same reporting strategy independently of the realization of their shock; on the other extreme, agents with sufficiently high promise v reveal full information to the government (at least as long as U exhibits decreasing absolute risk aversion) just as in Section 2.2. They show that the government's participation constraints imply the existence of an endogenous lower bound in the invariant distribution below which agents' promised utility never falls, preventing the emergence of long-run immiseration which was obtained in Section 2.4. Golosov and Iovino (2014) further generalize their analysis by considering Markov shocks and obtaining a recursive characterization along the lines of Section 2.5.[ad]

3.2.1 Optimal Insurance with a Mediator

In the game described in the previous section we assumed a particular communication protocol between the agents and the government: agents first report some information to the government, then the government takes some action. In settings where the government could commit, as in Section 2.5, restricting attention to such communication protocols was without loss of generality due to Theorem 1. As we saw, Theorem 1 fails when the government cannot commit. One may wonder if better outcomes can be attained if richer ways to communicate between agents and the government are available.

[ad] The problem of information revelation with persistent shocks is related to the literature on the ratchet effect (see Freixas et al., 1985; Laffont and Tirole, 1988).

The answer to this question turns out to be yes. Here we describe what the optimal communication devices are and how to characterize the optimal contracts in such settings.

Suppose agents and the government can communicate indirectly, using a third party called a "mediator." The mediator can be a trusted third person with no stake in the outcome of the game, or simply a machine that takes reports from the agents and recommends the action to the government as a function of those reports using a predetermined rule. Thus, the game is essentially the same as in the previous section, with the following modification. In each period, the agents first send reports $\tilde{\sigma}_t : M^{t-1} \times \Theta^t \to \Delta(M)$ to the mediator, then the mediator makes recommendations $\tilde{\sigma}_t^{med} : M^t \to \Delta(\mathbb{R}_+)$ to the government about which consumption allocation the government should pick. The government is then free to make any choice it wants.

Studying equilibria in this communication game using a mediator is interesting for the following reason. First, without loss of generality we can restrict attention to direct truth-telling strategies $\boldsymbol{\sigma}^{truth}$ for the agents, as defined in Section 2.2 (and hence we can assume that $\tilde{\sigma}_t^{med}$ is a mapping from Θ^t to $\Delta(\mathbb{R}_+)$). Moreover, with a mediator we can replicate the outcome of any PBE with any other communication device. Thus, we get a version of Theorem 1 for Bayesian Nash equilibria (see Myerson, 1982, 1986 and Mas-Colell et al., 1995 (Sec. 23.D)). Therefore, the equilibrium with a mediator provides an upper bound on what can be achieved using any other communicating device.

We want to make two observations about games with a mediator. First, while without loss of generality we can assume that agents report their types truthfully to the mediator, the mediator generally randomizes to garble the information that the government receives—otherwise the government would be able to learn information perfectly about the agent's type and this mechanism would be equivalent to the direct truthtelling mechanism discussed in the previous section. Second, while any PBE (using arbitrary communicating devices) can be implemented as a PBE in a game with a mediator, the converse is not true. Thus, whether the equilibrium with a mediator provides a reasonable description of the optimal insurance arrangement often depends on the context. For example, many negotiations of the resolutions of conflicts between countries already use mediators, so that this approach may be natural. On the other hand, in many political settings it seems often difficult to introduce an uninterested third party outside of the politician's control, and the approach we described in the previous section may be preferable.

To see how this approach alters the incentive constraints, we consider the analogue of the recursive problem (103). The mediator generally needs to randomize between different allocations that it recommends to the government. For simplicity we assume that the mediator offers finitely many recommendations $m_1,\ldots,m_I$ to the government for each agent's report. The reporting strategies of the mediator are now simply $\tilde{\sigma}_t^{med}\left(m \middle| m^{t-1}, \theta^t\right)$ for $m \in M \equiv \{m_1, \ldots, m_I\}$. Assuming that the one-shot deviation principle holds and that the dependence of period-t strategies on θ^{t-1} is redundant, we can write the agents' incentive constraint as (99) and

$$\begin{aligned}&\sum_m \tilde{\sigma}_{t+1}^{med}(m|m^t,\theta)\left[\theta u\left(m^{t+1}\right)+\beta v_{t+1}\left(m^{t+1}\right)\right]\\ \geq &\sum_m \tilde{\sigma}_{t+1}^{med}(m|m^t,\theta')\left[\theta u\left(m^{t+1}\right)+\beta v_{t+1}\left(m^{t+1}\right)\right],\quad \forall\theta'.\end{aligned} \tag{104}$$

Constraint (104) is weaker than constraints (100) and (101), so that more allocations are incentive compatible when a mediator is used. One way to understand the intuition is as follows. When an agent communicates using a mediator, he has no control over which recommendation the mediator makes to the government. Thus his incentive constraint (104) should hold in expectation, over all the recommendations that the mediator may make. When an agent communicates with the government directly, he would never send any message to the government which is dominated by another message. Therefore his incentive constraint (100), (101) should be satisfied for all the messages sent to the government.

It remains to describe how the government forms posterior beliefs based on the mediator's recommendations. The government's behavior is formally identical to that in (96) except that the value of the best deviation is now simply $\tilde{W}_t\left(\left\{\tilde{\sigma}_s^{med}\right\}_{s=1}^{t}\right)$ rather than $\tilde{W}_t\left(\left\{\tilde{\sigma}_s\right\}_{s=1}^{t}\right)$, so that the government uses the mediator's recommendations described by $\tilde{\boldsymbol{\sigma}}^{med}$ rather than agents' reports $\tilde{\boldsymbol{\sigma}}$ to form its posterior beliefs. Nevertheless the mathematical structure of the two problems is identical and we obtain a similar recursive representation as in (103), except that the incentive constraints are replaced by: for all θ',

$$\sum_m \sigma^{med}(m|\theta)[\theta u(m)+\beta w(m)]\geq \sum_m \sigma^{med}(m|\theta')[\theta u(m)+\beta w(m)].$$

A Word of Caution

We conclude this section with a word of caution about the usage of the term "Revelation Principle" in the literature. Some authors reserve this term only for principal–agent models and Theorem 1. Since Theorem 1 does not hold if the principal cannot commit, those authors often say that "the Revelation Principle fails without commitment" (see, eg, Laffont and Tirole, 1988 or Bester and Strausz, 2001). Other authors use this term more broadly as in Section 3.2.1, when the mechanism designer is thought not as a principal per se but rather as a mechanical randomizing device. In such settings truthful direct revelation holds both when the agents and the principal can and cannot commit (see, eg, Myerson, 1982, 1986 and Mas-Colell et al., 1995), and one often hears that "the Revelation Principle always holds." While it may be confusing, there is no disagreement about the mathematical facts, and one just needs to be careful about which version of the Revelation Principle one refers to.

3.3 Martingale Methods in Continuous Time

We now show how dynamic contracting problems can be conveniently analyzed in continuous time frameworks. We only briefly touch on this literature here. Sannikov (2008, 2014) analyzed a continuous-time dynamic moral hazard problem where observable output follows a Brownian motion whose drift is given by the agent's unobservable effort. Williams (2009, 2011) uses the stochastic Pontryagin principle based on the work of Bismut (1973, 1978) to analyze a continuous-time version of the Thomas and Worrall (1990) endowment shock model, and Cvitanić and Zhang (2013) apply the same techniques to moral hazard and adverse selection problems. Zhang (2009) considers a dynamic contracting problem with a finite Markov chain for the types. Miao and Zhang (2015) extend the Lagrangian techniques introduced in Section 3.1 to a model of limited commitment (see Section 4) in continuous time.

Here we follow Sannikov (2008) who uses the dynamic programming principle in continuous time to analyze the moral hazard model described in the discrete time setting in Section 2.7. We start with a short section on the mathematical techniques that allow us to solve this problem. A fully rigorous exposition of these techniques is beyond the scope of this paper, but we present the main tools that allow us to describe Sannikov (2008)'s model in a self-contained way.

3.3.1 *Mathematical Background*

For the basics of Brownian motion and stochastic processes, see, eg, Revuz and Yor (1999), Øksendal (2003), or Karatzas and Shreve (2012). For an exposition of the theory of stochastic optimal control, see, eg, Yong and Zhou (1999). In this section, after briefly introducing the basics of stochastic processes, we simply state the three fundamental theorems that will be important in the analysis of the continuous-time dynamic contracting model below, namely Itô's lemma, the Martingale Representation Theorem, and Girsanov's theorem. We also describe heuristically the dynamic programming principle in continuous time.

A stochastic process X is a family of random variables $\{X_t\}_{t\geq 0}$ on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Define the filtration $\{\mathscr{F}_t\}_{t\geq 0}$ such that for all t, $\mathscr{F}_t = \sigma\left(\{X_s\}_{s\leq t}\right)$ is the σ-algebra generated by X from time 0 to time t. We say that the process X is Markovian if $\mathbb{P}(X_t \in A | \mathscr{F}_s) = \mathbb{P}(X_t \in A | X_s)$ for all $t > s$ and all Borel sets A. The process X is a martingale (resp., submartingale) if $\mathbb{E}[|X_t|] < \infty$ for all $t \geq 0$ and $\mathbb{E}[X_t | \mathscr{F}_s] = X_s$ (resp., $\mathbb{E}[X_t | \mathscr{F}_s] \geq X_s$) for all $t > s$. An important example of martingale is the Brownian motion. A stochastic process $\mathcal{Z} = \{\mathcal{Z}_t\}_{t\geq 0}$ on $(\Omega, \mathscr{F}, \mathbb{P})$ is a *Brownian motion* if it satisfies (see Section 37 in Billingsley, 2008):

(i) The process starts at 0: $\mathbb{P}(\mathcal{Z}_0 = 0) = 1$;

(ii) The increments are independent: if $0 \leq t_0 \leq \cdots \leq t_n$, then

$$\mathbb{P}(\mathcal{Z}_{t_k} - \mathcal{Z}_{t_{k-1}} \in A_k, \quad \forall k \leq n) = \prod_{k=1}^{n} \mathbb{P}(\mathcal{Z}_{t_k} - \mathcal{Z}_{t_{k-1}} \in A_k);$$

(iii) For $0 \leq s < t$ the increment $\mathcal{Z}_t - \mathcal{Z}_s$ is normally distributed with mean 0 and variance $t - s$:

$$\mathbb{P}(\mathcal{Z}_t - \mathcal{Z}_s \in A) = \frac{1}{\sqrt{2\pi(t-s)}} \int_A e^{-x^2/2(t-s)} dx;$$

(iv) The sample paths are continuous: for each $\omega \in \Omega$, the function $t \mapsto \mathcal{Z}_t(\omega)$ is continuous.

We now define the concept of quadratic variation of a martingale. Consider a martingale M that has continuous sample paths. Consider a partition $\pi_t = \{t_0, \ldots, t_n\}$ of the interval $[0, t]$ with $0 = t_0 < t_1 < \cdots < t_n = t$, and denote its mesh by $\|\pi_t\| \equiv \max_{1 \leq k \leq n}(t_k - t_{k-1})$. Denoting by $\mathbb{P}\lim$ the limit of a process in the sense of the convergence in probability, we can show that

$$\mathbb{P}\lim_{\|\pi_t\| \to 0} \sum_{k=1}^{n} (M_{t_k} - M_{t_{k-1}})^2 = \langle M \rangle_t,$$

where $\langle M \rangle$ is an adapted process with continuous and nondecreasing sample paths, called the *quadratic variation* of the martingale M. In particular, in the case where M is a Brownian motion, $\langle M \rangle$ is the deterministic process $\langle M \rangle_t = t$, and the convergence holds almost surely. Since $\langle M \rangle$ has nondecreasing sample paths ω, we can define the (path-by-path) Lebesgue–Stieltjes integral $\int_0^t X_s(\omega) d\langle M \rangle_s(\omega)$ for each ω of a stochastic process X on an interval $[0, T]$ with $T < \infty$ (in the case where M is a Brownian motion, $d\langle M \rangle_s = ds$ is simply the Lebesgue measure).

We refer to Revuz and Yor (1999) for the rigorous construction of the stochastic integral $\int_0^t X_s dM_s$ of a process X with respect to a martingale M that has continuous sample paths (eg, a Brownian motion). For such a martingale M, let $L^2(M)$ denote the (Hilbert) space of processes X such that for all $t \geq 0$ the map $(\omega, s) \mapsto X_s(\omega)$ defined on $\Omega \times [0, t]$ is measurable with respect to $\mathscr{F}_t \otimes \mathcal{B}([0, t])$, and $\mathbb{E}[\int_0^T X_s^2 d\langle M \rangle_s] < \infty$. The construction of the stochastic integral involves several steps. Suppose first that X is a "simple" process, in the sense that there exists a partition $0 = t_0 < t_1 < \cdots < t_n = T$ of $[0, T]$ such that $X_s = \xi_j$ for all $s \in (t_j, t_{j+1}]$, where ξ_j is a bounded $\mathscr{F}_{t_j}$-measurable random variable. That is, X can be written as

$$X_s(\omega) = \sum_{j=0}^{n-1} \xi_j(\omega) \mathbb{I}_{(t_j, t_{j+1}]}(s).$$

We can then define, for $t_k < t \leq t_{k+1}$,

$$I_t(X) = \int_0^t X_s dM_s \equiv \sum_{j=0}^{k-1} \xi_j \left(M_{t_j+1} - M_{t_j} \right) + \xi_k (M_t - M_{t_k}).$$

The integral $I(X)$ is then a square integrable continuous martingale with quadratic variation given by $\langle I(X) \rangle_t = \int_0^t X_s^2 d\langle M \rangle_s$. Next, any process $X \in L^2(M)$ can be approximated by a sequence of simple processes $\{X^n\}_{n \geq 0}$ in the sense that $\mathbb{E}[\int_0^T \left(X_s^n - X_s \right)^2 d\langle M \rangle_s] \to 0$. We can then show that the sequence of integrals $I(X^n)$ is a Cauchy sequence in the complete space $L^2(M)$. Its limit defines the stochastic integral. It satisfies $\mathbb{E}[I_t(X)] = 0$ and is a martingale.

We now state the three main theorems which we use in our analysis. The first, *Itô's lemma*, is an extension of the chain rule from standard calculus:

Theorem 6 (Itô's lemma) *Let f be a deterministic $\mathcal{C}^2$ function and M a squared integrable martingale. We have:*

$$f(M_t) = f(M_0) + \int_0^t f'(M_s) dM_s + \frac{1}{2} \int_0^t f''(M_s) d\langle M \rangle_s.$$

The second important result is the *Martingale Representation Theorem*. If M is a martingale, define the *exponential martingale*

$$\mathcal{E}(M)_t = \exp\left(M_t - \frac{1}{2} \langle M \rangle_t \right). \tag{105}$$

We can then show that $\mathcal{E}(M)_t$ is a supermartingale, and it is a martingale if in addition $\mathbb{E}\left[\exp\left(\frac{1}{2} \langle M \rangle_T \right) \right] < \infty$. In particular, if M_t is defined as a stochastic integral with respect to a Brownian motion $\mathcal{Z}$, ie, $M_t = \int_0^t \mu_s d\mathcal{Z}_s$ with $\int_0^t \mu_s^2 ds < \infty$ a.s., then

$$\mathcal{E}(M)_t = \exp\left(\int_0^t \mu_s d\mathcal{Z}_s - \frac{1}{2} \int_0^t \mu_s^2 ds \right) \tag{106}$$

is a martingale if $\mathbb{E}\left[\exp\left(\frac{1}{2} \int_0^T \mu_s^2 ds \right) \right] < \infty$.

Theorem 7 (Martingale Representation Theorem) *Let $\mathcal{Z}$ be a given Brownian motion. Every square integrable continuous martingale M adapted to the filtration $\mathscr{F}^{\mathcal{Z}}$ generated by $\mathcal{Z}$ admits a unique representation*

$$M_t = M_0 + \int_0^t \beta_s d\mathcal{Z}_s$$

for some process β *adapted to* $\mathscr{F}^{\mathcal{Z}}$ *that satisfies* $\mathbb{E}\left[\exp\left(\int_0^T \beta_s^2 ds\right)\right] < \infty$.

Finally, the third important result that we will use is *Girsanov's theorem*, which concerns the changes of measures.

Theorem 8 (Girsanov theorem) *Let* $\mathcal{Z}$ *be a Brownian motion and* μ *be an adapted process with* $\int_0^t \mu_s^2 ds < \infty$ *a.s. Let* $\mathcal{E}(M)_t$ *be defined by (106). If* $\mathbb{E}\left[\mathcal{E}(M)_T\right] = 1$ *(which implies that* $\mathcal{E}(M)$ *is a martingale) then, under*

$$\widetilde{\mathbb{P}}\,(d\omega) = \mathcal{E}(M)_T(\omega) \times \mathbb{P}(d\omega),$$

the process

$$\widetilde{\mathcal{Z}} = \mathcal{Z} - \int_0^t \mu_s ds$$

is a Brownian motion.

Finally we describe heuristically the dynamic programming principle in continuous time. We skip many of the technicalities and refer to Yong and Zhou (1999) for a rigorous exposition. Consider a filtered probability space $\left(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geq 0}, \mathbb{P}\right)$, on which a Brownian motion $\mathcal{Z}$ is defined, and let $T \in (0, \infty)$ and $A \subset \mathbb{R}$ be a given Borel set. The state of a system at time t is described by a stochastic process $X_t \in \mathbb{R}$ that evolves according to

$$X_{t'} = x + \int_t^{t'} b(s, X_s, u_s) ds + \int_t^{t'} \sigma(s, X_s, u_s) d\mathcal{Z}_s, \quad 0 \leq t \leq t' \leq T, \tag{107}$$

where $\boldsymbol{u} : [0, T] \times \Omega \to A$ is the *control process*, and $b, \sigma : [0, T] \times \mathbb{R} \times A \to \mathbb{R}$. The goal is to choose $\boldsymbol{u}$ to maximize the functional

$$J(\boldsymbol{u}) \equiv \mathbb{E}\left[\int_t^T f(s, X_s, u_s) ds + g(X_T)\right], \tag{108}$$

where $f : [0, T] \times \mathbb{R} \times A \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$.[ae] We assume that the functions b,σ,f,g, satisfy suitable conditions ensuring that there exists a unique solution X to (107) for any $t,x,\boldsymbol{u}$ and that the functional $J(\boldsymbol{u})$ in (108) is well defined (see Definition 6.15. and Conditions (S1)$'$ and (S2)$'$ p. 177 in Yong and Zhou, 1999). The control process $\boldsymbol{u}$ is *admissible* if: (i) u_t

[ae] Here the control problem ends at a fixed duration T. In our analysis of the moral hazard problem we will deal instead with random horizons T optimally chosen by the principal (retirement), that is, where T is the stopping time $T \equiv \inf\{t \geq 0 : x_t \notin \mathcal{O}\}$ for some open set $\mathcal{O} \subset \mathbb{R}$. The dynamic programming principle can be extended to this case, see, eg, Section 2.7. in Yong and Zhou (1999) and Chapter 4 in Øksendal and Sulem (2007).

is $\{\mathscr{F}_t\}_{t\geq 0}$-adapted; (ii) X is the unique solution of Eq. (107); and (iii) the functions $f(\cdot, X., u.)$ and $g(X_T)$ are in $L^1_{\mathscr{F}}([0,T],\mathbb{R})$ and $L^1_{\mathscr{F}_T}(\boldsymbol{\Omega},\mathbb{R})$, respectively. The value function of the stochastic control problem that we consider is

$$V(t,x) = \sup_{\boldsymbol{u}} J(\boldsymbol{u}),$$

where the supremum is over all admissible controls $\boldsymbol{u}$.[af]

Theorem 9 (Dynamic Programming Principle) *For any stopping time τ with values in $[0,T]$, the value function $V(t,x)$ is equal to*

$$V(t,x) = \sup_{\boldsymbol{u}} \mathbb{E}\left[\int_t^\tau f(s,X_s,u_s)ds + V(\tau,X_\tau)|X_t = x\right].$$

Moreover, for all admissible controls $\boldsymbol{u}$,

$$M_{t'} \equiv \int_t^{t'} f(s,X_s,u_s)ds + V(t',X_{t'})$$

is a supermartingale (ie, $-M_{t'}$ is a submartingale), and it is a martingale if and only if $\boldsymbol{u}$ is optimal. Suppose that the value function $V \in \mathcal{C}^{1,2}([0,T]\times\mathbb{R})$. Then V is a solution to the following second-order Hamilton–Jacobi–Bellman partial differential equation:

$$\begin{cases} -\dfrac{\partial V}{\partial t} + \sup\limits_{u\in A}\left[f(t,x,u) + b(t,x,u)\dfrac{\partial V}{\partial x} + \dfrac{1}{2}\sigma^2(t,x,u)\dfrac{\partial^2 V}{\partial x^2}\right] = 0, & \forall (t,x)\in[0,T]\times\mathbb{R}, \\ V(T,x) = g(x), & \forall x\in\mathbb{R}. \end{cases}$$

Note that the last statement assumes smoothness conditions about the value function V, which is endogenous.[ag]

3.3.2 Moral Hazard in Continuous Time

We now analyze the moral hazard problem in a continuous time framework (see Section 2.7 for the discrete time version of the model), following Sannikov (2008)'s exposition. Our aim is to derive and explain the main results with the minimum of

[af] Rigorously, it is often natural and necessary to consider a *weak formulation* of the problem, in which the filtered probability space $(\Omega,\mathcal{F},\{\mathcal{F}_t\}_{t\geq 0},\mathbb{P})$ and the Brownian motion $\mathcal{Z}$ are not fixed, but parts of the control (see Sections 2.4.4. and 4.3.1. in Yong and Zhou, 1999). This is because the objective of the stochastic control problem is to minimize the expectation of a random variable that depends only on the distribution of the processes involved. We ignore this distinction in the sequel.

[ag] There exist other notions of solutions to stochastic differential equations, called viscosity solutions, which avoid making such assumptions, see, eg, Section 4.5. in Yong and Zhou (1999).

technicalities. Therefore we omit many technical details and refer to Sannikov's work for the fully rigorous proofs.

We analyze a model where the agent's current effort affects only current output. The agent derives utility $U(c_t) - h(\theta_t)$ from consumption $c_t \geq 0$ and effort $\theta_t \in \left[0, \overline{\theta}\right]$ at time t, where U is twice continuously differentiable, increasing, and concave with $U(0) = 0$ and $\lim_{c\to\infty} U'(c) = 0$, and h is differentiable, increasing, and convex with $h(0) = 0$ and $h'(0) > 0$.

Fix a reference probability space $(\boldsymbol{\Omega}, \mathscr{F}, \mathbb{P})$ with a standard Brownian motion $\mathcal{Z}$ under $\mathbb{P}$. If the agent works according to the effort process $\boldsymbol{\theta} = \{\theta_t\}_{t\in[0,\,\infty)}$ with $0 \leq \theta_t \leq \overline{\theta}$ for all t, he generates an output y_t given by $y_t = \int_0^t \theta_s ds + \sigma \mathcal{Z}_t$, ie,

$$dy_t = \theta_t dt + \sigma d\mathcal{Z}_t,$$

where $\sigma > 0$ is a constant. The principal observes y_t, but not θ_t or $\mathcal{Z}_t$, and compensates the agent with a consumption process $\mathbf{c} = \{c_t\}_{t\geq 0}$ with $c_t \geq 0$ for all t. Denoting by $\mathscr{F}_t^y$ the filtration generated by y_t, we impose that the process c_t is $\mathscr{F}_t^y$-adapted, ie, the agent's compensation c_t is conditional on past output $\{y_s\}_{s\leq t}$.

Rather than solving for the agent's effort choice $\boldsymbol{\theta}$ as a function of the fixed underlying Brownian motion $\mathcal{Z}$, we can instead view the agent as choosing a probability measure $\mathbb{P}^{\boldsymbol{\theta}}$ on the output space.[ah] That is, for each effort process $\boldsymbol{\theta}$ we can define a process $\mathcal{Z}_t^{\boldsymbol{\theta}} = \sigma^{-1}\left(y_t - \displaystyle\int_0^t \theta_s ds\right)$. By Girsanov's theorem, $\mathcal{Z}_t^{\boldsymbol{\theta}}$ is a Brownian motion under the measure $\mathbb{P}^{\boldsymbol{\theta}}$, where

$$\mathbb{P}^{\boldsymbol{\theta}}(d\omega) = \mathcal{E}(\mathcal{Z})_t \mathbb{P}(d\omega) = e^{\int_0^t \theta_s d\mathcal{Z}_s - \frac{1}{2}\int_0^t \theta_s^2 ds} \mathbb{P}(d\omega).$$

A change of measure from $\mathbb{P}^{\boldsymbol{\theta}}$ to $\mathbb{P}^{\hat{\boldsymbol{\theta}}}$ on the space of output paths corresponds to a change in the drift of the output process from $\boldsymbol{\theta}$ to $\hat{\boldsymbol{\theta}}$.

Planner's Problem

If he receives consumption $\mathbf{c} = \{c_t\}_{t\geq 0}$ and provides effort $\boldsymbol{\theta} = \{\theta_t\}_{t\geq 0}$, the agent gets the expected utility

$$U(\mathbf{c}, \boldsymbol{\theta}) = \mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^\infty e^{-rt}(U(c_t) - h(\theta_t))dt\right], \tag{109}$$

where $\mathbb{E}^{\boldsymbol{\theta}}$ denotes the expectation under the probability measure $\mathbb{P}^{\boldsymbol{\theta}}$ induced by the strategy $\boldsymbol{\theta}$, as defined above. The superscript $\boldsymbol{\theta}$ over the expectation $\mathbb{E}^{\boldsymbol{\theta}}$ highlights that the agent's strategy affects the probability distribution over the paths of output, and thus over

[ah] Similarly, in the standard static moral hazard problem, we can view the agent as choosing the probability distribution $\mathbb{P}(y|\theta)$ over output values y generated by his effort θ.

the compensation realizations. Thus, the utility depends on the agent's effort directly, as it enters the cost of effort $h(\theta_t)$, and indirectly through its effect on the probability distribution over the paths of y_t.

The principal gets expected profit

$$\mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^\infty e^{-rt}(dy_t - c_t dt)\right] = \mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^\infty e^{-rt}(\theta_t - c_t)dt\right]. \tag{110}$$

A contract $(\mathbf{c},\boldsymbol{\theta})$ is *incentive compatible* if the agent finds it optimal to exert the contractual effort θ_t at every t, ie, if $\{\theta_t\}_{t\geq 0}$ maximizes his expected utility $U(\mathbf{c},\boldsymbol{\theta})$ given $\{c_t\}_{t\geq 0}$:

$$\mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^\infty e^{-rt}(U(c_t) - h(\theta_t))dt\right] \geq \mathbb{E}^{\hat{\boldsymbol{\theta}}}\left[\int_0^\infty e^{-rt}\left(U(c_t) - h\left(\hat{\theta}_t\right)\right)dt\right], \forall \hat{\boldsymbol{\theta}}. \tag{111}$$

The contract must deliver initial promised utility $\hat{v}_0$, ie,

$$\mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^\infty e^{-rt}(U(c_t) - h(\theta_t))dt\right] \geq \hat{v}_0. \tag{112}$$

The principal's problem consists of choosing the contract $(\mathbf{c},\boldsymbol{\theta})$ that maximizes his expected profit (110) among all the contracts that satisfy the incentive-compatibility (111) and promise-keeping (112) constraints, that is,

$$\max_{\mathbf{c},\boldsymbol{\theta}} \ \mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^\infty e^{-rt}(\theta_t - c_t)dt\right]$$
$$\text{subject to} \quad (111),\ (112).$$

The principal can commit to the contract he offers.

Reducing the Planner's Problem to an Optimal Stochastic Control Problem

The planner's problem can be solved by reducing it to an optimal stochastic control problem. As in the discrete time framework, we use the agent's continuation utility v_t (defined formally below) as state variable. The key simplification of the planner's problem comes again from the (continuous-time equivalent of the) one-shot deviation principle, which substantially reduces the set of incentive constraints: the agent's incentive constraints hold for all alternative strategies $\hat{\boldsymbol{\theta}} = \left\{\hat{\theta}_t\right\}_{t\geq 0}$ if they hold just for strategies that differ from $\boldsymbol{\theta} = \{\theta_t\}_{t\geq 0}$ for an instant. The Martingale Representation Theorem then allows us to express the instantaneous incentive constraints in terms of v_t.

Fix an arbitrary consumption process $\mathbf{c} = \{c_t\}_{t\geq 0}$ and an effort strategy $\boldsymbol{\theta} = \{\theta_t\}_{t\geq 0}$ (not necessarily optimal for the agent given $\mathbf{c}$). The agent's *continuation value* $v_t(\mathbf{c},\boldsymbol{\theta})$, defined as his expected future payoff from $(\mathbf{c},\boldsymbol{\theta})$ after time t (ie, after a given history of output $\{y_s\}_{s\leq t}$), is given by

$$v_t(\mathbf{c},\boldsymbol{\theta}) = \mathbb{E}^{\boldsymbol{\theta}}\left[\int_t^\infty e^{-r(s-t)}(U(c_s)-h(\theta_s))ds|\mathscr{F}_t\right]. \tag{113}$$

Throughout this section, for a given time t and contract $(\mathbf{c},\boldsymbol{\theta})$, we also define the agent's *total* expected payoff from the contract $(\mathbf{c},\boldsymbol{\theta})$ given the information at time t as:[ai]

$$V_t^{\mathbf{c},\boldsymbol{\theta}} = \mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^\infty e^{-rs}(U(c_s)-h(\theta_s))ds|\mathscr{F}_t\right] = \int_0^t e^{-rs}(U(c_s)-h(\theta_s))ds + e^{-rt}v_t(\mathbf{c},\boldsymbol{\theta}). \tag{114}$$

We first derive the law of motion of $v_t(\mathbf{c},\boldsymbol{\theta})$ by applying the Martingale Representation Theorem.

Proposition 12 *Fix a contract $(\mathbf{c},\boldsymbol{\theta})$ with finite expected payoff to the agent. An adapted process v_t is the continuation value process (as defined in (113)) associated with the contract $(\mathbf{c},\boldsymbol{\theta})$ if and only if there exists an $\mathscr{F}_t$-adapted process $\boldsymbol{\beta} = \{\beta_t\}_{t\geq 0}$ with $\mathbb{E}\left[\int_0^t \beta_s^2 ds\right] < \infty$ for all t such that, for all $t \geq 0$,*

$$dv_t = (rv_t - U(c_t) + h(\theta_t))dt + \beta_t(dy_t - \theta_t dt) \tag{115}$$

and the transversality condition $\lim_{t\to\infty}\mathbb{E}^{\boldsymbol{\theta}}[e^{-rt}v_{t_0+t}|\mathscr{F}_{t_0}] = 0$ holds almost everywhere.

Proof Fix a contract $(\mathbf{c},\boldsymbol{\theta})$. The process $V_t^{\mathbf{c},\boldsymbol{\theta}}$ defined in (114) is a martingale under the probability measure $\mathbb{P}^{\boldsymbol{\theta}}$. Hence by the Martingale Representation Theorem there exists an adapted process β_t such that

$$V_t^{\mathbf{c},\boldsymbol{\theta}} = V_0^{\mathbf{c},\boldsymbol{\theta}} + \int_0^t e^{-rs}\beta_s\sigma d\mathcal{Z}_s^{\boldsymbol{\theta}}, \quad 0 \leq t < \infty,$$

where $\mathcal{Z}_t^{\boldsymbol{\theta}} = \sigma^{-1}\left(y_t - \int_0^t \theta_s ds\right)$ is a Brownian motion under $\mathbb{P}^{\boldsymbol{\theta}}$. Differentiating both expressions for $V_t^{\mathbf{c},\boldsymbol{\theta}}$ with respect to t and equating them implies that $v_t(\mathbf{c},\boldsymbol{\theta})$ satisfies (115). The transversality condition (for simplicity with $t_0 = 0$) follows from

$$\lim_{t\to\infty}\mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^t e^{-rs}(U(c_s)-h(\theta_s))ds\right] = \mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^\infty e^{-rs}(U(c_s)-h(\theta_s))ds\right],$$

by the Dominated Convergence Theorem using that θ_s, and thus $\int_0^t e^{-rs}(U(c_s)-h(\theta_s))ds$, is bounded. A similar argument shows that $\lim_{t\to\infty}\mathbb{E}^{\boldsymbol{\theta}}[e^{-rt}v_{t_0+t}|\mathscr{F}_{t_0}] = 0$ for all times $t_0 \geq 0$.

[ai] See Theorem 9 above.

Conversely, suppose that v_t is a process that satisfies (115) (for some starting value v_0 and some volatility process β_t) and the transversality condition. Define V_t as

$$V_t = \int_0^t e^{-rs}(U(c_s) - h(\theta_s))ds + e^{-rt}v_t.$$

Differentiating V_t implies that it is a martingale when the agent is following the effort strategy $\boldsymbol{\theta}$, ie, under the probability measure $\mathbb{P}^{\boldsymbol{\theta}}$. Therefore

$$v_0 = \quad V_0 = \mathbb{E}^{\boldsymbol{\theta}}[V_t|\mathscr{F}_0] = \mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^t e^{-rs}(U(c_s) - h(\theta_s))ds|\mathscr{F}_0\right] + \mathbb{E}^{\boldsymbol{\theta}}[e^{-rt}v_t|\mathscr{F}_0].$$

Since the transversality condition is satisfied (for $t_0 = 0$), taking limits as $t \to \infty$ in the previous equation implies that $v_0 = v_0(\mathbf{c}, \boldsymbol{\theta})$. A similar argument shows that v_t is the continuation value process $v_t(\mathbf{c}, \boldsymbol{\theta})$ defined by (113) at any time $t \geq 0$. □

The law of motion (115) of the continuation utility has the following interpretation. Since $dy_t - \theta_t dt = \sigma d\mathcal{Z}_t^{\boldsymbol{\theta}}$ is a Brownian motion when the agent takes the recommended effort level $\boldsymbol{\theta}$, $[rv_t(\mathbf{c}, \boldsymbol{\theta}) - (U(c_t) - h(\theta_t))]$ is the drift of the agent's continuation value. The value that the principal owes to the agent (future expected payoff), $v_t(\mathbf{c}, \boldsymbol{\theta})$, grows at the rate of interest r, and falls due to the flow of repayments $(U(c_t) - h(\theta_t))$. The transversality condition has to hold if the debt is eventually repaid. Since the agent's compensation and recommended effort are determined by output y_t, his continuation payoff $v_t(\mathbf{c}, \boldsymbol{\theta})$ is also determined by output, and the process β_t then expresses the sensitivity of the agent's continuation value to output at a given time, which will be the key to affect the agent's incentives.

The previous lemma is useful because it allows us to simplify the set of incentive constraints with a version of the one-shot deviation principle (Proposition 13), which shows that the agent's incentive constraints hold for all alternative strategies $\hat{\boldsymbol{\theta}}$ if they hold for all strategies which differ from $\boldsymbol{\theta}$ for an infinitesimally small amount of time. Heuristically, suppose that the agent has conformed to the contract (c_s, θ_s) for $s \leq t$ and cheats by performing effort $\hat{\theta}$ in the interval $[t, t + dt]$ and reverting to $\{\theta_s\}$ for $s \geq t + dt$. His immediate consumption c_t is unaffected, his cost on $[t, t + dt]$ is $h(\hat{\theta})dt$, and his expected benefit on $[0, \infty)$, ie, the expected impact of effort on his continuation value, is $\mathbb{E}^{\hat{\boldsymbol{\theta}}}[\beta_t dy_t] = \beta_t \hat{\theta} dt$. Hence for the contract to be incentive compatible we must have

$$\beta_t \theta_t - h(\theta_t) = \max_{\hat{\theta} \geq 0} \left\{-h(\hat{\theta}) + \beta_t \hat{\theta}\right\},$$

almost everywhere.[aj] This argument can be made rigorous, and in addition the condition is not only necessary but also sufficient: if this one-shot condition holds at each instant t, then any dynamic deviation strategy $\hat{\boldsymbol{\theta}} = \left\{\hat{\theta}_s\right\}_{s \geq 0}$ is suboptimal.

[aj] Note the fixed point nature of the argument: θ_t generates $v_t(\mathbf{c}, \boldsymbol{\theta})$ which yields β_t; in turn, the incentives have to be satisfied given this process β_t.

Proposition 13 *Let $(\mathbf{c},\boldsymbol{\theta})$ be a contract with agent's continuation value $v_t(\mathbf{c},\boldsymbol{\theta})$ and let β_t be the process from Proposition 12 that represents $v_t(\mathbf{c},\boldsymbol{\theta})$. Then $(\mathbf{c},\boldsymbol{\theta})$ is incentive compatible if and only if $\forall\hat{\theta}\in[0,\overline{\theta}]$, $\forall t\geq 0$,*

$$\theta_t \in \arg\max_{\hat{\theta}\geq 0}\left\{\beta_t\hat{\theta}-h\left(\hat{\theta}\right)\right\}, \text{a.e.} \tag{116}$$

Proof Suppose that (116) is satisfied. Suppose that an agent follows the alternative effort process $\hat{\boldsymbol{\theta}}=\left\{\hat{\theta}_s\right\}_{s\geq 0}$ until time t and reverts back to $\boldsymbol{\theta}$ thereafter; denote by $\hat{\boldsymbol{\theta}}^t$ this strategy. The time-t expectation of his total payoff is given by $V_t^{\mathbf{c},\hat{\boldsymbol{\theta}}^t}$ defined in (114),

$$V_t^{\mathbf{c},\hat{\boldsymbol{\theta}}^t}=\int_0^t e^{-rs}\left(U(c_s)-h\left(\hat{\theta}_s\right)\right)ds+e^{-rt}v_t(\mathbf{c},\boldsymbol{\theta}).$$

Differentiating $V_t^{\mathbf{c},\hat{\boldsymbol{\theta}}^t}$ and using Eq. (115) to compute $d[e^{-rt}v_t(\mathbf{c},\boldsymbol{\theta})]$, we find[ak]

$$dV_t^{\mathbf{c},\hat{\boldsymbol{\theta}}^t}=e^{-rt}\left\{\left(\beta_t\hat{\theta}_t-h\left(\hat{\theta}_t\right)\right)-\left(\beta_t\theta_t-h(\theta_t)\right)\right\}dt+e^{-rt}\beta_t\left(dy_t-\hat{\theta}_t dt\right).$$

Thus, since $\left(dy_t-\hat{\theta}_t dt\right)$ is a Brownian motion under $\mathbb{P}^{\hat{\boldsymbol{\theta}}}$, if (116) holds the drift of $V_t^{\mathbf{c},\hat{\boldsymbol{\theta}}^t}$ under the probability measure $\mathbb{P}^{\hat{\boldsymbol{\theta}}}$ is nonpositive and thus $V_t^{\mathbf{c},\hat{\boldsymbol{\theta}}^t}$ is a $\mathbb{P}^{\hat{\boldsymbol{\theta}}}$-supermartingale. Hence we have

$$\mathbb{E}^{\hat{\boldsymbol{\theta}}}\left[\int_0^t e^{-rs}\left(U(c_s)-h\left(\hat{\theta}_s\right)\right)ds\right]+\mathbb{E}^{\hat{\boldsymbol{\theta}}}[e^{-rt}v_t(\mathbf{c},\boldsymbol{\theta})]=\mathbb{E}^{\hat{\boldsymbol{\theta}}}\left[V_t^{\mathbf{c},\hat{\boldsymbol{\theta}}^t}|\mathscr{F}_0\right]\leq V_0^{\mathbf{c},\hat{\boldsymbol{\theta}}^0}=v_0(\mathbf{c},\boldsymbol{\theta}).$$

Taking the limit as $t\to\infty$ using the fact that $\mathbb{E}^{\hat{\boldsymbol{\theta}}}[e^{-rt}v_t(\mathbf{c},\boldsymbol{\theta})]\geq -e^{-rt}h\left(\overline{\theta}\right)$, we obtain $v_0\left(\mathbf{c},\hat{\boldsymbol{\theta}}\right)\leq v_0(\mathbf{c},\boldsymbol{\theta})$.

Conversely, if (116) does not hold on a set of times and sample paths with positive measure, then pick a deviation $\hat{\boldsymbol{\theta}}$ defined as $\hat{\theta}_t=\arg\max_{\hat{\theta}}\left(-h\left(\hat{\theta}\right)+\beta_t\hat{\theta}\right)$ everywhere. The drift of $V_t^{\mathbf{c},\hat{\boldsymbol{\theta}}^t}$ under $\mathbb{P}^{\hat{\boldsymbol{\theta}}}$ is nonnegative and positive on a set of positive measure, so that for t large enough the time-0 expected payoff from following $\hat{\boldsymbol{\theta}}$ until time t and switching to $\boldsymbol{\theta}$ thereafter is $v_0\left(\mathbf{c},\hat{\boldsymbol{\theta}}^t\right)=\mathbb{E}^{\hat{\boldsymbol{\theta}}}\left[V_t^{\mathbf{c},\hat{\boldsymbol{\theta}}^t}|\mathscr{F}_0\right]>V_0^{\mathbf{c},\hat{\boldsymbol{\theta}}^0}=v_0(\mathbf{c},\boldsymbol{\theta})$. Thus the strategy $\boldsymbol{\theta}$ is suboptimal. □

For a given sensitivity β, denote by $\theta(\beta)$ the effort that maximizes $(-h(\theta)+\beta\theta)$, namely $\theta(\beta)=h'^{-1}(\beta)$ if $\beta>0$ and $\theta(\beta)=0$ if $\beta=0$. Conversely, for a given effort level θ define the sensitivity $\beta(\theta)$ that ensures incentive compatibility as $\beta(\theta)=h'(\theta)$ if $\theta>0$, and $\beta(\theta)=0$ if $\theta=0$.

[ak] This equation evaluates the incremental change in the agent's utility from pursuing the alternative effort strategy $\hat{\boldsymbol{\theta}}$ for an additional unit of time during $[t,t+dt]$, and shows that in expectation such an incremental deviation hurts the agent. The next equation then uses a supermartingale argument to obtain inductively that the whole deviation strategy $\left\{\hat{\theta}_t\right\}_{t\geq 0}$ is worse than $\{\theta_t\}_{t\geq 0}$.

We are now ready to reformulate the planner's problem as a stochastic control problem, using the continuation value v_t as the single state variable.

Solution to the Optimal Stochastic Control Problem

The planner maximizes his expected profit (110) over incentive-compatible contracts $(\mathbf{c},\boldsymbol{\theta})$ subject to the law of motion of v_t, the transversality condition, and delivering initial promised utility $\hat{v}_0$. We consider a relaxed problem without the transversality condition (to be checked ex post). Before we analyze this problem, note that as in the discrete time setting, the principal has the option of "retiring" the agent at a given time τ by allocating a constant consumption $c_t = c$ and recommending zero effort $\theta_t = 0$ for all $t \geq \tau$. The continuation value at retirement time τ is then $v_\tau = r^{-1}U(c)$, so that $c = U^{-1}(rv_\tau)$. The retirement time τ must be specified in the contract, so it is a stopping time with respect to the filtration $\mathscr{F}_t$ generated by the output process y. We can thus write the principal's value of the optimal contract as

$$K(\hat{v}_0) = \max_{\mathbf{c},\boldsymbol{\theta},\tau} \mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^\tau e^{-rt}\left((\theta_t - c_t)dt + \sigma d\mathcal{Z}_t^{\boldsymbol{\theta}}\right) - \frac{e^{-r\tau}}{r}U^{-1}(rv_\tau)\right] \tag{117}$$

$$\text{subject to } \; dv_t = (rv_t - U(c_t) + h(\theta_t))dt + \beta(\theta_t)\sigma d\mathcal{Z}_t^{\boldsymbol{\theta}} \tag{118}$$

$$v_0 = \hat{v}_0. \tag{119}$$

Note in particular that the incentive constraints (111) are automatically satisfied if the constraint (118) holds. The function $K(v)$ can be found using standard optimal control and optimal stopping techniques, where the control variables are θ_t, c_t, τ and the state variable is v_t. The principal's problem can be solved in two steps: first, guess an optimal contract using the appropriate Bellman equation; second, verify ex post that this contract is indeed optimal.

We start by conjecturing the optimal contract. The function K is continuous on $[0, \infty)$ with $K(v) \geq -r^{-1}U^{-1}(rv)$ for all v. It satisfies the following Hamilton–Jacobi–Bellman (HJB) equation:[al]

[al] In fact, this is a Hamilton–Jacobi–Bellman Variational Inequality (HJBVI), where the HJB comes from the optimal stochastic control problem and the VI comes from the optimal stopping problem. See Chapter 4 in Øksendal and Sulem (2005).

$$rK(v) = \max\left\{ -U^{-1}(rv); \max_{\substack{0\le\theta\le\bar{\theta}\\ c\ge 0}} (\theta - c) + (rv - U(c) + h(\theta))K'(v) + \frac{1}{2}\sigma^2(\beta(\theta))^2 K''(v) \right\} \tag{120}$$

with the three boundary conditions

$$K(0) = 0, \quad K(\bar{v}) = -r^{-1}U^{-1}(r\bar{v}), \quad K'(\bar{v}) = -U^{-1\prime}(r\bar{v}), \tag{121}$$

for some $\bar{v} \ge 0$. Intuitively, (120) means that the principal maximizes the expected current flow of profit $(\theta - c)$ plus the expected change of future profit due to the drift and volatility of the agent's continuation value, until the stopping time τ at which the principal either retires the agent (if $v_\tau = \bar{v}$) or fires him (if $v_\tau = 0$). The second and third boundary conditions in (121) mean that the optimal retirement time occurs at the continuation value $\bar{v}$ where the value-matching condition (which equates the value of retiring the agent with that of continuing with positive effort) and the smooth-pasting condition (which equates the marginal values of retiring and continuing) are satisfied. We can show that there exists a unique function K that satisfies the HJB equation (120) with the three boundary conditions (121). The stopping time $\tau = \inf\{t \ge 0 : K(v) \le -r^{-1}U^{-1}(rv)\}$ satisfies $\tau < \infty$ a.s.,[am] and the function K is concave.

Define, for an arbitrary control policy $(\boldsymbol{c}, \boldsymbol{\theta})$, the process

$$G_t^{\boldsymbol{c},\boldsymbol{\theta}} = \int_0^t e^{-rs}(\theta_s - c_s)ds + e^{-rt}K(v_t). \tag{122}$$

The following proposition conjectures the optimal contract from the solution to (120), (121) and then verifies that it is indeed optimal using martingale techniques:

Proposition 14 *Denote by $\theta(v), c(v)$ the maximizers in the right hand side of the HJB equation.*[an] *Consider the unique solution $K(v) \ge -r^{-1}U^{-1}(rv)$ to the HJB equation (120) that satisfies the conditions (121) for some $\bar{v} \ge 0$. For any $\hat{v}_0 \in [0, \bar{v}]$, define the process v_t by $v_0 = \hat{v}_0$ and*

[am] If $h'(0) = 0$, the retirement point $\bar{v}$ may not be finite, so that $K(v)$, $c(v)$, $\theta(v)$ asymptote to $-r^{-1}U^{-1}(rv)$, ∞, 0 as $v \to \infty$.

[an] The optimal effort maximizes the difference between the expected flow of output θ, and the costs of compensating the agent for his effort, $-h(\theta)K'(v)$, and of exposing him to income uncertainty to provide incentives, $-\frac{\sigma^2}{2}\beta(\theta)^2K''(v)$. The optimal consumption is 0 for v small enough (ie, for $K'(v) \ge -1/u'(0)$), and it is increasing in v according to $K'(v) = -1/U'(c)$ otherwise, where $1/U'(c)$ and $-K'(v)$ are the marginal costs of giving the agent value through current consumption and through his continuation payoff, respectively.

$$dv_t = r(v_t - u(c(v_t)) + h(\theta(v_t)))dt + r\beta(\theta(v_t))(dy_t - \theta(v_t)dt) \tag{123}$$

until the stopping time τ when v_τ hits 0 or $\bar{v}$. Define the contract $(\mathbf{c},\boldsymbol{\theta})$ with payments $c_t = c(v_t)$ and recommended effort $\theta_t = \theta(v_t)$ for $t < \tau$, and $c_t = U^{-1}(rv_\tau)$ and $\theta_t = 0$ for $t \geq \tau$. Then $(\mathbf{c},\boldsymbol{\theta})$ is incentive compatible and it has a value $\hat{v}_0 = v_0(\mathbf{c},\boldsymbol{\theta})$ to the agent and profit $K(\hat{v}_0)$ to the principal. Moreover, consider a concave solution K of the HJB equation (120). Any incentive-compatible contract $(\mathbf{c},\boldsymbol{\theta})$ yields to the principal a profit less than or equal to $K(v_0(\mathbf{c},\boldsymbol{\theta}))$.

Proof Let v_t be given by the stochastic differential equation (123) for $t \leq \tau$ and $v_t = v_\tau$ for $t > \tau$ (note in particular that $v_t \in [0,\bar{v}]$ is bounded). We show that $v_t = v_t(\mathbf{c},\boldsymbol{\theta})$ for all $t \geq 0$, where $v_t(\mathbf{c},\boldsymbol{\theta})$ is the agent's true continuation value in the contract $(\mathbf{c},\boldsymbol{\theta})$ constructed above. This will imply in particular that the agent gets value $v_0(\mathbf{c},\boldsymbol{\theta}) = \hat{v}_0$ from the contract. From the representation of $v_t(\mathbf{c},\boldsymbol{\theta})$ in Proposition 12, we have

$$d(v_t(\mathbf{c},\boldsymbol{\theta}) - v_t) = r(v_t(\mathbf{c},\boldsymbol{\theta}) - v_t)dt + (\beta_t - \beta(\theta(v_t)))\sigma d\mathcal{Z}_t^{\boldsymbol{\theta}},$$

hence for all $s \geq 0$, $\mathbb{E}^{\boldsymbol{\theta}}[v_{t+s}(\mathbf{c},\boldsymbol{\theta}) - v_{t+s}] = e^{rs}(v_t(\mathbf{c},\boldsymbol{\theta}) - v_t)$. But $\mathbb{E}^{\boldsymbol{\theta}}[v_{t+s}(\mathbf{c},\boldsymbol{\theta}) - v_{t+s}]$ is bounded, hence $v_t = v_t(\mathbf{c},\boldsymbol{\theta})$. Moreover, the contract $(\mathbf{c},\boldsymbol{\theta})$ is incentive compatible by construction since the process from Proposition 12 that represents $v_t(\mathbf{c},\boldsymbol{\theta})$ is $\beta_t = \beta(\theta_t)$.

Next we show that the principal gets expected profit $K(\hat{v}_0)$ from the contract. Differentiating expression (122) and applying Itô's lemma to $K(v_t)$ yields that the drift of $G_t^{\mathbf{c},\boldsymbol{\theta}}$ under $\mathbb{P}^{\boldsymbol{\theta}}$ is

$$e^{-rt}\left\{(\theta_t - c_t - rK(v_t)) + (rv_t - U(c_t) + h(\theta_t))K'(v_t) + \frac{1}{2}\sigma^2(\beta(\theta_t))^2 K''(v_t)\right\}.$$

Thus, when $c_t = c(v_t)$ and $\theta_t = \theta(v_t)$, the drift of $G_t^{\mathbf{c},\boldsymbol{\theta}}$ under $\mathbb{P}^{\boldsymbol{\theta}}$ is equal to zero before time τ, so that $G_t^{\mathbf{c},\boldsymbol{\theta}}$ is a martingale. By the Optional Stopping Theorem, we thus obtain that the principal's profit from the contract is

$$\mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^\tau e^{-rs}(\theta_s - c_s)ds\right] + \mathbb{E}^{\boldsymbol{\theta}}[e^{-r\tau}K(v_\tau)] = \mathbb{E}\left[G_\tau^{\mathbf{c},\boldsymbol{\theta}}\right] = G_0^{\mathbf{c},\boldsymbol{\theta}} = K(v_0(\mathbf{c},\boldsymbol{\theta})).$$

Finally, consider an alternative incentive-compatible contract $(\mathbf{c},\boldsymbol{\theta})$. Then (120) implies that the drift of $G_t^{\mathbf{c},\boldsymbol{\theta}}$ under $\mathbb{P}^{\boldsymbol{\theta}}$ is smaller than zero, so that $G_t^{\mathbf{c},\boldsymbol{\theta}}$ is a bounded supermartingale. By the Optional Stopping Theorem, we obtain that the principal's expected profit at time 0 is less than or equal to $G_0^{\mathbf{c},\boldsymbol{\theta}} = K(v_0(\mathbf{c},\boldsymbol{\theta}))$. We refer to Sannikov (2008) for the technical details omitted in this sketch of proof. □

For any $v_0 > \bar{v}$, the function $K(v)$ is negative and is an upper bound on the principal's value function; thus, there is no profitable contract with positive profit to the principal in that range. In the range $(0,\bar{v})$, substituting for the optimal consumption $c(v)$ and effort

$\theta(v)$ into the HJB equation, we obtain a nonlinear second-order differential equation for $K(v)$ which can be solved numerically. Finally, note that the envelope theorem applied to the HJB equation before retirement implies

$$(rv - U(c) + h(\theta))K''(v) + \frac{1}{2}\sigma^2(\beta(\theta))^2 K'''(v) = 0.$$

By Itô's lemma, the left hand side is the drift of $K'(v_t) = -1/U'(c_t)$ on the interval $[\underline{v}, \bar{v}]$. Thus the inverse of the agent's marginal utility is a martingale when the agent's consumption is positive, a result that parallels the Inverse Euler Equation (77) found in the discrete-time model.

Sannikov (2014) extends the analysis of the moral hazard model in continuous time to the case where current actions affect not only current output but also future output. The solution to this problem is more involved than that of Sannikov (2008), but the steps and the martingale techniques (using the Martingale Representation Theorem to simplify the set of incentive constraints and reduce the problem to a stochastic control problem) are similar.

We conclude this section with a brief discussion of the benefits of using a continuous-time rather than discrete-time framework to analyze dynamic contracting problems. First, the Hamilton–Jacobi–Bellman equation is more tractable analytically than the discrete-time Bellman equation (23). In particular DeMarzo and Sannikov (2006) show how differentiating the HJB equation and its boundary conditions that characterize the optimal contract allows us to derive comparative statics results analytically. Here we illustrate their method on a simple example. Suppose, for instance, that we are interested in the effect of the volatility σ^2 on the principal's profit $K(v)$. Differentiating (120) yields, for $v \in (0, \bar{v})$,

$$r\frac{\partial K(v)}{\partial \sigma^2} = \frac{1}{2}(\beta(\theta))^2 K''(v) + (rv - U(c) + h(\theta))\left(\frac{\partial K(v)}{\partial \sigma^2}\right)' + \frac{1}{2}\sigma^2(\beta(\theta))^2\left(\frac{\partial K(v)}{\partial \sigma^2}\right)'',$$

with the following boundary condition, obtained by differentiating the value-matching condition (121) at $\bar{v}$:

$$\frac{\partial K}{\partial \sigma^2}(\bar{v}) = -\left(K'(\bar{v}) + U^{-1\prime}(r\bar{v})\right)\frac{\partial \bar{v}}{\partial \sigma^2} = 0.$$

A generalization of the Feynman–Kac formula (see DeMarzo and Sannikov, 2006 for the technical details) implies that the solution to this differential equation can be written as a conditional expectation:

$$\frac{\partial K(v)}{\partial \sigma^2} = \frac{1}{2}\mathbb{E}^{\boldsymbol{\theta}}\left[\int_0^\tau e^{-rt}\beta^2(\theta_t)K''(v_t)dt + e^{-r\tau}\frac{\partial K}{\partial \sigma^2}(\bar{v})|v_0 = v\right] < 0,$$

where v_t evolves according to (118), and where the inequality follows from the strict concavity of the profit function K. Intuitively, the right hand side of this equation sums the profit gains and losses along the path of v_t due to an increase in σ^2. This shows that a higher volatility σ^2 reduces the principal's profit. We can similarly evaluate the effects of all the parameters of the model on the principal's profit, the agent's time-0 utility (by differentiating the optimality condition $K'(v_0)=0$), or the value at retirement $\bar{v}$ (by differentiating the boundary conditions (121)).

Finally, another advantage of the continuous-time problem is that it is also more tractable computationally. In particular, the continuous-time formulation (120) can be computed more easily as the solution to an ordinary differential equation with a free boundary, while computing the solution to the discrete time Bellman equation (23) is more involved.

4. APPLICATIONS

In this section we discuss several applications of the theory of recursive contracts. The methods developed in the previous sections can be used to analyze questions in public economics, corporate finance, development, international finance, and political economy. Our goal is not to provide a comprehensive overview of those fields. Rather we want to show how several general principles emphasized above can be used to obtain rich insights in very different areas and relate those insights to empirical observations.

4.1 Public Finance

Individuals are subject to a variety of idiosyncratic shocks. Illness, disability, job loss, structural changes in the economy that diminish the value of human capital, unexpected promotions and demotions, success and failure in business ventures, all significantly affect individuals' incomes. It has been recognized at least since the work of Vickrey (1947) that the tax and transfer system can provide insurance against such shocks and help individuals smooth their consumption across different dates and states. A natural question is then how to design the optimal social insurance system that provides the best insurance given the distortions imposed by those programs.

Diamond and Mirrlees (1978, 1986) and Diamond et al. (1980) were some of the first papers to systematically study this question. At the same time, solving these problems either analytically or computationally is very difficult even in relatively simple dynamic settings. The advances in the theory of recursive contracts in the late 1980s and 1990s delivered a set of tools that allowed researchers to overcome many of the difficulties. The New Dynamic Public Finance literature applied those tools to the study of dynamic optimal taxation: see, eg, Golosov et al. (2003, 2006, 2016), Albanesi and Sleet (2006), Golosov and Tsyvinski (2006, 2007), Farhi and Werning (2013, 2012, 2007), Werning (2009), Kocherlakota (2010), Stantcheva (2014). In what follows, we describe a model that illustrates some of the main results of this literature.

We focus on a partial equilibrium model in which individuals are subject to idiosyncratic shocks to labor productivity.[ao] The economy lasts T periods, where T can be finite or infinite. Each agent's preferences are described by a time separable utility function over consumption $c_t \geq 0$ and labor supply $l_t \geq 0$,

$$\mathbb{E}_0\left[\sum_{t=1}^{T}\beta^{t-1}U(c_t,l_t)\right], \tag{124}$$

where $\beta \in (0,1)$ is a discount factor, $\mathbb{E}_0$ is a period-0 expectation operator conditional on the shock at date $t = 0$, and $U : \mathbb{R}^2_+ \rightarrow \mathbb{R}$ is differentiable, strictly increasing, and concave in consumption, and decreasing and concave in labor supply. The partial derivatives of the utility function are denoted by U_c and U_l.

Agents draw their initial type (skill) θ_1 from a distribution $\pi_1(\cdot)$ in period 1. From then on skills follow a Markov process $\pi_t(\theta_t|\theta_{t-1})$, where θ_{t-1} is the agent's skill realization in period $t-1$. We denote the probability density function of period-t types conditional on θ_{t-1} by $\pi_t(\cdot|\theta_{t-1})$. Skills are nonnegative: $\theta_t \in \Theta \subset \mathbb{R}_+$ for all t. At this stage we are agnostic about the dimensionality of Θ and allow Θ to be discrete or continuous. The set of possible histories up to period t is denoted by Θ^t. An agent of type θ_t who supplies l_t units of labor produces $y_t = \theta_t l_t$ units of output.

In this partial equilibrium economy, y_t also denotes the labor income of individuals. Individuals can freely borrow and lend at an exogenous interest rate R. We assume that there is no insurance available to individuals except self-insurance through borrowing and lending and through taxes and transfers provided by the government. We are interested in understanding how the government can design the optimal tax system $\mathcal{T}_t(\cdot)$ as a function of the information it has about individuals. We are thinking of the function $\mathcal{T}_t$ in very general terms: it is a combination of all taxes and transfers that individuals pay to or receive from the government. We are seeking a function $\mathcal{T}_t$ that maximizes welfare given by (124) in a competitive equilibrium.[ap]

If individuals' skills are observable, the optimal tax function is very simple: $\mathcal{T}_t$ should depend on the realization of the shocks θ^t and prescribe positive or negative transfers without distorting either labor supply or savings decisions. In reality idiosyncratic shocks are difficult to observe. Even disability insurance programs which extensively employ medical examinations to determine whether an applicant is subject to medical conditions that make a person unable to work are subject to substantial moral hazard problems and asymmetric information (see Golosov and Tsyvinski, 2006 and references therein).

[ao] See Albanesi (2011), Shourideh (2010), and Abraham and Pavoni (2008) for applications of recursive contracting tools to taxation with shocks to savings, Stantcheva (2014) for human capital accumulation, and Hosseini et al. (2013) for fertility choices.

[ap] It is straightforward to extend this analysis and allow other welfare criteria or expenditures on public goods (see, eg, Golosov et al., 2003).

Therefore we make the assumption that the realizations of θ_t are not observed by the government and the only observable choices are labor income, consumption, and capital.

We study the optimal taxes using a two-step procedure. In the first step, we invoke the Revelation Principle (see Section 2.2) and write the problem as a mechanism design program whose solution can be characterized using recursive techniques. In the second step, we back out a tax function $\mathcal{T}_t$ that can implement that solution in a competitive equilibrium.

The mechanism design problem is as follows. Let reports be given by $\sigma_t : \Theta^t \to \Theta$ and allocations by $c_t : \Theta^t \to \mathbb{R}_+$, $y_t : \Theta^t \to \mathbb{R}_+$, for all $t \geq 1$. The incentive constraint (8) writes

$$\begin{aligned} & \mathbb{E}_0\left[\sum_{t=1}^{T}\beta^{t-1}U\left(c_t(\theta^t), \frac{y_t(\theta^t)}{\theta_t}\right)\right] \\ \geq\ & \mathbb{E}_0\left[\sum_{t=1}^{T}\beta^{t-1}U\left(c_t(\sigma^t(\theta^t)), \frac{y_t(\sigma^t(\theta^t))}{\theta_t}\right)\right], \forall \sigma^T \in \Sigma^T, \end{aligned} \tag{125}$$

and the feasibility constraint (2) becomes

$$\mathbb{E}_0\left[\sum_{t=1}^{T}R^{1-t}c_t(\theta^t)\right] \leq \mathbb{E}_0\left[\sum_{t=1}^{T}R^{1-t}y_t(\theta^t)\right]. \tag{126}$$

The planner maximizes the ex ante expected utility (124) of the agents, ie, provides optimal ex ante insurance. This problem is thus similar to that analyzed in Section 2 (see Eqs. (9) or (74)). Solving this problem directly is difficult. There are prohibitively many incentive constraints (125) either for analytical or numerical analysis in most applications. In the next sections we overcome this problem using recursive techniques. For concreteness, we assume separable isoelastic preferences

$$U(c,l) = \frac{c^{1-\sigma}-1}{1-\sigma} - \frac{l^{1+\varepsilon}}{1+\varepsilon}. \tag{127}$$

While these preferences are not needed for most of the insights, they simplify the exposition of the main results.

4.1.1 Analysis with i.i.d. Shocks

We start the analysis by assuming that shocks are independent and identically distributed over time, so that the probability of realization of any $\theta \in \Theta$ in any period can be written as $\pi(\theta)$. This assumption, although unrealistic, allows us to illustrate many insights very transparently.

We follow the steps familiar from the analysis in Sections 2.3 and 2.4. To ensure convexity, we rewrite our maximization problem in terms of utils of consumption and leisure rather than c and l. To this end we define the functions $C(u) = [1 + (1-\sigma)u]^{1/(1-\sigma)}$ and

$Y(h)=[(1+\varepsilon)h]^{1/(1+\varepsilon)}$. We apply the one-shot deviation result from Propositions 2 and 3 to write the incentive-compatibility and promise-keeping constraints as:

$$u_t(\theta^t)-\theta_t^{-(1+\varepsilon)}h(\theta^t)+\beta v_t(\theta^t)\geq u_t\left(\theta^{t-1},\hat{\theta}\right)-\theta_t^{-(1+\varepsilon)}h\left(\theta^{t-1},\hat{\theta}\right)+\beta v_t\left(\theta^{t-1},\hat{\theta}\right)$$

for all $\theta^{t-1}\in\Theta^{t-1}$, $\theta_t\in\Theta$, $\hat{\theta}\in\Theta$, with

$$v_{t-1}\left(\theta^{t-1}\right)=\int_{\Theta}\left[u_t\left(\theta^{t-1},\theta\right)-\theta^{-(1+\varepsilon)}h\left(\theta^{t-1},\theta\right)+\beta v_t\left(\theta^{t-1},\theta\right)\right]d\pi(\theta).$$

Following the same steps as in Section 2.3 we write the Bellman equation as

$$K_t(v)=\min_{\{u(\theta),h(\theta),w(\theta)\}_{\theta\in\Theta}}\int_{\Theta}\left[C(u(\theta))-Y(h(\theta))+R^{-1}K_{t+1}(w(\theta))\right]d\pi(\theta) \tag{128}$$

subject to the incentive constraints: for all $\theta,\hat{\theta}\in\Theta$,

$$u(\theta)-\theta^{-(1+\varepsilon)}h(\theta)+\beta w(\theta)\geq u\left(\hat{\theta}\right)-\theta^{-(1+\varepsilon)}h\left(\hat{\theta}\right)+\beta w\left(\hat{\theta}\right),$$

the promise-keeping constraint:

$$v=\int_{\Theta}\left[u(\theta)-\theta^{-(1+\varepsilon)}h(\theta)+\beta w(\theta)\right]d\pi(\theta),$$

and $K_{T+1}(w)=0$ for all w if T is finite. When T is infinite, the subscript t drops out of the Bellman equation above.

Many of the qualitative properties of this model can be obtained along the lines of Proposition 5. For example, using steps analogous to those of Section 2.4, it is easy to show the analogue of Eqs. (29) and (76):[aq]

$$K_t'(v)=\mathbb{E}[C'(u_{v,t})]=(\beta R)^{-1}\mathbb{E}\left[K_{t+1}'(w_{v,t})|v\right]. \tag{129}$$

Moreover, optimality also requires: for all θ,t,v,

$$C'(u_{v,t}(\theta))=(\beta R)^{-1}K_{t+1}'(w_{v,t}(\theta)). \tag{130}$$

The intuition for this result is simple. The planner can provide incentives to reveal information either intratemporally, by giving an agent higher contemporaneous utility, or intertemporally, by giving higher future promises. Condition (130) implies that it is optimal to equalize the marginal costs of the two ways of providing incentives.

These conditions have some immediate but unexpected implications for taxation. Note that $C'=\frac{1}{U_c}$, where U_c is the marginal utility of consumption, and hence (129) can be rewritten as

[aq] This condition is particularly easy to derive if Θ is finite, in which case it can be obtained by simple manipulation of the Lagrangians on the incentive constraints.

$$\frac{\beta R}{U_c(c_{v,t}(\theta))} = \mathbb{E}\left[\frac{1}{U_c(c_{v,t+1})}\middle| v\right], \quad \forall v, t, \theta.$$

The policy functions generate the constrained-optimal stochastic processes $\{c_t^*, y_t^*\}_{t=1}^T$ which satisfy the Inverse Euler Equation (see our discussion in Section 2.7 as well as Golosov et al., 2003):

$$\frac{\beta R}{U_c(c_t^*)} = \mathbb{E}_t\left[\frac{1}{U_c(c_{t+1}^*)}\right]. \tag{131}$$

By Jensen's inequality, we have $\mathbb{E}\left[\frac{1}{X}\right] \geq \frac{1}{\mathbb{E}[X]}$ for any random variable X, with strict inequality if X is nondeterministic. Therefore this equation implies that at the optimum,

$$U_c(c_t^*) \leq \beta R \mathbb{E}_t\left[U_c(c_{t+1}^*)\right],$$

with strict inequality if future consumption is uncertain. Therefore, it follows that the optimal tax system must introduce positive savings distortions in this economy. One useful way to summarize the distortions introduced by the tax system is to define the *savings wedge* as

$$1 - \tau_t^s(\theta^t) \equiv \frac{1}{\beta R} \frac{U_c(c_t^*(\theta^t), y_t^*(\theta^t)/\theta_t)}{\mathbb{E}_t\left[U_c(c_{t+1}^*(\theta^{t+1}), y_{t+1}^*(\theta^{t+1})/\theta_{t+1})\right]}. \tag{132}$$

Optimality implies that $\tau_t^s(\theta^t) \geq 0$ for all θ^t, with strict equality if consumption in $t+1$ is uncertain.

Decentralization

We now describe how the government can design a tax system $\mathcal{T}_t$ such that agents optimally choose consumption and income $\{c_t^*, y_t^*\}_{t=1}^T$ given a budget constraint

$$c_t + k_{t+1} \leq y_t + Rk_t - \mathcal{T}_t.$$

That is, this tax function $\mathcal{T}_t$ is an *implementation* or *decentralization* of the constrained optimum. We want to understand on what arguments $\mathcal{T}_t(\cdot)$ should depend, and how to construct this function.

In general, there are many tax systems that implement the same allocation.[ar] Here we consider a particularly simple implementation that arises naturally from the recursive problem. Observe that to find the optimal allocations in period t in the Bellman equation

[ar] For example, an extreme tax system $\mathcal{T}_t(\{y_s\}_{s=1}^t)$ defined as $\mathcal{T}_t(\{y_s\}_{s=1}^t) = y_t^*(\theta^t) - c_t^*(\theta^t)$ if $y_s = y_s^*(\theta^s)$ for all $\theta^s \leq \theta^t$ and $\mathcal{T}_t(\{y_s\}_{s=1}^t) = \infty$ otherwise ensures that the only feasible choices for a consumer are $\{c_t^*, y_t^*\}_{t=1}^T$. Then the incentive-compatibility constraint ensures that $\mathcal{T}_t(\{y_s\}_{s=1}^t)$ implements $\{c_t^*, y_t^*\}_{t=1}^T$.

(128), we did not need to know the whole past history θ^t. It was sufficient to know the summary statistics $v_{t-1}\left(\theta^{t-1}\right)$ together with the current period shock θ_t. A natural analogue of the promised utility in competitive equilibrium is the agent's savings. Albanesi and Sleet (2006) use this insight to show that when types are i.i.d. and the utility function is separable between consumption and labor supply we can construct an optimal tax system in which taxes in period t depend only on labor income y_t and on savings k_t at the beginning of that period.

Proposition 15 *Assume that shocks are i.i.d. and preferences are separable in consumption and labor. The optimal allocations can be implemented by a tax system $\mathcal{T}_t(k_t, y_t)$.*

Proof We show this result in a two-period economy. Let $K_2(w_2)$ denote the planner's minimized cost function (128) in period 2, and $u^*_{w_2}(\theta), h^*_{w_2}(\theta)$ denote the policy functions that solve the second-period planner's problem.

In period 2, consider an individual who enters the period with savings k_2 and chooses labor income y_2. Suppose that $k_2 = K_2(w_2)$ for some promised utility w_2, and $y_2 = Y\left(h^*_{w_2}(\theta)\right)$ for some $\theta \in \Theta = \left[\underline{\theta}, \overline{\theta}\right]$. We then define the tax function $T_2(k_2, y_2)$ as[as]

$$T_2\left(K_2(w_2), Y\left(h^*_{w_2}(\theta)\right)\right) = K_2(w_2) + Y\left(h^*_{w_2}(\theta)\right) - C\left(u^*_{w_2}(\theta)\right).$$

By incentive compatibility, an agent with savings $k_2 = K_2(w_2)$ and type θ in period 2 chooses labor supply and consumption $(y_2, c_2) = \left(Y\left(h^*_{w_2}(\theta)\right), C\left(u^*_{w_2}(\theta)\right)\right)$, that is, the levels optimally assigned to his promised utility-type pair (w_2, θ).

In period 1, consider an individual who enters the period with savings k_1 and chooses labor income $y'{}_1$ (which may or may not be optimal given his first-period type θ). Denote by $\left(c_1', R^{-1}k_2'\right)$ his optimal consumption-savings choice given $\left(k_1, y_1'\right)$, and by $\tilde{u}' = U\left(c_1'\right) + \beta\mathbb{E}\left[V_2\left(k_2', \theta_2\right)\right]$ (where V_2 is the maximized objective of the agent in period 2) the utility that he achieves with this combination, gross of the first-period disutility of labor. We can show that the cost-minimizing way for the planner to deliver utility $\tilde{u}'$ to the agent is to offer the pair $\left(u_1', w_2'\right) = \left(U\left(c_1'\right), K_2^{-1}\left(k_2'\right)\right)$, and the corresponding cost is $C_{\tilde{u}'} = c_1' + R^{-1}k_2'$.

Now suppose that $y_1' = y_1^*(k_1, \theta')$ for some $\theta' \in \Theta = \left[\underline{\theta}, \overline{\theta}\right]$, where $y_1^*(k_1, \theta')$ denotes the first-period income optimally allocated to type θ' in the solution to the planner's problem. Define the tax function $T_1\left(k_1, y_1'\right)$ as

$$T_1\left(k_1, y_1^*(k_1, \theta')\right) = k_1 + y_1^*(k_1, \theta') - C_{\tilde{u}'}.$$

[as] The tax function can be easily extended to deter any move $y_2 < Y\left(h^*_{w_2}(\underline{\theta})\right)$ and $y_2 > Y\left(h^*_{w_2}(\overline{\theta})\right)$.

If the individual's true type is $\theta \neq \theta'$, by lying he reaches utility $\tilde{u}' - \theta^{-(1+\varepsilon)} h\left(y_1^*(k_1, \theta')\right)$. But by incentive compatibility this is smaller than the utility he gets by reporting his true type, namely $\tilde{u} - \theta^{-(1+\varepsilon)} h\left(y_1^*(k_1, \theta)\right)$. Thus under this tax function the agent finds it optimal to choose the income that corresponds to his true type in period 1, and his choice of savings will be exactly equal to $k_2 = K_2(w_2(\theta))$, since his net income is $C_{\tilde{u}}$. □

This proposition shows simultaneously that optimal allocations can be implemented by a joint tax on current period savings and labor income, and provides a method of constructing this tax.

When thinking about the relationship between this tax $\mathcal{T}_t$ and taxes in the data, it is important to keep in mind that $\mathcal{T}_t$ in the model corresponds to the sum of all taxes and transfers in the data. The marginal distortions with respect to capital and labor income, $\frac{\partial \mathcal{T}_t}{\partial k_t}$ and $\frac{\partial \mathcal{T}_t}{\partial y_t}$, correspond to the effective marginal tax rates in the data, which are a sum of statutory tax rates and the rates of phasing out of transfers in capital and labor income, respectively. Because of the phasing out of transfers, there is no reason to expect a priori that marginal taxes in the model and effective marginal taxes in the data are progressive.[at] For example, if individuals with more wealth receive less insurance against labor income shocks (eg, if they are not eligible to some welfare programs because of means-testing), we should expect the marginal labor taxes to be decreasing in capital.

4.1.2 Persistent Shocks

An important limitation of the previous discussion is the assumption that shocks are i.i.d. The empirical labor literature has emphasized that idiosyncratic shocks are highly persistent (for example, Storesletten et al., 2004 or Guvenen et al., 2015). In this section we discuss how to extend our analysis to persistent (Markov) shocks.

It is useful to assume, both for analytical tractability and for connecting the analysis to the empirical literature, that shocks are drawn from a continuous distribution. We focus on a family of stochastic processes frequently used in the applied labor and public finance literatures.[au]

[at] In the United States there is significant heterogeneity in the shapes of the effective tax rates as a function of income as they vary by state, family status, age, type of residence a person lives in, etc. Some typical patterns of the effective marginal rates in the US data are increasing, U-shaped, and inverted S-shaped (see CBO, 2007 and Maag et al., 2012).

[au] For example, Storesletten et al. (2004) and Farhi and Werning (2013) use lognormal distributions, Badel and Huggett (2014) and Lockwood et al. (2014) use Pareto-lognormal distributions, Geweke and Keane (2000) and Guvenen et al. (2015) use mixtures of lognormals.

Assumption 6 Suppose that shocks θ_t evolve according to

$$\ln\theta_t = b_t + \rho\ \ln\theta_{t-1} + \eta_t,$$

where e^{η_t} is drawn from one of the following three distributions:

(a) lognormal: $\eta_t \sim \mathcal{N}(0,\nu)$;

(b) Pareto-lognormal: $\eta_t \sim \mathcal{NE}(\mu,\nu,a)$, where $\mathcal{NE}$ is a normal-exponential distribution;

(c) mixture of lognormals: $\eta_t \sim \mathcal{N}(\mu_i,\nu_i)$ with probability p_i for $i=1,\ldots,I$; let $\nu = \max_i \nu_i$.

We can write the planner's problem recursively by applying the first-order approach discussed in Section 2.5.2. Under these assumptions the Bellman equation writes:

$$K_t(v,\hat{v},\theta_-) = \max_{\{u(\theta),h(\theta),w(\theta),\hat{w}(\theta)\}_{\theta\in\Theta}} \cdots \\ \int_0^\infty \big(Y(h(\theta)) - C(u(\theta)) + R^{-1}K_{t+1}(w(\theta),\hat{w}(\theta),\theta)\big)\pi(\theta|\theta_-)d\theta \tag{133}$$

subject to the promise-keeping and marginal promise-keeping constraints

$$v \ = \textstyle\int_0^\infty \varpi(\theta)\pi(\theta|\theta_-)d\theta, \tag{134}$$

$$\hat{v} \ = \textstyle\int_0^\infty \varpi(\theta)\hat{\pi}(\theta|\theta_-)d\theta, \tag{135}$$

$$\varpi(\theta) \ = u(\theta) - \theta^{-(1+\varepsilon)}h(\theta) + \beta w(\theta), \tag{136}$$

and the envelope condition

$$\dot{\varpi}(\theta) = (1+\varepsilon)\theta^{-(2+\varepsilon)}h(\theta) + \beta\hat{w}(\theta). \tag{137}$$

This problem can then be analyzed using optimal control techniques (see Golosov et al., 2016).

The analysis of savings distortions remains unchanged. In particular, the Inverse Euler Equation (131) continues to hold in this economy. The same arguments as in the previous section immediately imply the optimality of savings distortions.

We now turn to the analysis of labor distortions. We define the *labor wedge* as

$$1 - \tau_t^y(\theta^t) \equiv \frac{-U_l\big(c_t^*(\theta^t), y_t^*(\theta^t)/\theta_t\big)}{\theta_t U_c\big(c_t^*(\theta^t), y_t^*(\theta^t)/\theta_t\big)}. \tag{138}$$

To simplify the notations, for any history $\theta^t = \big(\theta^{t-1},\theta\big)$ and random variable x_t, we use the short-hand notations $x_t(\theta)$ to denote $x_t\big(\theta^{t-1},\theta\big)$ and x_{t-1} to denote $x_{t-1}\big(\theta^{t-1}\big)$. Manipulating the first-order conditions we obtain

$$\frac{\tau_t^y(\theta)}{1-\tau_t^y(\theta)}=(1+\varepsilon)\frac{\int_\theta^\infty \pi_t(x')dx'}{\theta\pi_t(\theta)}\int_\theta^\infty \frac{U_{c,t}(\theta)}{U_{c,t}(x)}\left(1-\int_0^\infty \frac{U_{c,t}(x)}{U_{c,t}(x')}\pi_t(x')dx'\right)\frac{\pi_t(x)dx}{\int_\theta^\infty \pi_t(x')dx'}$$
$$+\rho\beta R\frac{\tau_{t-1}^y}{1-\tau_{t-1}^y}\frac{U_{c,t}(\theta)}{U_{c,t-1}}. \tag{139}$$

Eq. (139) shows that the optimal labor distortion is the sum of two terms. The first (intratemporal) term on the right hand side captures the costs and benefits of labor distortions in providing insurance against period-t shocks. A labor distortion for type θ discourages that type's labor supply, as captured by the Frisch elasticity of labor supply $1/\varepsilon$. This lowers total output in proportion to $\theta\pi_t(\theta)$ but allows the planner to relax the incentive constraints for all types above θ, a trade-off summarized by the hazard ratio (of period-t shocks conditional on a given history θ^{t-1}), $\frac{\int_\theta^\infty \pi_t(x')dx'}{\theta\pi_t(\theta)}$. Finally, the relaxed incentive constraints allow the planner to extract more resources from individuals with skills above θ and transfer them to all agents. The social value of this transfer is captured by the integral term on the r.h.s., which depends on the marginal utilities of consumption of agents with skills above θ, weighted by the average marginal utility. The second term (intertemporal) on the right hand side captures how the planner uses distortions in the current period t to provide incentives for information revelation in earlier periods. It depends on the information that the period-t shock carries about θ^{t-1}, summarized by the coefficient ρ, and on the ratio $\frac{U_{c,t}(\theta)}{U_{c,t-1}}$ which captures the fact that it is cheaper to provide incentives in those states in which the marginal utility of consumption is high.

We can also use the decomposition (139) to obtain insights about the time series properties of the optimal labor distortions, as studied by Farhi and Werning (2013). Multiplying the expression above by $\frac{1}{U_{c,t}}\pi_t(\theta)$ and integrating by parts yields

$$\mathbb{E}_{t-1}\left[\frac{\tau_t^y}{1-\tau_t^y}\frac{1}{U_{c,t}}\right]=\rho\beta R\frac{\tau_{t-1}^y}{1-\tau_{t-1}^y}\frac{1}{U_{c,t-1}}+(1+\varepsilon)\mathrm{Cov}_{t-1}\left(\ln\theta,\frac{1}{U_{c,t}}\right). \tag{140}$$

Eq. (140) shows that the marginal utility-adjusted labor distortions follow an AR(1) process with a drift. The persistence of that process is determined by the persistence of the shock process ρ, and its drift is strictly positive since we should generally expect that $\mathrm{Cov}_{t-1}\left(\ln\theta,\frac{1}{U_{c,t}}\right)>0$. Farhi and Werning (2013) conclude that the optimal labor distortions should increase with age.

Golosov et al. (2016) use condition (139) to characterize the dependence of labor wedges on the realization of the shock θ. In particular they show the asymptotic laws of motion[av]

$$\frac{\tau_t^y(\theta)}{1-\tau_t^y(\theta)} \underset{\theta\to\infty}{\sim} \begin{cases} \left(\frac{a}{1+\varepsilon}-\frac{\sigma}{\sigma+\varepsilon}\right)^{-1}, & \text{if } \eta_t \text{ is Pareto} - \text{lognormal}, \dfrac{a}{1+\varepsilon}-\dfrac{\sigma}{\sigma+\varepsilon}>0 \\ \left(\frac{\ln\theta}{\nu^2}\frac{1}{1+\varepsilon}\right)^{-1}, & \text{if } \eta_t \text{ is lognormal or a mixture}, \end{cases} \tag{141}$$

and

$$\frac{\tau_t^y(\theta)}{1-\tau_t^y(\theta)} \underset{\theta\to 0}{\sim} \rho\beta R \frac{\tau_{t-1}^y}{1-\tau_{t-1}^y}\left(\frac{c_t(0)}{c_{t-1}}\right)^{-\sigma}. \tag{142}$$

Given the fact that $(\ln\theta)^{-1}$ is very slowly moving, Eq. (141) implies that the labor distortions are approximately flat for high realizations of θ_t for all three classes of distributions (although in the cases of lognormal and mixture of lognormal distributions they eventually converge to zero), they do not depend on the past history of shocks, and they are given by relatively simple closed-form expressions. Eq. (142) shows that the labor distortions for low shocks depend on the persistence, the past history, and the growth rate of consumption, and are generally increasing in age.

Another implication of these equations is that the higher moments, such as the kurtosis, play an important qualitative and quantitative role for the size of the labor distortions. Some of the best estimates of those moments are obtained by Guvenen et al. (2014, 2015) who use US administrative data on a random sample of 10% of the US male taxpayers to estimate the stochastic process for labor earnings. Golosov et al. (2016) use that finding to calibrate their model using newly available estimates of idiosyncratic shocks. The optimal labor distortions are U-shape, while savings distortions are increasing in current earnings. Welfare in the constrained optimum is 2–4% higher than in the equilibrium with affine taxes. These findings (both the U-shaped and the relatively high welfare gains from nonlinear, history-dependent taxation) are largely driven by the high kurtosis found in the labor earnings process in the data. This suggests that a system of progressive taxes and history-dependent transfers that are being phased out relatively quickly with income can capture most of the welfare gains in this economy.

4.2 Corporate Finance

In this section we describe some applications of the recursive contract theory to corporate finance. We show how financing frictions arise endogenously from agency problems, leading to implications for the capital structure and dynamics of firms. To cite only a few papers in this literature, such models have been analyzed by Albuquerque and Hopenhayn (2004), Clementi and Hopenhayn (2006), and DeMarzo and Fishman

[av] For any functions h,g and $c\in\bar{\mathbb{R}}$, $h(x)\underset{x\to c}{\sim}g(x)$ if $\lim_{x\to c}h(x)/g(x)=1$.

(2007a,b) in discrete-time environments, and by DeMarzo and Sannikov (2006), Biais et al. (2007), DeMarzo et al. (2012), Biais et al. (2010), and He (2009) in continuous-time environments.

Endogenous Financing Frictions and Firm Dynamics

A large empirical literature (see, eg, Caves, 1998 for a survey) describes the properties of firm dynamics, eg, the characteristics and evolution of their size, growth rate, and survival probability. In particular, as firms get older, their size and survival probability increase, the mean and variance of their growth rates decrease, and the hazard rates for exit first increase and then decrease. Moreover, starting with the work of Fazzari et al. (1988), many authors have found that firms' investment responds positively to innovations in the cash-flow process (after controlling for Tobin's q), suggesting the importance of borrowing constraints, and that the investment-cash flow sensitivity decreases with the firm's age and size.

Clementi and Hopenhayn (2006) analyze a dynamic moral hazard model where such features arise *endogenously* in the optimal contract between a borrower (a firm, or agent) and a lender (bank, or principal) who cannot observe the outcome of the project. They describe the optimal contract and show that the model yields rich testable predictions about firm dynamics that are in line with the evidence presented above.

In their model, the agent's project requires a fixed initial investment $I_0 > 0$, and subsequently a per-period investment of capital which we denote by k_t. At the beginning of each period the bank can liquidate the project, with scrap value $S \geq 0$. If the bank decides to finance the project, the firm's revenues are stochastic (i.i.d.) and increase with the amount of capital k_t advanced by the lender. Specifically, in each period t, with probability π the project is successful and yields revenue $R(k_t)$, where R is continuous, bounded, and concave, whereas with probability $(1-\pi)$ it yields zero revenues. Denote the outcome of the project in period t by $\theta_t \in \Theta = \{\theta_{(1)}, \theta_{(2)}\} = \{0, 1\}$, where $\theta_{(1)} = 0$ is failure and $\theta_{(2)} = 1$ is success, and histories up to period t by θ^t. The borrower and the lender are both risk-neutral, have the same discount factor β, and have the ability to commit to contracts.

Suppose first that revenues are observable. The efficient amount of capital is advanced in every period, $k^* = \arg\max(\pi R(k_t) - k_t)$, and running the project is efficient if $W^* \equiv \dfrac{1}{1-\beta}(\pi R(k^*) - k^*) > I_0$. Thus, in the benchmark complete-information version of the model, the firm neither grows, nor shrinks, nor exits: its size k^* is constant. This feature allows us to cleanly analyze the implications of informational frictions on the firm's dynamics.

Now suppose that revenues are private information to the agent, so that the lender must rely on the borrower's reports of the outcome of the project in each period. Denote by $\boldsymbol{\sigma} = \{\sigma_t(\theta^t)\}_{t \geq 1}$ the borrower's reporting strategy.

The timing of events is as follows. At the beginning of each period t, the bank decides whether (and if so, with which probability) to liquidate the firm, in which case it gets the scrap value S and compensates the agent with a transfer $Q_t \geq 0$. Denote by $\boldsymbol{\alpha} = \left\{\alpha_t\left(\sigma^{t-1}\left(\theta^{t-1}\right)\right)\right\}_{t\geq 1}$ the liquidation probabilities and by $\mathbf{Q} = \left\{Q_t\left(\sigma^{t-1}\left(\theta^{t-1}\right)\right)\right\}_{t\geq 1}$ the transfers from the lender to the borrower in case of liquidation. Then, if the firm is not liquidated, the bank chooses the amount of capital k_t it lends to the firm, and the borrower's repayment τ_t if the project is successful. Denote by $\boldsymbol{k} = \left\{k_t\left(\sigma^{t-1}\left(\theta^{t-1}\right)\right)\right\}_{t\geq 1}$ the capital advancements and by $\boldsymbol{\tau} = \left\{\tau_t\left(\sigma^{t-1}\left(\theta^{t-1}\right), \sigma_t(\theta_t)\right)\right\}_{t\geq 1}$ the contingent payments from the borrower to the lender in case of success (there is no transfer in case of failure). The firm is restricted at all times to have a nonnegative cash flow, ie, the following limited-liability constraint must be satisfied: $\tau_t\left(\sigma^{t-1}\left(\theta^{t-1}\right), \theta_{(2)}\right) \leq R\left(k_t\left(\sigma^{t-1}\left(\theta^{t-1}\right)\right)\right)$ for all t, θ^{t-1}. The outcome θ_t of the project is then realized and privately observed by the borrower, who sends a report $\sigma_t(\theta^t)$ and transfers $\tau_t(\sigma^t(\theta^t))$ to the bank in case of a (truthfully reported) success.

We define "equity" as the entrepreneur's share of the total firm's value, and "debt" as the lender's share. That is, equity $V_t\left(\{\boldsymbol{k}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \mathbf{Q}, \boldsymbol{\sigma}\}, \sigma^{t-1}\left(\theta^{t-1}\right)\right)$ and debt $B_t\left(\{\boldsymbol{k}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \mathbf{Q}, \boldsymbol{\sigma}\}, \sigma^{t-1}\left(\theta^{t-1}\right)\right)$ are the expected discounted cash flows (or continuation values) accruing to the borrower and the lender, respectively, under the contract $\{\boldsymbol{k}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \mathbf{Q}\}$ and reporting strategy $\boldsymbol{\sigma}$, given the time-t history of reports $\sigma^{t-1}\left(\theta^{t-1}\right)$. Note that the value of equity corresponds to the promised utility variable (11) in the taste shock model of Section 2.

This setup is formally similar to that in Section 2.3 and can be analyzed using the same recursive techniques. Specifically, we write the problem in recursive form using the value of equity v as the state variable. We can show that the set of continuation values v that can be supported by a feasible contract is $[0, \infty)$. A constrained-efficient contract maximizes the value obtained by the lender, $B(v_0)$, in the space of incentive-compatible and feasible contracts, subject to delivering some utility $v_0 \geq 0$ to the entrepreneur. The pair $(v_0, B(v_0))$ defines the capital structure of the firm (equity and debt) and implies a total value for the firm $W(v_0) = v_0 + B(v_0)$.

Denote by $W(\cdot)$ the total value of the firm prior to the liquidation decision, and by $\hat{W}(\cdot)$ the total value of the firm conditional on not being liquidated. Following the steps of Proposition 3, we obtain that the latter value function is given by the following Bellman equation:

$$\hat{W}(\hat{v}) = \max_{k, \tau, \{w(\theta)\}} (\pi R(k) - k) + \beta\left[(1-\pi) W\left(w\left(\theta_{(1)}\right)\right) + \pi W\left(w\left(\theta_{(2)}\right)\right)\right]$$

subject to the promise-keeping constraint:

$$\hat{v} = \pi(R(k) - \tau) + \beta\left[(1-\pi) w\left(\theta_{(1)}\right) + \pi w\left(\theta_{(2)}\right)\right], \tag{143}$$

the incentive-compatibility constraint in the high state:

$$R(k)-\tau+\beta w\left(\theta_{(2)}\right)\geq R(k)+\beta w\left(\theta_{(1)}\right), \tag{144}$$

and the limited liability constraint:

$$\tau\leq R(k). \tag{145}$$

The liquidation decision of the firm can be formalized as follows. At the beginning of the period, the firm is liquidated with probability α, in which case the borrower receives Q, and it is kept in operation with probability $1-\alpha$, in which case the borrower receives the continuation value $\hat{v}$. The value function $W(\cdot)$ then solves the following Bellman equation:

$$W(v)=\max_{\alpha,\,Q,\,\hat{v}}\alpha S+(1-\alpha)\hat{W}(\hat{v})$$

subject to

$$v=\alpha Q+(1-\alpha)\hat{v}.$$

Clementi and Hopenhayn (2006) characterize the solution to this problem as follows.

First, if the equity (or promised value) v is large enough, then the policy of providing the unconstrained efficient level of capital k^* in every period is both feasible and incentive compatible. The minimum value v^* for which this is the case is given by the solution to the following problem:

$$\begin{aligned} v^*\equiv \min_{\tau,\{w(\theta)\}_{\theta\in\Theta}} \quad & \pi(R(k^*)-\tau)+\beta\left[(1-\pi)w\left(\theta_{(1)}\right)+\pi w\left(\theta_{(2)}\right)\right] \\ \text{subject to} \quad & R(k^*)-\tau+\beta w\left(\theta_{(2)}\right)\geq R(k^*)+\beta w\left(\theta_{(1)}\right), \\ & \tau\leq R(k^*),\quad w\left(\theta_{(1)}\right)\geq v^*,\quad w\left(\theta_{(2)}\right)\geq v^*. \end{aligned}$$

Solving this problem yields $v^*\equiv\dfrac{1}{1-\beta}\pi R(k^*)=W^*+\dfrac{1}{1-\beta}k^*$.

Second, we can show that there exists a value $v_*\in(0,v^*)$ such that:

(i) when $v\geq v^*$, the firm's value $W(v)$ is equal to W^*. Letting $k_v=k^*$ at any future date, with $\tau_v=0$ and $w_v\left(\theta_{(j)}\right)=v^*$ for all $j\in\{1,2\}$, is optimal.

(ii) when $v\in[v_*,v^*)$, the value function $W(v)$ is strictly increasing and concave and the policy functions are $\alpha(v)=0$, $k_v<k^*$, $\tau_v=R(k_v)$, and $w_v\left(\theta_{(1)}\right)<v<w_v\left(\theta_{(2)}\right)$. The values $w_v\left(\theta_{(1)}\right),w_v\left(\theta_{(2)}\right)$ are given as a function of k_v by the promise-keeping and incentive-compatibility constraints (143), (144), which both hold with equality.[aw]

[aw] If these values are such that $w_v\left(\theta_{(2)}\right)>v^*$, then other values for the transfer τ_v (along with $w_v\left(\theta_{(2)}\right)$) are also optimal.

Moreover, k_v is increasing in v for v close enough to v^*, $w_v(\theta_{(1)}), w_v(\theta_{(2)})$ are increasing in v, and equity is a submartingale, ie, $v < \mathbb{E}[w_v]$.

(iii) when $v < v_*$, the firm is liquidated with positive probability $\alpha(v) = 1 - v/v_*$ and transfer $Q = 0$, and continues at value $\hat{v} = v_*$ with probability $(1 - \alpha(v))$. The firm's value is equal to $W(v) = \alpha(v)S + (1 - \alpha(v))\hat{W}(v_*)$.

This characterization of the optimal contract has the following interpretation and implications. The contract determines stochastic processes for the firm size k_t, equity v_t, and debt $B(v_t) = W(v_t) - v_t$. Specifically, consider an entrepreneur who starts with equity $v_0 \in (v_*, v^*)$. Starting from this region, a good shock raises the value of equity to $w_v(\theta_{(2)}) > v$, and a bad shock reduces it. The submartingale property (which follows from Eq. (143)) implies that the equity v_t of surviving firms on average increases over time, and the monotonicity of the functions $w_v(\theta)$ implies that this process v_t displays persistence. Eventually, equity reaches either the lower threshold v_* (after a series of negative shocks), leading to the region where it is optimal to liquidate the firm with positive probability, or the upper threshold v^* (after a series of positive shocks), at which point the incentive constraints no longer bind and the unconstrained efficient level of capital k^* is advanced from then on. There are therefore two absorbing states: either the firm is liquidated or it attains its efficient size. In the transition, the transfer τ_v in the event of a good shock is set equal to the maximum possible amount $R(k_v)$. This is because the bank and the firm are both risk-neutral, so that it is optimal to backload the distribution of dividends to the borrower (by choosing the highest possible value of transfers τ and raising $w_v(\theta_{(j)})$ accordingly) in order to allow the equity to reach v^* as fast as possible. Finally, when v^* is attained, the firm's future cash flows are equal to $v^* = W(v^*) + \frac{k^*}{1-\beta} = W^* + \frac{k^*}{1-\beta}$, and the lender's continuation value is $B(v^*) = -\frac{k^*}{1-\beta}$. This means that the entrepreneur has accumulated assets at the bank (at the interest rate r such that $\beta = \frac{1}{1+r}$) up to the positive balance $k^*/(1-\beta)$, while his payments were being postponed and all the cash flows were received by the lender; this balance is exactly enough to self-finance the project at the efficient scale from then on.

Next, the optimal contract shows that when equity v is below the threshold v^*, the amount of capital advanced by the bank is strictly smaller than the unconstrained efficient level: $k_v < k^*$. We can interpret this result as an (endogenous) borrowing constraint to which the entrepreneur is subject. Moreover, if v is close enough to v^*, higher equity relaxes the borrowing constraint and allows the entrepreneur to finance the project on a more efficient scale, as k_v is increasing in v. Such financing frictions arise endogenously in the optimal contract due to moral hazard. To provide incentives for the successful entrepreneur to truthfully report the (good) outcome of his project, the optimal contract requires the borrower's compensation to be sensitive to reported output, which necessitates a spread $w_v(\theta_{(2)}) - w_v(\theta_{(1)})$ between the future equity values in the successful versus unsuccessful states. Moreover, advancing more capital today tightens the incentive constraint (as the borrower will have to repay more in case of success, since $\tau = R(k)$) and thus requires a larger spread between future continuation values. But this spread is costly, because the marginal revenue is decreasing in capital and hence the firm's total value $W(\cdot)$ is concave. Therefore, the trade-off between higher capital and profits today against a lower firm's value in future periods implies an inefficient level of financing $k_v < k^*$ in the optimal contract.

These results imply that revenue shocks affect the financial structure $(v, B(v))$ of the firm, and yield rich implications for firm dynamics (size, growth, and survival probability). Defining the firm's size as the level of capital k_t invested in the project, investment as $k_t - k_{t-1}$, and simulating a calibrated version of the model, the authors obtain the following testable predictions. First, firm age and size are positively correlated. Second, the mean and variance of growth decrease with size and age. Third, the survival probability $\mathbb{P}(T > t|v)$, where T is the stopping time for exit, increases with the value of equity v and thus with age. The hazard rates for exit follow an inverted U-shaped function of age, as it takes a few periods for young firms to reach the liquidation region from their initial value v_0, and a selection effect implies that older (surviving) firms have on average higher values and hence lower hazard rates. All these properties are consistent with the empirical evidence on firm dynamics (see the references in Clementi and Hopenhayn, 2006 for a survey of the empirical literature). Finally, the authors argue that simulated data generated using the policy functions of the model would reproduce the empirical prediction that investment responds positively to innovations in the cash-flow process, and that the sensitivity of investment to cash flows decreases with the age and size of the firm. Importantly, in the model, the financing frictions (borrowing constraints) arise *endogenously* as a feature of the optimal contract.

Optimal Capital Structure

We now describe another application of recursive contracts to corporate finance in a continuous-time framework using the techniques described in Section 3.3.2, following a simple version of DeMarzo and Sannikov (2006).[ax] In their model the agent (firm) can unobservably divert cash flows for its private benefit; investors control its wage and choose when to liquidate the project. While the closely related framework of Clementi and Hopenhayn (2006) focused on the importance of informational frictions for firm investment and growth as a function of the history of profit realizations (so that

[ax] A discrete time version of this problem has been analyzed by DeMarzo and Fishman (2007b).

the scale, ie, the capital, of the firm is an endogenous part of the optimal incentive contract), DeMarzo and Sannikov (2006) assume instead that the firm has a fixed size, and they focus on the optimal choice of the firm's capital structure.[ay] Specifically, they propose an implementation of the optimal contract using simple financial instruments. This implementation is composed of a combination of long-term debt with a constant coupon, a credit line, and equity. In this implementation the firm is compensated by holding a fraction of the equity, and defaults if debt service payments are not made or the credit line is overdrawn; dividends are paid when cash flows exceed debt payments and the credit line is paid off. This analysis can therefore help understand the choice between various forms of borrowing for firms, in particular the characteristics of credit line contracts, an empirically important component of firm financing. Finally, as we saw in Section 3.3.2, setting the model in continuous time allows the authors to obtain both a clean characterization of the optimal contract through an ordinary differential equation, and analytical comparative statics of the optimal contract with respect to the parameters of the model.

We now turn to a formal description of the model. An agent manages a project that generates stochastic cash flows given by:

$$d\hat{y}_t = (\mu - \theta_t)dt + \sigma d\mathcal{Z}_t,$$

where $\mathcal{Z}_t$ is a standard Brownian motion, and $\theta_t \geq 0$ is the agent's private action, which can be interpreted as cash flow diversion. This unobserved diversion generates private benefit to the agent at rate $\lambda\theta_t$, with $\lambda \in (0,1]$. The principal observes only the reported cash flows $\{\hat{y}_t\}_{t\geq 0}$.[az] The principal and the agent are risk-neutral and discount the future at rate r and γ, respectively, with $r < \gamma$. The project requires an external capital of $I_0 \geq 0$ to be started. The principal offers a contract $(\mathbf{c}, \tau)$ that specifies the agent's compensation $dc_t \geq 0$ for all t and a termination date τ, as functions of the histories $\{\hat{y}_s\}_{s\leq t}$. In the event of termination, the agent gets his outside option $R \geq 0$ and the principal receives the liquidation payoff $L \geq 0$.

The optimal contract maximizes the principal's expected profit subject to delivering expected utility $\hat{v}_0$ to the agent and the incentive-compatibility constraints. We can show that in the optimal contract we have $\theta_t = 0$ for all $t \geq 0$. The problem is similar to the model analyzed in Section 3.3.2 and can be expressed as:

$$\max_{\mathbf{c},\tau} \mathbb{E}^{\boldsymbol{\theta}=0}\left[\int_0^{\tau} e^{-rt}(d\hat{y}_t - dc_t) + e^{-r\tau}L\right]$$

[ay] DeMarzo et al. (2012) extend this model to include investment and nonconstant firm size.

[az] DeMarzo and Sannikov (2006) consider a more general model in which the agent can secretly save and thus overreport, ie, $\theta_t < 0$, but show that in the optimal contract the agent always chooses to maintain zero savings.

subject to the promise-keeping constraint:

$$\hat{v}_0 = \mathbb{E}^{\boldsymbol{\theta}=0}\left[\int_0^{\tau} e^{-\gamma t} dc_t + e^{-\gamma \tau} R\right]$$

and the incentive-compatibility constraints:

$$\hat{v}_0 \geq \mathbb{E}^{\hat{\boldsymbol{\theta}}}\left[\int_0^{\tau} e^{-\gamma t}\left(dc_t + \lambda \hat{\theta}_t dt\right) + e^{-\gamma \tau} R\right],$$

for any deviation strategy $\hat{\boldsymbol{\theta}}$.

Following identical steps as in Section 3.3.2 (see in particular Proposition 12), we find that there is a one-to-one correspondence between incentive-compatible contracts $(\mathbf{c}, \tau)$ and controlled processes (with controls (c_t, β_t))

$$dv_t = \gamma v_t dt - dc_t + \beta_t (d\hat{y}_t - \mu dt), \tag{146}$$

where the sensitivity of the agent's promised value to his report satisfies $\beta_t \geq \lambda$ for all $t \leq \tau$. The termination time τ is the earliest time that the agent's promised value v_t reaches R. The one-shot incentive constraint (Proposition 13) here says that truthtelling is incentive compatible if and only if $\beta_t \geq \lambda$ for all t, since the agent has incentives not to steal cash flows if he gets at least λ of promised value for each reported dollar.

DeMarzo and Sannikov (2006) characterize the optimal contract as follows. Denote by $K(v)$ the principal's value function. It is easy to see that the optimal contract must satisfy $K'(v) \geq -1$ for all v. This is because the principal can always give to the agent with current promised utility v a lump-sum transfer $dc > 0$ and then revert to the optimal contract with utility $v - dc$, so that $K(v) \geq K(v - dc) - dc$. Defining $\bar{v}$ as the lowest value such that $K'(\bar{v}) = -1$, it is optimal to keep the agent's promised utility in the range $[R, \bar{v}]$ and to set $dc_v = (v - \bar{v})\mathbb{I}_{\{v \geq \bar{v}\}}$. The function $K(v)$ can then be characterized recursively as in Section 3.3.2. The Hamilton–Jacobi–Bellman equation is

$$rK(v) = \max_{\beta \geq \lambda} \mu + \gamma v K'(v) + \frac{1}{2}\beta^2 \sigma^2 K''(v),$$

$$\text{with} \quad K(v) = K(\bar{v}) - (v - \bar{v}) \text{ for } v > \bar{v},$$

with the following value-matching, smooth-pasting, and super-contact conditions

$$K(R) = L, \quad K'(\bar{v}) = -1, \quad K''(\bar{v}) = 0.$$

The function $K(\cdot)$ is concave so that it is optimal to set $\beta_t = \lambda$ for all t. The optimal contract (with $\hat{v}_0 \in [R, \bar{v}]$) is such that v_t evolves according to (146) with $dc_t = 0$ when

$v_t \in [R, \bar{v})$. If $v_t = \bar{v}$, payments dc_t cause v_t to reflect at $\bar{v}$. The contract is terminated at time τ when v_t reaches R.[ba]

DeMarzo and Sannikov (2006) propose an implementation of the optimal contract using equity, long-term debt D, and a credit line C^L. If the agent defaults on a debt coupon payment or his credit balance exceeds C^L, the project is terminated. The idea behind this implementation is to map the interval of continuation values $[R, \bar{v}]$ into a credit line, with point $\bar{v}$ corresponding to balance 0. From (146), we can write the evolution of the credit balance $\lambda^{-1}(\bar{v} - v_t)$ (where λ is simply a normalization) as

$$d\left(\frac{\bar{v} - v_t}{\lambda}\right) = -d\hat{y}_t + \left\{\gamma\left(\frac{\bar{v} - v_t}{\lambda}\right)dt + \left(\mu - \frac{\gamma}{\lambda}\bar{v}\right)dt + \frac{dc_t}{\lambda}\right\}.$$

The first term in the right hand side of this expression, $-d\hat{y}_t$, is the credit balance reduction due to the cash flows, where each dollar of cash flow subtracts exactly one dollar from the credit line balance. The next three terms in the right hand side (inside the brackets) are the three components that compose the implementation of the contract. The first term inside the brackets is the interest charged on the credit balance $\lambda^{-1}(\bar{v} - v_t)$, so that the implementation of the optimal contract has a credit line $C^L = \lambda^{-1}(\bar{v} - R)$, up to which credit is available to the firm at interest rate γ. The second term inside the brackets is the coupon rD on long-term debt, so that the face value of the debt is $D = r^{-1}(\mu - \gamma\bar{v}/\lambda)$. Finally the third term inside the brackets consists of the dividend payments made by the firm, ie, the equity. The agent gets a fraction λ of the dividends dc_t, while outside investors hold the remaining firm's equity, debt, and credit line. Cash flows in excess of the debt coupon payments are issued as dividends once the credit line is fully repaid. Termination occurs when the credit line balance reaches the credit limit C^L. Observe that the balance on the credit line fluctuates with the past performance of the firm, in particular leverage decreases with its profitability since the firm pays off the credit line when it makes profits.

DeMarzo and Sannikov (2006) further analyze this optimal capital structure, ie, how the amount of long-term debt and the size of the credit line depend on the parameters of the model, by deriving analytical comparative statics using the techniques described in Section 3.3.2. We refer the reader to the original paper for an in-depth analysis of these questions.

[ba] In the discrete-time setting described in the previous section (based on the work of Clementi and Hopenhayn (2006)), allowing for randomization over the decision to terminate the project could improve the contract. DeMarzo and Sannikov (2006) show that in the continuous-time framework, such randomization is not necessary: without loss of generality the termination time τ is based only on the firm's (reported) past performance.

4.3 Development Economics

There is a large literature that studies informal insurance arrangements in the context of village economies. An early work of Townsend (1994), for example, showed that in rural India idiosyncratic variation in consumption is systematically related to idiosyncratic variation in income, implying that households can only achieve partial insurance against their idiosyncratic risks. Models of limited commitment developed by Thomas and Worrall (1988), Kehoe and Levine (1993), Kocherlakota (1996), Alvarez and Jermann (2000), and Ligon et al. (2000, 2002) can potentially explain these observations. In these models, all the information is public (there is no information friction); instead there is an *enforcement* friction: agents are free to walk away from the insurance contract at any time. Nevertheless, these models can be analyzed using the same recursive techniques as those described in Section 2. Analogous to the asymmetric information models we analyzed, the state variable is the utility promised to the agent. The only formal difference is that the incentive-compatibility constraints (8) are replaced by participation constraints that we formally define in Eq. (147).

Here we describe the two-sided limited commitment framework analyzed by Ligon et al. (2002). The (observable) period-t state of nature $\theta_t \in \Theta = \left\{\theta_{(1)}, \ldots, \theta_{(|\Theta|)}\right\}$ is stochastic and follows a Markov process with transition probability $\pi\left(\theta_{(i)} \middle| \theta_{(j)}\right) > 0$ for all i, j. There are two agents with period-t utilities $U^1\left(c_t^1\right), U^2\left(c_t^2\right)$ and exogenous nonstorable endowments $\left(y_t^1, y_t^2\right)$ determined by θ_t. At least one of the two households is risk averse, and they both discount the future at rate β. A risk-sharing contract $\boldsymbol{\tau}$ specifies for every date t and history θ^t a (possibly negative) transfer $\tau_t(\theta^t)$ from household 1 to household 2. A first-best, or full risk-pooling, contract $\boldsymbol{\tau}$ is such that the ratio of marginal utilities $\dfrac{U^{2\prime}\left(y_t^2(\theta_t) + \tau_t(\theta^t)\right)}{U^{1\prime}\left(y_t^1(\theta_t) - \tau_t(\theta^t)\right)}$ is constant across all histories and dates, so that each individual's consumption is only a function of the aggregate endowment.

The key friction of the model is that agents can walk away from the insurance contract, after which both households consume at autarky levels forever after, ie, $\tau_t(\theta^t) = 0$ for all t, θ^t. Household $j \in \{1, 2\}$ has no incentive to break the contract if the following *sustainability constraint* holds: for all $\theta^t \in \Theta^t$,

$$U^j\left(c_t^j(\theta^t)\right) + \mathbb{E}_t\left[\sum_{s=1}^{\infty} \beta^s U^j\left(c_{t+s}^j(\theta^{t+s})\right)\right] \geq U^j\left(y_t^j(\theta_t)\right) + \mathbb{E}_t\left[\sum_{s=1}^{\infty} \beta^s U^j\left(y_{t+s}^j(\theta_{t+s})\right)\right], \tag{147}$$

where $c_s^1(\theta^s) = y_s^1(\theta_s) - \tau_s(\theta^s)$ and $c_s^2(\theta^s) = y_s^2(\theta_s) + \tau_s(\theta^s)$ for all s, and where $\mathbb{E}_t$ is the expectation conditional on θ^t.

As in Section 3.2 (where the government was unable to commit), it is useful to describe the present environment with bilateral lack of commitment as a repeated game between the two agents. Since reversion to autarky is the worst subgame-perfect punishment, there is a one-to-one relationship between sustainable contracts and subgame-perfect equilibria (see Abreu, 1988).

We now show how to characterize the set of constrained-efficient sustainable contracts, using recursive arguments formally similar to those we used in Section 2. The constrained efficient allocations maximize the expected lifetime utility of agent 2 subject to both sustainability constraints (147), and to delivering at least a given utility level v^1 to agent 1, given that the current state is θ. Before we formally write this problem, we describe the space of discounted expected utilities v^1,v^2 for each agent (defined as in (11)) for which there exists a sustainable contract that delivers those values, given that the current state is θ. We can show that this set is an interval of the form $\left[\underline{v}^j(\theta), \bar{v}^j(\theta)\right]$ for each agent $j \in \{1,2\}$, where the minimum sustainable utilities when the current state is θ are

$$\underline{v}^j(\theta) = U^j\left(y^j(\theta)\right) + \mathbb{E}\left[\sum_{s=1}^{\infty} \beta^s U^j\left(y_s^j(\theta_s)\right) | \theta\right]$$

for $j \in \{1,2\}$, that is, the value of autarky for agent j from state θ onward.

The ex post efficiency frontier, calculated once the current state θ is known, can then be characterized in recursive form as follows: for $v^1 \in \left[\underline{v}^1(\theta), \bar{v}^1(\theta)\right]$,

$$V\left(v^1,\theta\right) = \max_{\tau(\theta),\{w^1(\theta')\}_{\theta' \in \Theta}} \quad U^2\left(y^2(\theta) + \tau(\theta)\right) + \beta \sum_{\theta' \in \Theta} \pi(\theta'|\theta) V\left(w^1(\theta'),\theta'\right)$$

subject to the promise-keeping constraint

$$U^1\left(y^1(\theta) - \tau(\theta)\right) + \beta \sum_{\theta' \in \Theta} \pi(\theta'|\theta) w^1(\theta') = v^1, \tag{148}$$

the sustainability constraints

$$w^1(\theta') \geq \underline{v}^1(\theta'), \forall \theta', \tag{149}$$

$$V\left(w^1(\theta'),\theta'\right) \geq \underline{v}^2(\theta'), \forall \theta', \tag{150}$$

(the constraint $w^1(\theta') \leq \bar{v}^1(\theta')$ is equivalent to (150)), and the nonnegativity constraints

$$y^1(\theta) - \tau(\theta) \geq 0 \quad \text{and} \quad y^2(\theta) + \tau(\theta) \geq 0. \tag{151}$$

The Lagrange multiplier λ associated with the constraint (148) is the key variable in the analysis of optimal insurance contracts. The first-order conditions and envelope condition of the problem imply that λ is related to the ratio of the marginal utilities of consumption by

$$\lambda = -\frac{\partial}{\partial v} V\left(v^1,\theta\right) = \frac{U^{2\prime}(y^2(\theta) + \tau(\theta))}{U^{1\prime}(y^1(\theta) - \tau(\theta))} + \frac{\psi_2 - \psi_1}{U^{1\prime}(y^1(\theta) - \tau(\theta))}, \tag{152}$$

where ψ_1,ψ_2 are the Lagrange multipliers associated with the nonnegativity constraints (151).

Suppose that the value of λ is known. If λ is in the set of marginal utility ratios $\frac{U^{2\prime}(y^2(\theta)+\tau(\theta))}{U^{1\prime}(y^1(\theta)-\tau(\theta))}$ which can be generated by feasible transfers in state θ (ie, by $\tau(\theta)\in[-y_2(\theta),y_1(\theta)]$), then there is a unique interior solution and the value of the transfer $\tau(\theta)$ is pinned down by Eq. (152) with $\psi_1 = \psi_2 = 0$. Otherwise, there is a corner solution with all income going to one of the households, ie, $\tau(\theta) = -y_2(\theta)$ or $\tau(\theta) = y_1(\theta)$ (with a positive multiplier ψ_2 or ψ_1, respectively).

Therefore the constrained efficient contracts can be fully characterized by the evolution of the multiplier $\lambda(\theta^t)$ (along with an initial value λ_0). This can be easily done using the first-order conditions with respect to $w^1(\theta')$, which writes, for all $\theta'\in\Theta$,

$$-\frac{\partial}{\partial v}V\left(w^1(\theta'),\theta'\right)=\frac{\lambda+\chi_1(\theta')}{1+\chi_2(\theta')}, \tag{153}$$

where $\beta\pi(\theta'|\theta)\chi_1(\theta')$ and $\beta\pi(\theta'|\theta)\chi_2(\theta')$ are the multipliers associated with the constraints (149) and (150). For each $\theta\in\Theta$, we can then define an interval $\left[\underline{\lambda}_\theta,\overline{\lambda}_\theta\right]$ by

$$\underline{\lambda}_\theta \equiv -\frac{\partial}{\partial v}V\left(\underline{v}^1(\theta),\theta\right) \text{ and } \overline{\lambda}_\theta \equiv -\frac{\partial}{\partial v}V\left(\overline{v}^1(\theta),\theta\right),$$

where $\overline{v}^1(\theta)$ is the maximum feasible expected value for agent 1, which satisfies $V\left(\overline{v}^1(\theta),\theta\right)=\underline{v}^2(\theta)$. We thus obtain the following law of motion for $\lambda(\theta^t)$:

$$\lambda(\theta^t,\theta_{t+1})=\begin{cases}\underline{\lambda}_{\theta_{t+1}}, & \text{if } \lambda(\theta^t)<\underline{\lambda}_{\theta_{t+1}},\\ \lambda(\theta^t), & \text{if } \lambda(\theta^t)\in\left[\underline{\lambda}_{\theta_{t+1}},\overline{\lambda}_{\theta_{t+1}}\right],\\ \overline{\lambda}_{\theta_{t+1}}, & \text{if } \lambda(\theta^t)>\overline{\lambda}_{\theta_{t+1}}.\end{cases} \tag{154}$$

Finally, varying the initial value λ_0 in the interval $\left[\min_{\theta\in\Theta}\{\underline{\lambda}_\theta\},\max_{\theta\in\Theta}\{\overline{\lambda}_\theta\}\right]$ traces out the Pareto frontier.

To understand intuitively this characterization, suppose for simplicity that the nonnegativity constraints on consumption (151) never bind, ie, $\psi_1 = \psi_2 = 0$. We already argued that in a full risk-pooling contract, the current transfers in every period are chosen such that the ratio of the two households' marginal utilities (152) is constant. Now consider a constrained-efficient contract, where the evolution of this ratio is given by Eq. (154). Suppose that the marginal utility ratio last period was $\lambda(\theta^t)$, and that the current state is $\theta_{t+1}=\theta'$, which defines an interval of possible marginal utility ratios $\left[\underline{\lambda}_{\theta'},\overline{\lambda}_{\theta'}\right]$. If $\lambda(\theta^t)\in\left[\underline{\lambda}_{\theta'},\overline{\lambda}_{\theta'}\right]$, then we choose $\tau(\theta^t)$ so that $\lambda(\theta^t,\theta')=\lambda(\theta^t)$. If instead $\lambda(\theta^t)<\underline{\lambda}_{\theta'}$

(respectively, if $\lambda(\theta^t) > \overline{\lambda}_{\theta'}$), household 1 (resp., household 2) would want to break the contract if the ratio of marginal utilities remained constant, as the short-term costs of making the corresponding transfer in the current period would exceed the long-term insurance benefits coming from promises of future reciprocation. That is, the constraint (149) (resp., (150)) is binding and the multiplier $\chi_1(\theta')$ (resp., $\chi_2(\theta')$) is strictly positive, implying (from (153)) that $\lambda\left(\theta^{t+1}\right) > \lambda(\theta^t)$ (resp., $\lambda\left(\theta^{t+1}\right) < \lambda(\theta^t)$). Therefore full risk-pooling, which would occur with complete markets, is not feasible in this case. We then choose $\lambda(\theta^t, \theta') = \underline{\lambda}_{\theta'}$ (resp., $\lambda(\theta^t, \theta') = \overline{\lambda}_{\theta'}$). The value $\lambda = \underline{\lambda}_{\theta'}$ (respectively, $\lambda = \overline{\lambda}_{\theta'}$) corresponds to household 1 receiving its minimum possible sustainable surplus $\underline{v}^1(\theta')$ in state θ' (resp., its maximum surplus $\bar{v}^1(\theta')$), or equivalently household 2 getting $\bar{v}^2(\theta')$ (resp., $\underline{v}^2(\theta')$). In other words, if full risk sharing is not possible, the ratio of marginal utilities must change to an endpoint (ie, by the minimum possible amount) so that one of the households is just indifferent between staying in the contract and reneging.[bb]

Ligon et al. (2002) then test the model on the data for three Indian villages, using the model to predict consumption allocations (by estimating empirically the initial ratio of marginal utilities and values for the model's parameters that provide the best fit to the data), and measuring the difference between these predictions and the actual data. They find that the dynamic limited commitment model does a substantially better job at explaining the dynamic response of consumption to income than do models of full insurance, static limited commitment, or autarky.

In models of limited commitment, the key to the amount of informal insurance that can be provided in the optimal contract depends on how costly reneging is for the households. That is, the value of autarky is the most important determinant of the extent of insurance. Recent work by Morten (2013) studies a model of risk sharing with endogenous commitment in which temporary migration is possible. The possibility of migration has the unintended consequence of improving self insurance of individuals and the value of autarky, and worsening the risk sharing in the economy. She studies the joint determination of risk sharing and migration decisions and decomposes the welfare effects of migration between changes in income and changes in the endogenous structure of insurance. Morten (2013) further structurally estimates the model on a panel from rural India and argues that the possibility of migration may significantly reduce risk sharing.

[bb] We can show that for a sufficiently high discount factor $\beta \geq \beta^* \in [0,1)$ the λ-intervals overlap and thus there is some first-best contract which is sustainable, whereas if the households are sufficiently impatient, ie, $\beta \leq \beta_* \in (0,1)$, then no nonautarkic contract exists. In the former case, irrespective of the initial value of λ_0, and hence of the initial division of the surplus, the contract converges with probability 1 to a first-best contract. Thus, if people are sufficiently patient, absence of commitment cannot justify the observed lack of diversification in individual consumption as being efficient.

There is by now a large literature studying the predictions of models with contracting frictions in the development economics context. For example, Karaivanov and Townsend (2014) is a comprehensive study comparing exogenously incomplete markets to markets which are endogenously incomplete due to contractual frictions. Their focus is on consumption, income, investment, and asset behavior of small businesses in Thailand. They conclude that the exogenously incomplete market model has the best fit for their rural sample, while the dynamic moral hazard model is more appropriate for urban households. A recent paper by Kinnan (2011) develops a test to distinguish barriers to informal insurance in Thai villages for three types of models: limited commitment, moral hazard, and hidden income, based on the theoretical prediction (see, eg, Eq. (77)) that a single lag of inverse marginal utility is sufficient to forecast current inverse marginal utility, which is satisfied by the first two models but not the latter. She concludes that hidden income is more likely to be the cause of barriers to insurance.

4.4 International Finance

In this section, we describe an application of the recursive contract models to the international finance context, based on Kehoe and Perri (2002). The benchmark model is one of limited commitment similar to that studied in the previous section, but we now analyze it using the duality theory described in Section 3.1.4. Models of limited commitment are useful to analyze questions related to sovereign debt default as they provide a framework that can explain the mechanisms by which countries are induced to participate in contracts involving transfers backed only by promises of future repayment, ie, without a legal authority enforcing them. In such models, countries are free to renege on their debts; the only threat is exclusion from future participation in the financial market.

Standard international business cycle models with either complete or exogenously incomplete markets typically deliver predictions that are at odds with the data (see Backus et al., 1992), for instance, that cross-country correlations of consumption are much higher than those for output, and that both employment and investment in different countries comove negatively. Moreover, net exports and investment are much more volatile in these models than in the data. Kehoe and Perri (2002) show that introducing endogenously incomplete markets due to limited loan enforcement frictions in an otherwise standard international business cycle model can resolve these puzzles. This feature allows the model to reproduce the data's positive cross-country comovements of factors of production, consumption, and output.

Formally, the model consists of two countries $i = 1,2$ that produce their output using domestic labor and capital inputs and face exogenous idiosyncratic Markov technology shocks $A_i(\theta^t)$. (For simplicity, in this section we ignore the subscripts "t" when there is no ambiguity.) Output in country i after a history of shocks θ^t is given by

$F\left(k_i\left(\theta^{t-1}\right), A_i(\theta^t) l_i(\theta^t)\right)$. The social planner's problem consists of choosing allocations $\left\{c_i(\theta^t), l_i(\theta^t), k_i\left(\theta^{t-1}\right)\right\}_{i,t,\theta^t}$ to maximize a weighted (with weights λ_i) sum of utilities of the representative consumers in each country:

$$\max_{\mathbf{c},\boldsymbol{l},\boldsymbol{k}} \sum_{i=1,2} \lambda_i \left\{ \sum_{t=0}^{\infty} \sum_{\theta^t \in \Theta^t} \beta^t \pi_t(\theta^t) U(c_i(\theta^t), l_i(\theta^t)) \right\} \tag{155}$$

subject to the feasibility constraint

$$\sum_{i=1,2} (c_i(\theta^t) + k_i(\theta^t)) = \sum_{i=1,2} \left[F\left(k_i\left(\theta^{t-1}\right), A_i(\theta^t) l_i(\theta^t)\right) + (1-\delta) k_i\left(\theta^{t-1}\right)\right],$$

and the enforcement constraints (similar to (147)): for all $i = 1,2$ and t,θ^t,

$$\sum_{s=t}^{\infty} \sum_{\theta^s \geq \theta^t} \beta^{s-t} \pi_s(\theta^s|\theta^t) U(c_i(\theta^s), l_i(\theta^s)) \geq \underline{V}_i\left(k_i\left(\theta^{t-1}\right), \theta^t\right), \tag{156}$$

where $\underline{V}_i\left(k_i\left(\theta^{t-1}\right), \theta^t\right)$ denotes country i's value of autarky from θ^t onward, given by

$$\begin{aligned} \underline{V}_i\left(k_i\left(\theta^{t-1}\right), \theta^t\right) = \max_{\mathbf{c},\boldsymbol{l},\boldsymbol{k}} & \sum_{s=t}^{\infty} \sum_{\theta^s \geq \theta^t} \beta^{s-t} \pi_s(\theta^s|\theta^t) U(c_i(\theta^s), l_i(\theta^s)) \\ \text{subject to} \quad & c_i(\theta^s) + k_i(\theta^s) \leq F\left(k_i\left(\theta^{s-1}\right), A_i(\theta^s) l_i(\theta^s)\right) + (1-\delta) k_i\left(\theta^{s-1}\right). \end{aligned} \tag{157}$$

The enforcement constraints are formally derived from arguments similar to those we used to obtain (95) in Section 3.2. They ensure that it is the best response for each country to stick to their equilibrium strategies.

We can rewrite this problem recursively using the Marcet and Marimon (2015) approach (see Section 3.1.4). Letting $\beta^t \pi_t(\theta^t) \mu_i(\theta^t)$ denote the multipliers on the enforcement constraints (156), a similar derivation as Eq. (89) implies that we can write the Lagrangian of the social planner's problem as

$$\begin{aligned} \sum_{t=0}^{\infty} \sum_{\theta^t \in \Theta^t} \sum_{i=1,2} \beta^t \pi_t(\theta^t) \quad & \left\{ M_i\left(\theta^{t-1}\right) U(c_i(\theta^t), l_i(\theta^t)) \right. \\ & \left. + \mu_i(\theta^t) \left[U(c_i(\theta^t), l_i(\theta^t)) - \underline{V}_i\left(k_i\left(\theta^{t-1}\right), \theta^t\right) \right] \right\} \end{aligned} \tag{158}$$

subject to the feasibility constraint, where $M_i(\theta^t)$ is a cumulative Lagrange multiplier defined recursively as

$$M_i(\theta^t) = M_i\left(\theta^{t-1}\right) + \mu_i(\theta^t), \tag{159}$$

for $t \geq 0$, with $M_i\left(\theta^{-1}\right) = \lambda_i$. Thus the cumulative multiplier $M_i(\theta^t)$ is equal to the original planning weight λ_i at time 0, plus the sum of the past multipliers on the enforcement constraints at time t and history θ^t. Using the techniques described in

Section 3.1.4 and denoting by $z(\theta^t) = \frac{M_2(\theta^t)}{M_1(\theta^t)}$ the relative weight on country 2, this problem can be written recursively and its solution is stationary in the state space that consists of the current shock, the current capital stocks, and the relative weight, ie, $x_t = \left(\theta_t, k_1\left(\theta^{t-1}\right), k_2\left(\theta^{t-1}\right), z\left(\theta^{t-1}\right)\right)$.

It is instructive to compare this objective (158) with the unconstrained objective (155). The enforcement constraints introduce three key differences. First, starting at the beginning of the period, the cumulative Lagrange multiplier $M_i\left(\theta^{t-1}\right)$ shifts the (relative) weights of each agent. Second, the current Lagrange multiplier $\mu_i(\theta^t)$ on the sustainability constraint further changes the weight on current consumption (as well as on future consumption by affecting the future cumulative multiplier $M_i(\theta^t)$). These two forces translate in the first-order conditions into a distortion of the relative marginal utilities of consumption (letting $U_{ic}(\theta^t)$ denote the marginal utility of consumption in country i in history θ^t):

$$\frac{U_{1c}(\theta^t)}{U_{2c}(\theta^t)} = \frac{M_2\left(\theta^{t-1}\right) + \mu_2\left(\theta^t\right)}{M_1\left(\theta^{t-1}\right) + \mu_1\left(\theta^t\right)}. \tag{160}$$

Third, accumulating more capital $k_i\left(\theta^{t-1}\right)$ tightens the enforcement constraint by increasing the value of autarky. As a result, the Euler equation (and capital accumulation) is distorted as follows (letting $F_{ik}(\theta^t)$ denote the marginal product of capital in country i in history θ^t):

$$\begin{aligned} U_{ic}(\theta^t) = \beta \sum_{\theta_{t+1}} \pi(\theta_{t+1}|\theta_t) \\ \times \left[\frac{M_i\left(\theta^{t+1}\right)}{M_i(\theta^t)} U_{ic}\left(\theta^{t+1}\right)\left(F_{ik}\left(\theta^{t+1}\right) + 1 - \delta\right) - \frac{\mu_i\left(\theta^{t+1}\right)}{M_i(\theta^t)} \underline{V}_{ik}\left(\theta^{t+1}\right)\right]. \end{aligned} \tag{161}$$

The last first-order condition writes $\frac{U_{il}(\theta^t)}{U_{ic}(\theta^t)} = F_{il}(\theta^t)$ (letting $U_{il}(\theta^t)$ and $F_{il}(\theta^t)$ denote the marginal disutility and marginal product of labor in country i in history θ^t): there is no distortion in the consumption–labor decision, since this margin does not affect the enforcement constraint.[bc] These first-order conditions along with the transition law for $z_2(\theta^t)$ can be straightforwardly rewritten as functions of $z_2\left(\theta^{t-1}\right)$ and the normalized multipliers $\tilde{\mu}_i(\theta^t) \equiv \frac{\mu_i(\theta^t)}{M_i\left(\theta^{t-1}\right)}$. The solution to this problem can then be characterized

[bc] Kehoe and Perri (2004) show how to decentralize the constrained efficient allocation as a competitive equilibrium using a tax on capital income to replicate the wedge in the Euler equation (161) generated by the enforcement constraint.

by allocations of the form $(c_i(x_t), l_i(x_t), k_i(x_t))$, where the state vector is $x_t = \left(\theta_t, k_1\left(\theta^{t-1}\right), k_2\left(\theta^{t-1}\right), z\left(\theta^{t-1}\right)\right)$. These policy functions satisfy the first-order conditions above, the feasibility and enforcement constraints, and the complementary slackness conditions on the multipliers.

The model has the following implications. Suppose that the home country (say, country $i = 1$) is hit in history (t, θ^t) with a positive and persistent productivity shock $A_1(\theta^t) > 0$. Eq. (157) shows that such a shock increases the home country's value of autarky, and thus tightens its enforcement constraint (156). This may lead the enforcement constraint to bind, which translates into a positive multiplier $\mu_1(\theta^t)$ in the first-order condition (160). This in turn implies that the planner increases the relative weight to the home country in its objective and allocates it higher consumption $c_1(\theta^t)$ (ie, lower marginal utility $U_{1c}(\theta^t)$) to prevent it from defaulting. Moreover, this increase in consumption is persistent, because the productivity shock is persistent and the positive multiplier $\mu_1(\theta^t)$ raises the cumulative multiplier $M_1(\theta^s)$ of the home country (defined in (159)) in all future periods $s \geq t$. In contrast, consumption in the foreign country does not vary much, as risk sharing in this economy is limited. Finally, the planner optimally restricts the investment flow into country 1 in order to reduce the home country's future value of autarky in Eq. (161) and relax the enforcement constraint. It also increases labor effort and investment in the foreign country to raise country 1's value of participating into the contract, leading to positive cross-country correlations of investment and employment and to a trade surplus (positive net exports) in the home country.

Now compare these effects with those that would occur in an economy without enforcement frictions, ie, with complete markets. In response to a positive productivity shock in the home country, and hence a higher productivity of capital and labor, the planner optimally increases the domestic labor effort and the capital stock, both by saving more and increasing investment flowing from abroad. In contrast, foreign labor effort and investment decrease. Moreover, because of risk sharing, the domestic economy shares its consumption gains, leading to an increase in the consumption of the foreign country. The responses are qualitatively similar but muted in a model where markets are exogenously incomplete (only bonds are allowed). In such models, therefore, output is less correlated across countries than is consumption, the cross-country correlations of investment and employment are negative, and a positive productivity shock leads to a trade deficit in the home country (due to the net inflow of investment).

Kehoe and Perri (2002) calibrate the economy and analyze numerically these implications of the model with endogenously incomplete markets. They find that it matches the data's positive cross-country comovements of factors of production (employment, investment) and the cross-country comovements of consumption and output. This resolves several of the puzzles arising in standard (complete or exogenously incomplete market-)models of international finance described in the first paragraph of this section.

There is a large literature that analyzes questions of international debt and sovereign default using models of (one- or two-sided) limited commitment. The seminal paper is Eaton and Gersovitz (1981), and this literature has been comprehensively reviewed by Aguiar and Amador (2013). In particular, Aguiar et al. (2009) analyze the behavior of sovereign debt and foreign direct investment in a small open economy (rather than in a two-country general equilibrium environment as analyzed in the previous paragraphs) where the government lacks commitment (leading to potential default and expropriation of capital) and is more impatient than the market. While the standard one-sided limited commitment model (see Thomas and Worrall, 1988) predicts that the government will eventually accumulate enough assets to overcome its commitment problem,[bd] the additional assumption of a higher degree of impatience (and hence, the combination of front loading due to impatience and back loading due to limited commitment) leads to cycles in both sovereign debt and foreign direct investment, as well as a "debt overhang" effect whereby investment is distorted by more in recessions than in booms.

5. CONCLUSION

The theory of recursive contracts underpins a variety of applications in a range of fields, from public finance to development economics to corporate finance, international finance, and political economy. A unifying feature of these applications is that they feature frictions such as unobservability of shocks or actions or nonenforceability of contracts that endogenously limit the amount of risk sharing and insurance that can be achieved. This chapter provides a self-contained treatment of the fundamental techniques and the more advanced topics of recursive contracts. We also survey a number of applications through the lens of this unified theoretical treatment that illustrate the versatility of the theoretical apparatus.

[bd] One-sided limited commitment models generally imply that the optimal contract features a form *back loading*: the profile of consumption is shifted toward the future. The intuition is as follows. Additional consumption in a particular period helps ensure the agent's participation in the contract. Moreover, it also helps satisfy the enforcement constraints in all previous periods as well, since the left-hand side of the enforcement constraint (eg, (156)) is forward-looking. At the margin, therefore, consumption in the future is preferable as it relaxes all the preceding participation constraints. As a result the relevant Euler equation includes the cumulative sums of Lagrange multipliers that take into account all of the binding constraints in the previous periods. When the government and the market have the same degree of impatience, the economy will eventually achieve perfect risk sharing with constant consumption, so that a country has an incentive to save to grow out of the enforcement constraints if it is patient enough. Ray (2002) shows that the backloading result and eventual reaching of the unconstrained allocations apply in very general settings.

ACKNOWLEDGMENTS

M.G. and A.T. thank the NSF for financial support.

REFERENCES

Abraham, A., Pavoni, N., 2008. Efficient allocations with moral hazard and hidden borrowing and lending: a recursive formulation. Rev. Econ. Dyn. 11 (4), 781–803.

Abreu, D., 1988. On the theory of infinitely repeated games with discounting. Econometrica 56, 383–396.

Abreu, D., Pearce, D., Stacchetti, E., 1990. Toward a theory of discounted repeated games with imperfect monitoring. Econometrica 58, 1041–1063.

Acemoglu, D., Golosov, M., Tsyvinski, A., 2008. Political economy of mechanisms. Econometrica 76 (3), 619–641.

Acemoglu, D., Golosov, M., Tsyvinski, A., 2011. Power fluctuations and political economy. J. Econ. Theory 146 (3), 1009–1041.

Aguiar, M., Amador, M., 2013. Sovereign debt. Handbook of International Economics, vol. 4.

Aguiar, M., Amador, M., Gopinath, G., 2009. Investment cycles and sovereign debt overhang. Rev. Econ. Stud. 76 (1), 1–31.

Aiyagari, S.R., 1994. Uninsured idiosyncratic risk and aggregate saving. Q. J. Econ. 109, 659–684.

Aiyagari, S.R., Marcet, A., Sargent, T.J., Seppälä, J., 2002. Optimal taxation without state-contingent debt. J. Polit. Econ. 110 (6), 1220–1254.

Albanesi, S., 2011. Optimal taxation of entrepreneurial capital with private information. Working Paper.

Albanesi, S., Sleet, C., 2006. Dynamic optimal taxation with private information. Rev. Econ. Stud. 73 (1), 1–30.

Albuquerque, R., Hopenhayn, H., 2004. Optimal lending contracts and firm dynamics. Rev. Econ. Stud. 71 (2), 285–315.

Ales, L., Maziero, P., 2009. Non-exclusive dynamic contracts, competition, and the limits of insurance. Working Paper.

Allen, F., 1985. Repeated principal-agent relationships with lending and borrowing. Econ. Lett. 17 (1–2), 27–31.

Alvarez, F., Jermann, U.J., 2000. Efficiency, equilibrium, and asset pricing with risk of default. Econometrica 68, 775–797.

Atkeson, A., Lucas, R.E., 1992. On efficient distribution with private information. Rev. Econ. Stud. 59 (3), 427–453.

Atkeson, A., Lucas, R.E., 1995. Efficiency and equality in a simple model of efficient unemployment insurance. J. Econ. Theory 66 (1), 64–88.

Backus, D.K., Kehoe, P.J., Kydland, F.E., 1992. International real business cycles. J. Polit. Econ. 100, 745–775.

Badel, A., Huggett, M., 2014. Taxing top earners: a human capital perspective. Federal Reserve Bank of St. Louis. Working Paper.

Barro, R.J., 1979. On the determination of the public debt. J. Polit. Econ. 87, 940–971.

Battaglini, M., Lamba, R., 2015. Optimal dynamic contracting: the first-order approach and beyond. Working Paper.

Benveniste, L.M., Scheinkman, J.A., 1979. On the differentiability of the value function in dynamic models of economics. Econometrica 47 (3), 727–732.

Bertsekas, D.P., Nedi, A., Ozdaglar, A.E., 2003. Convex Analysis and Optimization. Athena Scientific, Boston.

Bester, H., Strausz, R., 2001. Contracting with imperfect commitment and the revelation principle: the single agent case. Econometrica 69 (4), 1077–1098.

Biais, B., Mariotti, T., Plantin, G., Rochet, J.C., 2007. Dynamic security design: convergence to continuous time and asset pricing implications. Rev. Econ. Stud. 74 (2), 345–390.

Biais, B., Mariotti, T., Rochet, J.C., Villeneuve, S., 2010. Large risks, limited liability, and dynamic moral hazard. Econometrica 80, 73–118.

Billingsley, P., 2008. Probability and Measure. John Wiley & Sons.

Bismut, J.M., 1973. Conjugate convex functions in optimal stochastic control. J. Math. Anal. Appl. 44 (2), 384–404.

Bismut, J.M., 1978. An introductory approach to duality in optimal stochastic control. SIAM Rev. 20 (1), 62–78.

Caves, R.E., 1998. Industrial organization and new findings on the turnover and mobility of firms. J. Econ. Lit. 36, 1947–1982.

CBO, 2007. Historical effective federal tax rates, 1979 to 2005. Congressional Budget Office.

Chamberlain, G., Wilson, C.A., 2000. Optimal intertemporal consumption under uncertainty. Rev. Econ. Dyn. 3 (3), 365–395.

Chari, V.V., Kehoe, P.J., 1990. Sustainable plans. J. Polit. Econ. 98 (4), 783–802.

Chari, V.V., Kehoe, P.J., 1993. Sustainable plans and debt. J. Econ. Theory 61 (2), 230–261.

Clementi, G.L., Hopenhayn, H.A., 2006. A theory of financing constraints and firm dynamics. Q. J. Econ. 121 (1), 229–265.

Cole, H., Kocherlakota, N., 2001. Efficient allocations with hidden income and hidden storage. Rev. Econ. Stud. 68 (3), 523–542.

Cole, H., Kubler, F., 2012. Recursive contracts, lotteries and weakly concave Pareto sets. Rev. Econ. Dyn. 15 (4), 479–500.

Cvitanić, J., Zhang, J., 2013. Contract Theory in Continuous-Time Models. Springer Science & Business Media.

DeMarzo, P.M., Fishman, M.J., 2007a. Agency and optimal investment dynamics. Rev. Financ. Stud. 20 (1), 151–188.

DeMarzo, P.M., Fishman, M.J., 2007b. Optimal long-term financial contracting. Rev. Financ. Stud. 20 (6), 2079–2128.

DeMarzo, P.M., Sannikov, Y., 2006. Optimal security design and dynamic capital structure in a continuous-time agency model. J. Finance 61 (6), 2681–2724.

DeMarzo, P.M., Fishman, M.J., He, Z., Wang, N., 2012. Dynamic agency and the q theory of investment. J. Finance 67 (6), 2295–2340.

Diamond, P., Mirrlees, J., 1978. A model of social insurance with variable retirement. J. Public Econ. 10 (3), 295–336.

Diamond, P., Mirrlees, J., 1986. Payroll-tax financed social insurance with variable retirement. Scand. J. Econ. 88 (1), 25–50.

Diamond, P.A., Helms, L.J., Mirrlees, J.A., 1980. Optimal taxation in a stochastic economy: a Cobb-Douglas example. J. Public Econ. 14 (1), 1–29.

Dovis, A., Golosov, M., Shourideh, A., 2015. Political economy of sovereign debt: cycles of debt crisis and inequality overhang. Working Paper.

Eaton, J., Gersovitz, M., 1981. Debt with potential repudiation: theoretical and empirical analysis. Rev. Econ. Stud. 48, 289–309.

Ekeland, I., Scheinkman, J.A., 1986. Transversality conditions for some infinite horizon discrete time optimization problems. Math. Oper. Res. 11 (2), 216–229.

Espino, E., Kozlowski, J., Sanchez, J.M., 2013. Too big to cheat: efficiency and investment in partnerships. FRB of St. Louis Working Paper No. 2013-001C.

Farhi, E., Golosov, M., Tsyvinski, A., 2009. A theory of liquidity and regulation of financial intermediation. Rev. Econ. Stud. 76 (3), 973–992.

Farhi, E., Sleet, C., Werning, I., Yeltekin, S., 2012. Non-linear capital taxation without commitment. Rev. Econ. Stud. 79 (4), 1469–1493.

Farhi, E., Werning, I., 2007. Inequality and social discounting. J. Polit. Econ. 115 (3), 365–402.

Farhi, E., Werning, I., 2012. Capital taxation: quantitative explorations of the inverse Euler equation. J. Polit. Econ. 120 (3), 398–445.

Farhi, E., Werning, I., 2013. Insurance and taxation over the life cycle. Rev. Econ. Stud. 80 (2), 596–635.

Fazzari, S.M., Hubbard, R.G., Petersen, B.C., Blinder, A.S., Poterba, J.M., 1988. Financing constraints and corporate investment. Brook. Pap. Econ. Act. 1, 141–206.

Fernandes, A., Phelan, C., 2000. A recursive formulation for repeated agency with history dependence. J. Econ. Theory 91 (2), 223–247.

Freixas, X., Guesnerie, R., Tirole, J., 1985. Planning under incomplete information and the ratchet effect. Rev. Econ. Stud. 52 (2), 173–191.

Friedman, M., 1957. A theory of the consumption function. National Bureau of Economic Research, Inc.

Geweke, J., Keane, M., 2000. An empirical analysis of earnings dynamics among men in the PSID: 1968–1989. J. Econ. 96 (2), 293–356.

Golosov, M., Iovino, L., 2014. Social insurance, information revelation, and lack of commitment. NBER Working Paper No. w20633.

Golosov, M., Kocherlakota, N., Tsyvinski, A., 2003. Optimal indirect and capital taxation. Rev. Econ. Stud. 70 (3), 569–587.

Golosov, M., Troshkin, M., Tsyvinski, A., 2016. Redistribution and social insurance. Am. Econ. Rev. 106, 359–386.

Golosov, M., Tsyvinski, A., 2006. Designing optimal disability insurance: a case for asset testing. J. Polit. Econ. 114 (2), 257–279.

Golosov, M., Tsyvinski, A., 2007. Optimal taxation with endogenous insurance markets. Q. J. Econ. 122 (2), 487–534.

Golosov, M., Tsyvinski, A., Werning, I., 2006. New dynamic public finance: a user's guide. NBER Macroecon. Annu. 21, 317–363.

Green, E.J., 1987. Lending and the smoothing of uninsurable income. In: Prescott, E.C., Wallace, N. (Eds.), Contractual Arrangements for Intertemporal Trade. University of Minnesota Press, Minneapolis, Minnesota.

Guvenen, F., Ozkan, S., Song, J., 2014. The nature of countercyclical income risk. J. Polit. Econ. 122 (3), 621–660.

Guvenen, F., Song, J., Ozkan, S., Karahan, F., 2015. What do data on millions of US workers reveal about life-cycle earnings risk? NBER Working Paper No. w20913.

Hall, R.E., 1978. Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. J. Polit. Econ. 86 (6), 971–987.

He, Z., 2009. Optimal executive compensation when firm size follows geometric Brownian motion. Rev. Financ. Stud. 22 (2), 859–892.

Hosseini, R., Jones, L.E., Shourideh, A., 2013. Optimal contracting with dynastic altruism: family size and per capita consumption. J. Econ. Theory 148 (5), 1806–1840.

Hurwicz, L., 1960. Optimality and informational efficiency in resource allocation processes. In: Mathematical Methods in the Social Sciences, 1959: Proceedings of the First Stanford Symposium. Stanford University Press, p. 27.

Hurwicz, L., 1972. On informationally decentralized systems. In: Decision and Organization: A Volume in Honor of Jacob Marschak. North-Holland.

Kapička, M., 2013. Efficient allocations in dynamic private information economies with persistent shocks: a first order approach. Rev. Econ. Stud. 80 (3), 1027–1054.

Karaivanov, A., Townsend, R.M., 2014. Dynamic financial constraints: distinguishing mechanism design from exogenously incomplete regimes. Econometrica 82 (3), 887–959.

Karatzas, I., Shreve, S., 2012. Brownian Motion and Stochastic Calculus, vol. 113. Springer Science & Business Media.

Kehoe, P.J., Perri, F., 2002. International business cycles with endogenous incomplete markets. Econometrica 70 (3), 907–928.

Kehoe, P.J., Perri, F., 2004. Competitive equilibria with limited enforcement. J. Econ. Theory 119 (1), 184–206.

Kehoe, T.J., Levine, D.K., 1993. Debt-constrained asset markets. Rev. Econ. Stud. 60, 865–888.

Kinnan, C., 2011. Distinguishing barriers to insurance in Thai villages. Working Paper.

Kocherlakota, N.R., 1996. Implications of efficient risk sharing without commitment. Rev. Econ. Stud. 63 (4), 595–609.

Kocherlakota, N.R., 2010. The New Dynamic Public Finance. Princeton University Press, Princeton, NJ.

Laffont, J.J., Tirole, J., 1988. The dynamics of incentive contracts. Econometrica 56 (5), 1153–1175.

Le Van, C., Saglam, H.C., 2004. Optimal growth models and the Lagrange multiplier. J. Math. Econ. 40 (3), 393–410.

Ligon, E., Thomas, J.P., Worrall, T., 2000. Mutual insurance, individual savings, and limited commitment. Rev. Econ. Dyn. 3 (2), 216–246.

Ligon, E., Thomas, J.P., Worrall, T., 2002. Informal insurance arrangements with limited commitment: theory and evidence from village economies. Rev. Econ. Stud. 69 (1), 209–244.

Lindbeck, A., Weibull, J.W., 1987. Balanced-budget redistribution as the outcome of political competition. Public Choice 52 (3), 273–297.

Ljungqvist, L., Sargent, T., 2012. Recursive Macroeconomic Theory. MIT Press, Cambridge, Massachusetts.

Lockwood, B.B., Nathanson, C.G., Weyl, E.G., 2014. Taxation and the allocation of talent. Working Paper.

Luenberger, D., 1969. Optimization by Vector Space Methods. Wiley-Interscience.

Maag, E., Steuerle, C.E., Chakravarti, R., Quakenbush, C., 2012. How marginal tax rates affect families at various levels of poverty. Natl. Tax J. 65 (4), 759–782.

Marcet, A., Marimon, R., 2015. Recursive contracts. European University Institute, Mimeo.

Mas-Colell, A., Whinston, M., Green, J., 1995. Microeconomic Theory. Oxford University Press, New York.

Messner, M., Pavoni, N., 2016. On the recursive saddle point method. Dyn. Games Appl. 6, 161–173.

Messner, M., Pavoni, N., Sleet, C., 2012. Recursive methods for incentive problems. Rev. Econ. Dyn. 15 (4), 501–525.

Messner, M., Pavoni, N., Sleet, C., 2014. The dual approach to recursive optimization: theory and examples. In: 2014 Meeting Papers, 1267.

Miao, J., Zhang, Y., 2015. A duality approach to continuous-time contracting problems with limited commitment. J. Econ. Theory 159 (Part B), 929–988.

Milgrom, P., Segal, I., 2002. Envelope theorems for arbitrary choice sets. Econometrica 70 (2), 583–601.

Morten, M., 2013. Temporary migration and endogenous risk sharing in village India. Working Paper.

Myerson, R.B., 1981. Optimal auction design. Math. Oper. Res. 6 (1), 58–73.

Myerson, R.B., 1982. Optimal coordination mechanisms in generalized principal-agent problems. J. Math. Econ. 10 (1), 67–81.

Myerson, R.B., 1986. Multistage games with communication. Econometrica 54, 323–358.

Øksendal, B., 2003. Stochastic Differential Equations: An Introduction with Applications. Springer Berlin Heidelberg, Berlin, Heidelberg.

Øksendal, B.K., Sulem, A., 2007. Applied Stochastic Control of Jump Diffusions. Springer Berlin Heidelberg, Berlin, Heidelberg.

Pavan, A., Segal, I., Toikka, J., 2014. Dynamic mechanism design: a myersonian approach. Econometrica 82 (2), 601–653. ISSN 1468-0262.

Phelan, C., 1995. Repeated moral hazard and one-sided commitment. J. Econ. Theory 66 (2), 488–506.

Phelan, C., Townsend, R.M., 1991. Computing multi-period, information-constrained optima. Rev. Econ. Stud. 58 (5), 853–881.

Ray, D., 2002. The time structure of self-enforcing agreements. Econometrica 70 (2), 547–582.

Revuz, D., Yor, M., 1999. Continuous Martingales and Brownian Motion. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 293. Springer-Verlag, Berlin.

Rogerson, W.P., 1985. Repeated moral hazard. Econometrica 53 (1), 69–76.

Rudin, W., 1976. Principles of Mathematical Analysis, third ed. McGraw-Hill Book Co., New York.

Rustichini, A., 1998. Lagrange multipliers in incentive-constrained problems. J. Math. Econ. 29 (4), 365–380.

Sannikov, Y., 2008. A continuous-time version of the principal-agent problem. Rev. Econ. Stud. 75 (3), 957–984.

Sannikov, Y., 2014. Moral hazard and long-run incentives. Princeton University, Mimeo.

Scheuer, F., Wolitzky, A., 2014. Capital taxation under political constraints. Working Paper Series. National Bureau of Economic Research.

Shourideh, A., 2010. Optimal taxation of capital income: a Mirrleesian approach to capital accumulation. Working Paper.

Sleet, C., Yeltekin, S., 2008. Politically credible social insurance. J. Monet. Econ. 55 (1), 129–151.

Spear, S., Srivastava, S., 1987. On repeated moral hazard with discounting. Rev. Econ. Stud. 54 (4), 599–617.

Stantcheva, S., 2014. Optimal taxation and human capital policies over the life cycle. Working Paper.

Stokey, N.L., Lucas, R.E., Prescott, E.C., 1989. Recursive Methods in Economic Dynamics. Harvard University Press, Cambridge, MA.

Storesletten, K., Telmer, C.I., Yaron, A., 2004. Consumption and risk sharing over the life cycle. J. Monet. Econ. 51, 609–633.

Sun, Y., 2006. The exact law of large numbers via Fubini extension and characterization of insurable risks. J. Econ. Theory 126 (1), 31–69.

Thomas, J., Worrall, T., 1988. Self-enforcing wage contracts. Rev. Econ. Stud. 55 (4), 541–554.

Thomas, J., Worrall, T., 1990. Income fluctuation and asymmetric information: an example of a repeated principal-agent problem. J. Econ. Theory 51 (2), 367–390.

Townsend, R.M., 1994. Risk and insurance in village India. Econometrica 62, 539–591.

Uhlig, H., 1996. A law of large numbers for large economies. Econ. Theory 8 (1), 41–50.

Vickrey, W., 1947. Agenda for Progressive Taxation. The Ronald Press Company, New York.

Werning, I., 2002. Optimal unemployment insurance with unobservable savings. MIT Working Paper.

Werning, I., 2009. Nonlinear capital taxation. MIT Working Paper.

Williams, N., 2009. On dynamic principal-agent problems in continuous time. Working Paper. Citeseer.

Williams, N., 2011. Persistent private information. Econometrica 79 (4), 1233–1275.

Yong, J., Zhou, X.Y., 1999. Stochastic Controls: Hamiltonian Systems and HJB Equations, vol. 43. Springer Science & Business Media.

Zhang, Y., 2009. Dynamic contracting with persistent shocks. J. Econ. Theory 144 (2), 635–675.

CHAPTER 11

Macroeconomics and Household Heterogeneity

D. Krueger[*,†,‡,§,¶], **K. Mitman**[†,||], **F. Perri**[†,**]

[*] University of Pennsylvania, Philadelphia, PA, United States
[†] CEPR, London, United Kingdom
[‡] CFS, Goethe University Frankfurt, Frankfurt, Germany
[§] NBER, Cambridge, MA, United States
[¶] Netspar, Tilburg, The Netherlands
[||] IIES, Stockholm University, Stockholm, Sweden
[**] Federal Reserve Bank of Minneapolis, Minneapolis, MN, United States

Contents

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.04.003

Abstract

The goal of this chapter is to study how, and by how much, household income, wealth, and preference heterogeneity amplify and propagate a macroeconomic shock. We focus on the US Great Recession of 2007–09 and proceed in two steps. First, using data from the Panel Study of Income Dynamics, we document the patterns of household income, consumption, and wealth inequality before and during the Great Recession. We then investigate how households in different segments of the wealth distribution were affected by income declines, and how they changed their expenditures differentially during the aggregate downturn. Motivated by this evidence, we study several variants of a standard heterogeneous household model with aggregate shocks and an endogenous cross-sectional wealth distribution. Our key finding is that wealth inequality can significantly amplify the impact of an aggregate shock, and it does so if the distribution features a sufficiently large fraction of households with very little net worth that sharply *increase* their saving (ie, they are not hand-to mouth) as the recession hits. We document that both these features are observed in the PSID. We also investigate the role that social insurance policies, such as unemployment insurance, play in shaping the cross-sectional income and wealth distribution, and through it, the dynamics of business cycles.

Keywords

Recessions, Wealth inequality, Social insurance

JEL Classification Codes:

E21, E32, J65

1. INTRODUCTION

How important is household heterogeneity for the amplification and propagation of macroeconomic shocks? The objective of this chapter is to give a quantitative answer to a narrower version of this broad question.[a] Specifically, we narrow the focus of this question along two dimensions. First, we mainly focus on a specific macroeconomic event, namely the US Great Recession of 2007–09.[b] Second, we focus on specific dimensions of household heterogeneity, namely that in earnings, wealth, and household preferences, and their associated correlations with, and consequences for, the cross-sectional inequality in disposable income and consumption expenditures.[c]

The Great Recession was the largest negative macroeconomic downturn the United States has experienced since World War II. The initial decline in economic activity was deep and had an impact on all macroeconomic aggregates—notably private aggregate consumption and employment—and the recovery has been slow. Is the cross-sectional distribution of wealth an important determinant of the dynamics of the initial downturn and the ensuing recovery? That is, does household heterogeneity matter in terms of aggregate economic activity (as measured by output and labor input), its composition

[a] In this chapter we focus on household heterogeneity. A sizeable literature has investigated similar questions in models with firm heterogeneity. Representative contributions from this literature include Khan and Thomas (2008) and Bachmallnn et al. (2013). We abstract from firm heterogeneity in this chapter, but note that the methodological challenges in computing these classes of models are very similar to the ones encountered here.

[b] By focusing on a business cycle event, and macroeconomic fluctuations more generally, we also abstract from the interaction between income or wealth inequality and aggregate income growth rates in the long run. See Kuznets (1955), Benabou (2002), or Piketty (2014) for important contributions to this large literature.

[c] Excellent earlier surveys of different aspects of the literature on macroeconomics with microeconomic heterogeneity are contained in Deaton (1992), Attanasio (1999), Krusell and Smith (2006), Heathcote et al. (2009), Attanasio and Weber (2010), Guvenen (2011) as well as Quadrini and Rios-Rull (2015).

between consumption and investment, and, eventually, the cross-sectional distribution of consumption and welfare?

To address these questions *empirically*, we make use of recent waves of the Panel Study of Income Dynamics (PSID), which provides household-level panel data on earnings, income, consumption expenditures, and wealth for the United States. To answer these questions *theoretically and quantitatively*, we then study various versions of the canonical real business cycle model with aggregate technology shocks and ex-ante household heterogeneity in preferences and ex-post household income heterogeneity induced by the realization of uninsurable idiosyncratic labor earnings shocks, as in Krusell and Smith (1998). In the model, a recession is associated with lower aggregate wages and higher unemployment (ie, a larger share of households with low labor income). The main empirical and model-based focus of the chapter is on the dynamics of macroeconomic variables—specifically, aggregate consumption, investment, and output—in response to such a business cycle shock. Specifically, we investigate the conditions under which the degree of wealth inequality plays a quantitatively important role for shaping this response. We also study how a stylized unemployment insurance program shapes the cross-sectional distribution of wealth and welfare, and how it affects the recovery of the aggregate economy after a Great Recession-like event.

We proceed in four steps: First, we make use of the PSID earnings, income, consumption and wealth data to document three sets of facts related to cross-sectional inequality. We summarize the key features of the joint distribution of income, wealth, and consumption prior to the Great Recession (ie, for the year 2006). Next, we show how this joint distribution changed during the recession—over the 2006–10 period—exploiting the panel dimension of the data to investigate how individual households fared and adjusted their consumption-savings behavior. The purpose of this empirical analysis is two fold. First, we believe the facts are interesting in their own right, as they characterize the distributional consequences of the Great Recession. Second, the facts serve as important moments for the evaluation of the different versions of the quantitative heterogeneous household model we study next.

In the second step, then, we construct, calibrate, and compute various versions of the canonical Krusell–Smith (1998) model and study its cross-sectional and dynamic properties. We first revisit the well-known finding that idiosyncratic unemployment risk and incomplete financial markets alone are insufficient to generate a sufficiently dispersed model-based cross-sectional wealth distribution. The problem is two fold: in the model, the very wealthy are not nearly wealthy enough, and the poor hold far too much wealth relative to the data. We argue that it is the discrepancy at bottom of the distribution that implies that the model generates an aggregate consumption response to a negative technology shock that is essentially identical to the response in a representative agent model.

We then study extensions of the model in which preference heterogeneity, idiosyncratic labor productivity risk conditional on employment, and a stylized life-cycle structure interact with the presence of unemployment insurance and social security to deliver a wealth distribution that is consistent with the data. In these economies, the decline in aggregate consumption is substantially larger than in the representative agent economy, by approximately 0.5 percentage points. This finding is primarily due to these economies now being populated by more wealth-poor households whose consumption responds strongly to the aggregate shock, both for those households that experience a transition from employment to unemployment, but also for households that have not lost their job, but understand they are facing a potentially long-lasting recession with elevated unemployment *risk*. We also stress that data and theory show that these wealth poor households do not behave as hand-to-mouth consumers, but are the group that reduces their expenditure rates strongly as their recession hits. This behavior implies that our benchmark model has quantitatively very different implications relative to a model where a large fraction of households is exogenously assumed to be hand-to-mouth consumers.

The more severe consumption declines in economies with larger wealth inequality imply a smaller collapse in investment, and thus a faster recovery from the recession, although this last effect is quantitatively small.

In light of the previous finding that larger wealth inequality—specifically, the importance of a large fraction of wealth-poor households—is an important contributor to an aggregate consumption collapse in the Great Recession, in the third step we determine whether public unemployment insurance is important for the dynamics of the economy in response to an aggregate shock. The answer to this question depends crucially on whether the distribution of household wealth has had a chance to respond to changes in the policy. In the short run, an unexpected cut or expiration of unemployment insurance benefits induces a significantly larger negative consumption response. These dynamics are explained by forward-looking households responding to lower public insurance by increasing their precautionary savings. The increased investment generates a medium-run boost to output, at the cost of a slow recovery of consumption.

In the long run, the new ergodic distribution of wealth features fewer people with zero or few assets. The consumption dynamics in response to a negative technology shock under this rightward shift in the wealth distribution are less severe than in they are in response to an unexpected shock, but still larger than in the economy with high unemployment insurance. Thus, for a *given wealth distribution* a cut in social insurance will result in a larger aggregate consumption drop. However, since social insurance policies themselves shape the ergodic distribution of wealth, and especially influence the share of households with zero or close to zero net worth, the aggregate consumption response across different economies is partially offset by these distributional shifts.

In the models considered thus far, the wealth distribution has had a potentially large effect on the *division* of aggregate output between consumption and investment, but not on output itself. In the final step, we therefore study an economy with a New Keynesian flavor—we introduce an aggregate demand externality that makes output partially demand-determined and generates an endogenous feedback effect from private consumption to total factor productivity, and thus output. In this model, social insurance policies might not only be beneficial in providing public insurance, but can also serve a potentially positive role for stabilizing aggregate output. We find that the output decline with an unemployment insurance benefit replacement rate of 50% to a Great Recession-like shock is 1 percentage point smaller on impact than in an economy with a replacement rate of 10%.

This work is part of a broader research agenda (and aims to partially synthesize it) that seeks to explore the importance of micro heterogeneity in general, and household income and wealth heterogeneity in particular, for classic macroeconomic questions (such as the impact of a particular aggregate shock) that have traditionally been answered within the representative agent paradigm (ie, goes from micro to macro). It also builds upon, and contributes to, the related but distinct literature that studies the distributional consequences of macroeconomic shocks (ie, goes from macro to micro).

The chapter is organized as follows. Section 2 documents key dimensions of heterogeneity among US households, prior to and during the Great Recession. Sections 3 and 4 present our benchmark real business cycle model with household heterogeneity and discuss how we calibrate it. Section 5 studies to what extent the benchmark model is consistent with the cross-sectional facts presented in Section 2, and Section 6 studies how the aggregate consumption response to a large shock depends on the cross-sectional wealth distribution. In Section 7 we augment the model with demand externalities in order to investigate the importance of cross-sectional wealth heterogeneity for the dynamics of aggregate output. Section 8 concludes, and the appendix contains details about the construction of the empirical facts, about the theory, and the computational algorithm used.

2. THE GREAT RECESSION: A HETEROGENEOUS HOUSEHOLD PERSPECTIVE

In this section, we present the basic facts about the cross-sectional distribution of earnings, income, consumption, and wealth before and during the Great Recession. The main data set we employ is the Panel Survey of Income Dynamics (PSID) for the years 2004, 2006, 2008, and 2010. This data set has two key advantages for the purpose of this study. First, it contains information about household earnings, income, a broad and comprehensive measure of consumption expenditures, and net worth for a sample of households representative of the US population. Second, it has a panel dimension so we can, in the same data set, both measure the key dimensions of cross-sectional household heterogeneity as well as investigate how different groups in the income and wealth distribution changed their consumption expenditure patterns during the Great Recession.[d]

The purpose of this empirical section is to provide simple and direct evidence for the importance of household heterogeneity for macroeconomic questions. It complements the large empirical literature documenting inequality trends in income, consumption and wealth in the United Sates and around the world.[e] If, as we will document, there are significant differences in behavior (for example, along the consumption and savings margin) across different groups of the earnings and wealth distribution during the Great Recession, then keeping track of the cross-sectional earnings and wealth distribution and understanding their dynamics is likely important for analyzing the unfolding of the Great Recession from a macroeconomic and distributional perspective.

2.1 Aggregates

We start our analysis by comparing the evolution of basic US macroeconomic aggregates from the National Income and Product Accounts (NIPA) with the aggregates for the same variables obtained from the PSID. In Fig. 1, we compare trends in aggregate per capita disposable income (panel A) and per capita consumption expenditures (panel B) from the Bureau of Economic Analysis (BEA) with the corresponding series obtained by aggregating household level in the PSID, for the years 2004 through 2010, the last available data point for the PSID.[f]

The main conclusion we draw from Fig. 1 is that both the NIPA and the PSID paint the same qualitative picture of the US macroeconomy over the period 2004–10. Both disposable income and consumption expenditures experience a slowdown, which is somewhat more pronounced in the PSID. Furthermore, PSID consumption expenditure data also display a much weaker aggregate recovery than what is observed in the NIPA data.[g]

2.2 Inequality Before the Great Recession

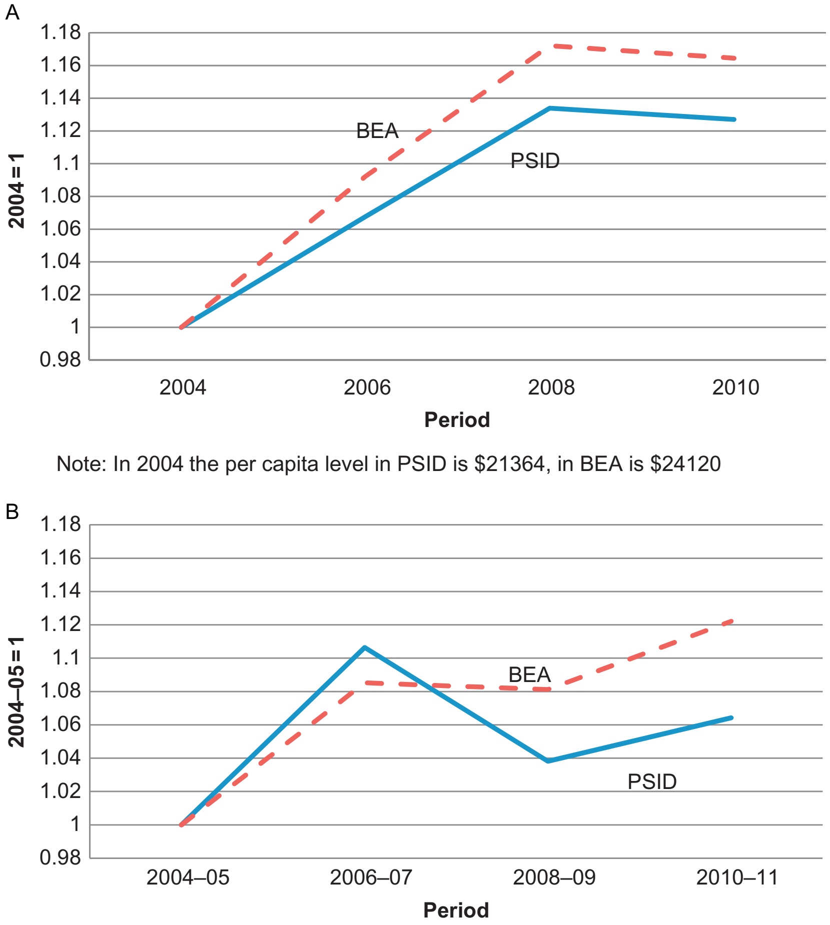
In this section, we document basic inequality facts in the United States for the year 2006, just before the Great Recession hit the economy. Since the Great Recession greatly affected households in the labor market, and our models below focus on labor earnings

[d] Empirical analyses of the joint wealth, income, and consumption distribution using the same panel data set are also contained in Fisher et al. (2015) for the United States, and in Krueger and Perri (2011) for Italy. See Skinner (1987), Blundell et al. (2008), and Smith and Tonetti (2014) for an alternative method for constructing an income-consumption panel using both the PSID and the Consumer Expenditure Survey (CE).

[e] For representative contributions, see eg, Piketty and Saez (2003), Krueger and Perri (2006), Krueger et al. (2010), Piketty (2014), Aguiar and Bils (2015), Atkinson and Bourguignon (2015), and Kuhn and Rios-Rull, 2015.

[f] In Section A.1, we describe in detail how these series are constructed.

[g] As Heathcote et al. (2010) document, this discrepancy between macro data and aggregated micro data is also observed in previous recoveries from US recessions.



Note: In 2004 the per capita level in PSID is $21364, in BEA is $24120

Note: In 2004–05 the per capita level in PSID is $15084, in BEA is $18705

Fig. 1 The Great Recession in the NIPA and in the PSID data. (A) Per capita disposable income. (B) Per capita consumption expenditures.

risk, we restrict attention to households with heads between ages 25 and 60, which in 2006 represents slightly less than 80% of total households in the PSID. Table 1 reports statistics that characterize, for this group of households, the distributions of four key variables: earnings, disposable income, consumption expenditures, and net worth. Our definition of earnings captures income sources that we will model as exogenous to household choices; they include all sources of labor income plus transfers (but not including unemployment benefits) minus tax liabilities.[h] Disposable income includes earnings

[h] During the Great Recession, transfers and taxes have played an important role in affecting household income dynamics. See, for example, Perri and Steinberg (2012).

Table 1 Means and Marginal Distributions in 2006

	Variable						
	Earn.	**Disp. *Y***		**Cons. Exp**		**Net Worth**	
Source	**PSID**	**PSID**	**CPS**	**PSID**	**CE**	**PSID**	**SCF (2007)**
Mean (2006$)	54,349	64,834	60,032	42,787	47,563	324,951	538,265
% Share by:							
Q1	3.6	4.5	4.4	5.6	6.5	−0.9	−0.2
Q2	9.9	9.9	10.5	10.7	11.4	0.8	1.2
Q3	15.3	15.3	15.9	15.6	16.4	4.4	4.6
Q4	22.7	22.8	23.1	22.4	23.3	13.0	11.9
Q5	48.5	47.5	46.0	45.6	42.4	82.7	82.5
90–95	10.9	10.8	10.1	10.3	10.2	13.7	11.1
95–99	13.1	12.8	12.8	11.3	11.1	22.8	25.3
Top 1%	8.0	8.0	7.2	8.2	5.1	30.9	33.5
Gini	0.43	0.42	0.40	0.40	0.36	0.77	0.78
Sample size	6232	6232	54,518	6232	4908	6232	2910

plus unemployment benefits, plus income from capital, including rental equivalent income of the main residence of the household. Consumption expenditures include all expenditure categories reported by the PSID, ie, cars and other vehicles purchases, food at home and away, clothing and apparel, housing including rent and imputed rental services for owners, household equipment, utilities and transportation expenses. Finally, net worth includes the value of the sums of households' assets minus liabilities.[i]

Table 1 reports, for each variable (earnings, disposable income, consumption expenditures, and net worth), the cross-sectional average (in 2006 dollars), as well as the share of the total value held by each of the five quintiles of the corresponding distribution. At the bottom of the table, we also report the share held by the households between the 90th and 95th percentile, between the 95th and 99th percentile, by those in the top 1% of the respective distribution, and the Gini index of concentration. All statistics are computed from PSID data, but for disposable income, consumption expenditures, and net worth we also compare the statistics from the PSID with the same statistics computed from alternative micro data sets. In particular, for disposable income we use households from the 2006 Current Population Survey (CPS), which is a much larger sample often used to compute income inequality statistics. For consumption expenditures, we use household

[i] Assets include the value of farms and of any businesses owned by the household, the value of checking/saving accounts, the value of stocks or bonds owned, the value of primary residence and of other real estate assets, the value of vehicles, and the value of individual retirement accounts. Liabilities include any form of debt including mortgages on the primary residence or on other real estate, vehicle debt, student loans, medical debt, and credit card debt.

data from the 2006 Consumer Expenditure Survey (CE). Finally, for net worth we use the 2007 Survey of Consumer Finances (SCF), which is the most commonly used dataset for studying the US wealth distribution.

The table reveals features that are typical of distributions of resources across households in developed economies. Earnings and disposable income are both quite concentrated, with the bottom quintiles of the respective distributions holding shares smaller than 5% (3.6% and 4.5% to be exact) and the top quantiles holding almost 50% (48.5% and 47.5% to be precise). The distributions of earnings and disposable income look quite similar, since for the households in our sample (ages 25 to 60), capital income is a fairly small share of total disposable income (constituting only roughly 1/6 of disposable income).[j] Note also that the distributions of disposable income in PSID and CPS look quite similar.[k]

The table also shows that consumption expenditures are less unequally distributed than earnings or income, with the bottom quintile accounting for a bigger fraction (5.6%) of total expenditures. The distributions of consumption expenditures in the PSID and the CE are also fairly comparable.

Finally, net worth is by far the most concentrated variable, especially at the top of the distribution. The bottom 40% of households hold essentially no net worth at all, whereas the top quintile owns 83% of all wealth, and the top 10% holds around 70% of total wealth. Comparing the last two columns demonstrates that, although the average level of wealth in the PSID is substantially lower than in the SCF, the distribution of wealth across the five quintiles lines up quite closely between the two data sets, suggesting that the potential underreporting or mismeasurement of wealth in the PSID might affect the overall amount of wealth measured in this data set, but not the cross-sectional distribution too significantly, which is remarkably comparable to that in the SCF.

Although the marginal distributions of earnings, income, and wealth are interesting in their own right, the more relevant object for our purposes is the joint distribution of wealth, earnings, disposable income, and consumption expenditures.[l] To document

[j] Recall that our definition of earnings is net of taxes and it already includes government transfers.

[k] The CPS income has a lower mean because it does not include the rental equivalent from the main residence. Notice also that both distributions are much less concentrated at the top than are income distributions computed by using tax returns, as in Piketty and Saez (2003). Two reasons account for this difference. The first is that Piketty–Saez focus on income measures before taxes and transfers, whereas here we restrict attention to after-tax and after-transfers income, which is less concentrated; the second is that they focus on tax units, which is a unit of a analysis different than households. See Burkhauser et al. (2012) for more on this distinction.

[l] The class of models we will construct below will have wealth—in addition to current earnings—as the crucial state variable, and thus we stress the correlation of net worth with earnings, income, and especially consumption here.

Table 2 PSID Households across the net worth distribution: 2006

	% Share of:			% Expend. Rate		Head's	
NW Q	**Earn.**	**Disp. *Y***	**Expend.**	**Earn.**	**Disp. *Y***	**Age**	**Edu. (yrs)**
Q1	9.8	8.7	11.3	95.1	90.0	39.2	12
Q2	12.9	11.2	12.4	79.3	76.4	40.3	12
Q3	18.0	16.7	16.8	77.5	69.8	42.3	12.4
Q4	22.3	22.1	22.4	82.3	69.6	46.2	12.7
Q5	37.0	41.2	37.2	83.0	62.5	48.8	13.9
	Correlation with net worth						
	0.26	0.42	0.20				

the salient features of this joint distribution, we divide the households in our 2006 PSID sample into net worth quintiles, and then for each *net worth quintile* we report, in Table 2, key differences across these wealth groups.

The table shows two important features of the data. The first is that, perhaps not surprisingly, households with higher net worth tend to have higher earnings and higher disposable incomes. The last row of the table shows more precisely the extent to which earnings and disposable income are positively correlated with net worth. One simple explanation for this is that wealthier households tend to be older and more educated, as confirmed by the last two columns of the table. The second observation is that consumption expenditures are also positively correlated with net worth, but less so than the two income variables. The reason is that, as can be seen in the last two columns of the table, the lower is net worth, the higher the consumption rate. We measure the consumption rate by computing total consumption expenditures for a specific wealth quintile and then dividing it by total earnings (or disposable income) in that wealth quintile. The differences in the consumption rates across wealth quintiles are economically significant: for example, between the bottom and the top wealth quintile, the differences in the consumption rates range between 20% and 30%.

Another way to look at the same issue is to notice that the households in the bottom two net worth quintiles, basically hold no wealth (see Table 1), but are responsible for 11.3% + 12.4% = 23.7% of total consumption expenditures (see Table 2), making this group quantitatively consequential for aggregate consumption dynamics. The differences across groups delineated by wealth constitute *prima facie* evidence that the shape of the wealth distribution *could* matter for the aggregate consumption response to macroeconomic shocks such as the ones responsible for the Great Recession.

In the next section, we will go beyond household heterogeneity at a given point in time and empirically evaluate how, during the Great Recession, expenditures and saving behavior changed differentially for households across the wealth distribution.

2.3 The Great Recession Across the Income and Wealth Distributions

In Table 3, we report for all households, and for households in each of the five quintiles of the net worth distribution, the changes (both percentages and absolute) in net worth, percentage changes in disposable income, and consumption expenditures and change in consumption expenditure rates (in percentage points).[m] For each variable we first establish a benchmark (the growth rate in a nonrecession period) by reporting the change or growth rate for the 2004–06 period, and then report the same variable for the 2006–10 period, which covers the whole recession. To make the two measures comparable, all changes are annualized.[n]

Table 3 reveals a number of interesting facts that we want to highlight. From the first four columns of the table, notice that all groups of households experienced increases in net worth between 2004 and 2006, likely mainly because of the rapid growth in asset prices (stock prices and especially real estate prices) during this period, with low-wealth households experiencing the strongest percentage growth in wealth (but of course starting from very low levels: see again Table 1). Turning to disposable income (second variable of Table 3), we observe that households originally at the bottom of the wealth

Table 3 Annualized changes in selected variables across PSID net worth

	Net worth[a]				Disp. *Y* (%)		Cons. Exp.(%)		Exp. Rate (pp)	
	(1) 04–06		(2) 06–10		(3) 04–06	(4) 06–10	(5) 04–06	(6) 06–10	(7) 04–06	(8) 06–10
All	**15.7**	**44.6**	**−3.0**	**−10**	**4.1**	**1.2**	**5.6**	**−1.3**	**0.9**	**−1.6**
NW Q										
Q1	NA	12.9	NA	6.6	7.4	6.7	7.1	0.6	−0.2	−4.2
Q2	121.9	19.5	24.4	3.7	6.7	4.1	7.2	2	0.3	−1.3
Q3	32.9	23.6	4.3	3.3	5.1	1.8	9	0	2.3	−1.1
Q4	17.0	34.7	1.7	3.8	5.0	1.7	5.9	−1.5	0.5	−2
Q5	11.6	132.2	−4.9	−68.4	1.8	−1.2	2.7	−3.5	0.5	−1.4

[a] The first figure is the percentage change (growth rate), the second is the change in 000's of dollars.

[m] To construct these changes, we keep the identity of the households fixed; for example, to compute the 2004–06 change in net worth for Q1 of the net worth distribution, we select all households in the bottom quintile of the wealth distribution in 2004, compute their average net worth (or income or consumption) in 2004 and 2006, and then calculate the percent difference between the two averages. For the consumption expenditure rates, we report percentage point differences.

[n] Table A.2 reports bootstrap standard errors for all figures in Table 3. In Tables A.3 and A.4, we separately report the changes for the 2006–08 and 2008–10 time periods.

distribution experience faster disposable income growth than those in the highest wealth quintile (7.4% vs 1.8%). This is most likely due to mean reversion in income: low-wealth households are also low-income households, and on average low income households experience faster income growth. Finally, expenditure growth roughly tracked the growth of income variables between 2004 and 2006, and as a result the consumption rates of each group remained roughly constant, perhaps with the exception of households initially in the middle quintile who experienced strong consumption expenditure growth, and thus their consumption rate displays a marked rise.

Now we turn to the dynamics in income, consumption, and wealth during the Great Recession. The columns labeled 06-10 display very significant changes in the dynamics of household income, consumption, and net worth throughout the wealth distribution, relative to the previous time period. Growth in net worth slowed down substantially for all households (it actually turned negative, from +15.6% to −3%) and most significantly so at the top of the wealth distribution. In fact, for households initially (that is, in 2006) in the top wealth quintile net worth fell 4.9% per year over the period 2006–10. Income growth also slowed down, although not uniformly across the wealth distribution. Table 3 shows that the slowdown in income growth is modest at the bottom of the wealth distribution (from 7.4% to 6.7%), whereas the middle and top quintiles experience a more substantial slowdown. For example, the fourth wealth quintile went from annual disposable income growth of 5% between 2004 and 2006 to a growth rate of 1.7% between 2006 and 2010.

Most important for our purposes is the change in consumption expenditures at different points in the wealth distribution, especially in relation to the magnitude of the associated earnings and disposable income changes (as evident in the movement of the consumption rates over time). The first fact we want to highlight is that, overall, PSID households cut the growth in expenditures from +5.6% to −1.3%. Although the decline in the growth rate of consumption expenditures is sizeable across all quintiles, the fall is most pronounced at the bottom of the wealth distribution. To highlight the starkest differences across the wealth distribution, focus on the difference between the top and the bottom wealth quintile. Between 2004 and 2006 the households in both the bottom and the top wealth quintiles display small (less than 0.5 percentage point) changes in the consumption rate (out of disposable income). By contrast, between 2006 and 2010, households at the bottom end of the 2006 wealth distribution reduced the change in their consumption rate by 4 percentage points (from −0.2% to −4.2%), whereas the top quintile's change in the consumption rate declined by only 1.9 percentage points (from 0.5% to −1.4%). In other words, during the Great Recession saving rates increased across the wealth distribution, but more strongly so at the bottom of the wealth distribution.[o]

[o] Heathcote and Perri (2015) also document a similar pattern using data from the Consumer Expenditure Survey.

Table 4 Decomposing changes in expenditure growth

	Change *C* growth	**Change *Y* growth**	**Change *C/Y* growth**
	$g_{c,t} - g_{c,t-1}$	$g_{y,t} - g_{y,t-1}$	$\frac{\rho_{it}-\rho_{it-1}}{\rho_{it-1}} - \frac{\rho_{it-1}-\rho_{it-2}}{\rho_{it-2}}$
All	−6.9	−2.9 (42%)	−3.8 (55%)
NW Q			
Q1	−6.5	−0.7 (11%)	−4.5 (69%)
Q2	−5.2	−2.6 (50%)	−2.3 (44%)
Q3	−9.0	−3.3 (37%)	−5.2 (58%)
Q4	−7.4	−3.3 (48%)	−3.8 (55%)
Q5	−6.2	−3.0 (42%)	−3.4 (55%)

To investigate the sources of the decline in expenditures growth across the wealth distribution in greater detail, we now decompose the difference in consumption growth across the two periods as follows:

$$g_{c,it} - g_{c,it-1} \simeq g_{y,it} - g_{y,it-1} + \frac{\rho_{it}-\rho_{it-1}}{\rho_{it-1}} - \frac{\rho_{it-1}-\rho_{it-2}}{\rho_{it-2}}, \tag{1}$$

where $g_{c,it} = \frac{C_{it}-C_{it-1}}{C_{it-1}}$ is the growth rate of consumption expenditure for group i (for example households in the first wealth quintile in period $t-1$) across periods t and $t-1$, $g_{y,it}$ is the same measure for disposable income, and $\rho_{it} = \frac{C_{it}}{Y_{it}}$ is the consumption rate out of disposable income for group i in period t.

The first column of Table 4 reports the changes in consumption growth rates for all households and for each group, ie, the term $g_{c,it} - g_{c,it-1}$, which is the difference between column (6) and column (5) in Table 3. The second and third columns of the table report the two right-hand-side terms from Eq. (1): the first term, labeled as change in disposable income growth Y, and the second term, labeled as change in the growth of the expenditure rate C/Y. Intuitively, if we see group i's consumption growth slowing down, it could be because its income growth is slowing down, ie, $g_{y,it} - g_{y,it-1}$ falls, or because, keeping income growth fixed, the growth in its expenditure rates, ie, $\frac{\rho_{it}-\rho_{it-1}}{\rho_{it-1}}$, falls. The numbers in parentheses in the table represent the relative contribution of each term.[p]

Overall Table 4 portrays a clear message. Households in the PSID reduce their expenditure growth significantly more than the slowdown in their disposable income alone would suggest (−6.9% vs 2.9%). This implies that, overall, households increase their

[p] The relative contributions do not sum to 1 as the decomposition in 1 is not exact, and it excludes terms that involve the product of growth rates.

saving rate. However, the increase in saving rates, although present among all wealth quintiles, is quantitatively most potent for the first quintile, ie, for those households entering the recession with the lowest net worth. Indeed, for these households the increase in the saving rate accounts for over two-thirds (69%) of the consumption growth decline, whereas for the other wealth groups consumption expenditure growth fell because both income growth slowed down and saving increased. We believe this fact is especially interesting, since it suggests that the decline in consumption at the bottom of the wealth distribution is not simply explained by standard hand-to-mouth behavior (ie, the decline in income of these households), but primarily by changes in consumption behavior though a decline in expenditure rates.

Having documented the salient features of the joint wealth, income, and consumption distribution in the United States prior to the Great Recession and their dynamics over the course of the downturn, we now proceed with a quantitative evaluation of how well standard economic theory, in the form of the canonical heterogeneous household business cycle model with uninsurable idiosyncratic earnings risk, can explain these patterns. We then use this model as a quantitative laboratory to assess the importance of cross-sectional household heterogeneity for aggregate business cycles.

3. A CANONICAL BUSINESS CYCLE MODEL WITH HOUSEHOLD HETEROGENEITY

In this section, we lay out the benchmark model on which this chapter is built. The model is a slightly modified version of the original Krusell and Smith (1998) real business cycle model with household wealth and preference heterogeneity[q] and shares many features of the model recently studied by Carroll et al. (2015).

3.1 Technology

In the spirit of real business cycle theory, aggregate shocks take the form of productivity shocks to the aggregate production function

$$Y = Z^{*}F(K,N). \tag{2}$$

Total factor productivity Z^{*} in turn is given by

$$Z^{*} = ZC^{\omega}, \tag{3}$$

where the exogenous part of technology Z follows a first-order Markov process with transition matrix $\pi(Z'|Z)$. Here C is aggregate consumption and the parameter $\omega \geq 0$

[q] Krusell and Smith (1998) in turn build on stationary versions of the model with household wealth heterogeneity, and thus on Bewley (1986), Imrohoroglu (1989), Huggett (1993), Huggett (1997), and Aiyagari (1994). See Deaton (1991) and Carroll (1992, 1997) for important early partial equilibrium treatments.

measures the importance of an aggregate demand externality. In the benchmark model, we consider the case of $\omega = 0$ in which case total factor productivity is exogenous and determined by the stochastic process for Z (and in which case we do not distinguish between Z and Z^*). In Section 7, we consider a situation with $\omega > 0$. In that case current TFP and thus output is partially determined by demand (aggregate consumption).

In either case, in order to aid the interpretation of the results, we will mainly focus on a situation in which the exogenous technology Z can take two values, $Z \in Z_l, Z_h$. We then interpret Z_l as a severe recession and Z_h as normal economic times.

Finally, we assume that capital depreciates at a constant rate $\delta \in [0, 1]$.

3.2 Household Demographics, Endowments, and Preferences

3.2.1 Demographics and the Life Cycle

In each period a measure 1 of potentially infinitely lived households populates the economy. Households are either young, working households (denoted by W) and participate in the labor market or are old and retired (and denoted by R). We denote a household's age by $j \in \{W, R\}$. Young households have a constant probability of retiring $1 - \theta \in [0, 1]$, and old households have a constant probability of dying $1 - \nu \in [0, 1]$. Deceased households are replaced by new young households. Given these assumptions, the distribution of the population across the two ages is given by

$$\Pi_W = \frac{1-\theta}{(1-\theta)+(1-\nu)}$$

$$\Pi_R = \frac{1-\nu}{(1-\theta)+(1-\nu)}.$$

This simple structure captures the life cycle of households and thus their life-cycle savings behavior in a parsimonious way.

3.2.2 Preferences

Households do not value leisure, but have preferences defined over stochastic consumption streams, determined by a period utility function $u(c)$ with the standard concavity and differentiability properties, as well as a time discount factor β that may be heterogeneous across households (but is fixed over time for a given household). Denote by B the finite set of possible time discount factors.

3.2.3 Endowments

Since households do not value leisure in the utility function, young households supply their entire time endowment (which is normalized to 1) to the market. However, they face idiosyncratic labor productivity and thus earnings risk. This earnings risk comes from two sources. First, households are subject to unemployment risk. We denote by $s \in S = \{u, e\}$ the current employment status of a household, with $s = u$ indicating unemployment. Employment follows a first-order Markov chain with transitions $\pi(s'|s, Z', Z)$ that depend on the aggregate state of the world. This permits the dependence of unemployment-employment transitions on the state of the aggregate business cycle.

In addition, conditional on being employed, a household's labor productivity $y \in Y$ is stochastic and follows a first order Markov chain; denote by $\pi(y'|y) > 0$ the conditional probability of transiting from state y today to y' tomorrow, and by $\Pi(y)$ the associated (unique) invariant distribution. In the benchmark model we assume that, conditional on being employed, transitions of labor productivity are independent of the aggregate state of the world.[r]

For both idiosyncratic shocks (s, y) we assume a law of large numbers, so that idiosyncratic risk averages out, and only aggregate risk determines the number of agents in a specific idiosyncratic state $(s, y) \in S \times Y$. Furthermore, we assume that the share of households in a given idiosyncratic employment state s only depends on the current aggregate state[s] Z, and thus denote by $\Pi_Z(s)$ the deterministic fraction of households with idiosyncratic unemployment state s if the aggregate state of the economy is given by Z. We denote the cross-sectional distribution over labor productivity by $\Pi(y)$; by assumption this distribution does not depend on the aggregate state Z.

Households can save (but not borrow)[t] by accumulating (moderately risky) physical capital[u] and have access to perfect annuity markets.[v] We denote by $a \in A$ the asset holdings of an individual household and by A the set of all possible asset holdings. Households are born with zero initial wealth, draw their unemployment status according to $\Pi_Z(s)$ and their initial labor productivity from $\Pi(y)$. The cross-sectional population distribution of employment status s, labor productivity y, asset holdings a, and discount factors β is denoted as Φ and summarizes, together with the aggregate shock Z, the aggregate state of the economy at any given point in time.

[r] Even for the unemployed, the potential labor productivity y evolves in the background and determines the productivity upon finding a job, as well as unemployment benefits while being unemployed, as described below.

[s] This assumption imposes consistency restrictions on the transition matrix $\pi(s'|s, Z', Z)$. By assumption, the cross-sectional distribution over y is independent of Z to start with.

[t] We therefore abstract from uncollateralized household debt, as modeled in Chatterjee et al. (2007) and Livshits et al. (2007). Herkenhoff (2015) provides an investigation of the impact of increased access to consumer credit on the US business cycle.

[u] We therefore abstract from household portfolio choice. See Cocco et al. (2005) for the analysis of portfolio choice in a canonical partial equilibrium model with idiosyncratic risk, and Krusell and Smith (1997) and Storesletten et al. (2007) for general equilibrium treatments.

[v] Thus the capital of the deceased is used to pay an extra return on capital $\frac{1}{\nu}$ of the retired survivors.

3.3 Government Policy

3.3.1 Unemployment Insurance

The government implements a balanced budget unemployment insurance system whose size is parameterized by a replacement rate $\rho = \frac{b(y, Z, \Phi)}{w(Z, \Phi)y}$ that gives benefits b as a fraction of potential earnings wy of a household, with $\rho = 0$ signifying the absence of public social insurance against unemployment risk.[w] These benefits are paid to households in the unemployment state $s = u$ and financed by proportional taxes on labor earnings with tax rate $\tau(Z, \Phi)$. Taxes are levied on both labor earnings and unemployment benefits.

Recall that by assumption the number of unemployed $\Pi_Z(u)$ only depends on the current aggregate state. The budget constraint of the unemployment insurance system then reads as

$$\Pi_Z(u)\sum_y \Pi(y)b(y, Z, \Phi) = \tau(Z, \Phi)\left[\sum_y \Pi(y)[\Pi_Z(u)b(y, Z, \Phi) + (1-\Pi_Z(u))w(Z, \Phi)y]\right].$$

Exploiting the fact that $b(y, Z, \Phi) = \rho w(Z, \Phi)y$ and that the cross-sectional distribution over y is identical among the employed and unemployed we can simply write

$$\Pi_Z(u)\rho = \tau(Z, \Phi)[\Pi_Z(u)\rho + (1-\Pi_Z(u))]$$

and conclude that the tax rate needed to balance the budget satisfies

$$\tau(Z, \Phi; \rho) = \left(\frac{\Pi_Z(u)\rho}{1-\Pi_Z(u)+\Pi_Z(u)\rho}\right) = \left(\frac{1}{1+\frac{1-\Pi_Z(u)}{\Pi_Z(u)\rho}}\right) = \tau(Z; \rho) \in (0, 1). \quad (4)$$

That is, the tax rate $\tau(Z; \rho)$ only depends (positively) on the exogenous policy parameter ρ measuring the size of the unemployment system as well as (negatively) on the exogenous ratio of employed to unemployed $\frac{1-\Pi_Z(u)}{\Pi_Z(u)}$, which in turn varies over the business cycle.

3.3.2 Social Security

The government runs a balanced budget PAYGO system whose size is determined by a constant payroll tax rate τ_{SS} (that applies only to labor earnings). Social security benefits $b_{SS}(Z, \Phi)$ of retirees are assumed to be independent of past contributions, but because of

[w] Recall that even unemployed households carry with them the idiosyncratic state y even though it does not affect their current labor earnings since they are unemployed.

fluctuations in the aggregate tax base will vary with the aggregate state of the economy Z. The budget constraint then determines the relationship between benefits and the tax rate according to

$$b_{SS}(Z,\Phi)\Pi_R = \tau_{SS}\Pi_W\left[\sum_{\gamma}\Pi(\gamma)(1-\Pi_Z(u))w(Z,\Phi)\gamma\right],$$

Note that in the absence of unemployment (and with average labor productivity of working people equal to 1), we have

$$\tau_{SS} = \frac{b_{SS}(Z,\Phi)}{w(Z,\Phi)}\frac{\Pi_R}{\Pi_W}$$

In this case, the social security tax rate is simply equal to the average replacement rate $\frac{b_{SS}(Z,\Phi)}{w(Z,\Phi)}$ times the old age dependency ratio $\frac{\Pi_R}{\Pi_W}$.

3.4 Recursive Competitive Equilibrium

As is well known, the state space in this economy includes the entire cross-sectional distribution Φ of individual characteristics,[x] (j, s, γ, a, β). Since the dynamic programming problems of young, working age households and retired households differ significantly from each other (in terms of both individual state variables as well the budget constraint) it makes notation easier to separate age $j \in \{W, R\}$ from the other state variables. The dynamic programming problem of retired households then reads as

$$v_R(a,\beta;Z,\Phi) = \max_{c,a'\geq 0}\left\{u(c) + \nu\beta\sum_{Z'\in Z}\pi(Z'|Z)v_R(a',\beta;Z',\Phi')\right\}$$

subject to

$$c + a' = b_{SS}(Z,\Phi) + (1 + r(Z,\Phi) - \delta)a/\nu$$
$$\Phi' = H(Z,\Phi',Z')$$

[x] In order to make the computation of a recursive competitive equilibrium feasible, we follow Krusell and Smith (1998), and many others since, and define and characterize a recursive competitive equilibrium with boundedly rational households who use only a small number of moments (and concretely here, just the mean) of the wealth distribution to forecast future prices. For a discussion of the various alternatives in computing equilibria in this class of models, see the January 2010 special issue of the *Journal of Economic Dynamics and Control*.

For working household households, the decision problem is given by

$$v_W(s,y,a,\beta;Z,\Phi)=\left\{\max_{c,a'\geq 0} u(c)+\beta\sum_{(Z',s',y')\in(Z,S,Y)}\pi(Z'|Z)\pi(s'|s,Z',Z)\pi(y'|y)\right.$$
$$\left.\times\left[\theta v_W(s',y',a',\beta;Z',\Phi')+(1-\theta)v_R(a',\beta;Z',\Phi')\right]\right\}$$

subject to

$$c+a'=(1-\tau(Z;\rho)-\tau_{SS})w(Z,\Phi)y[1-(1-\rho)1_{s=u}]+(1+r(Z,\Phi)-\delta)a$$
$$\Phi'=H(Z,\Phi',Z'),$$

where $1_{s=u}$ is the indicator function that takes the value 1 if the household is unemployed, and thus labor earnings equal unemployment benefits $b(y, Z, \Phi) = \rho w(Z, \Phi)y$.

Definition 1 A recursive competitive equilibrium is given by value and policy functions of working and retired households, v_j, c_j, a_j', pricing functions r, w, and an aggregate law of motion H such that

1. Given the pricing functions r, w, the tax rate given in Eq. (4), and the aggregate law of motion H, the value function v solves the household Bellman equation above and c, a' are the associated policy functions.
2. Factor prices are given by

$$w(Z,\Phi)=ZF_N(K(Z,\Phi),N(Z,\Phi))$$
$$r(Z,\Phi)=ZF_K(K(Z,\Phi),N(Z,\Phi)).$$

3. Budget balance in the unemployment system: Eq. (4) is satisfied
4. Market clearing

$$N(Z,\Phi)=(1-\Pi_Z(u))\sum_{y\in Y}y\Pi(y)$$
$$K(Z,\Phi)=\int a d\Phi.$$

5. The aggregate law of motion H is induced by the exogenous stochastic processes for idiosyncratic and aggregate risk as well as the optimal policy function a' for assets.[y]

3.5 A Taxonomy of Different Versions of the Model

Table 5 summarizes the different versions of the model we will study in this chapter, including the section of the chapter in which it will appear. We start with a version of the model in which total factor productivity is exogenous. The only source of propagation of the aggregate shocks is the capital stock, which is predetermined in the short run (and thus output is exogenous), but responds in the medium run to technology

[y] We give the explicit statement of the law of motion H in Appendix B.

Table 5 Taxonomy of different versions of the model used in the chapter

Name	Discounting	Techn.	Soc. Ins.	Section
KS	$\beta = \overline{\beta}$	$\omega = 0$	$\rho = 1\%$	Section 6.1
Het. β	$\beta \in [\overline{\beta} - \epsilon \overline{\beta} + \epsilon]$	$\omega = 0$	$\rho = 50\%$	Section 6.1
Het. β	$\beta \in [\overline{\beta} - \epsilon \overline{\beta} + \epsilon]$	$\omega = 0$	$\rho = 10\%$	Section 6.3
Dem. Ext.	$\beta \in [\overline{\beta} - \epsilon \overline{\beta} + \epsilon]$	$\omega > 0$	$\rho = 50\%$	Section 7

shocks and/or reforms of the social insurance system. We study two versions of the model, the original Krusell–Smith (1998) economy without preference heterogeneity (which we will alternatively refer to as the KS economy, the low–wealth inequality economy, or the homogeneous discount factor economy), and a model with permanent discount factor heterogeneity (which we refer to as the high–wealth inequality economy, the heterogeneous discount factor economy, or simply the benchmark economy). The latter economy also features an unemployment insurance system whose size is consistent with US data. In Section 5.1, we discuss the extent to which both versions of this model match the empirically observed US cross-sectional wealth distribution, and in Section 6.1 we trace out the model-implied aggregate consumption, investment, and output dynamics in response to a Great Recession type shock.

In order to assess the interaction of wealth inequality and social insurance policies for aggregate macro dynamics, in Section 6.3, we study a version of the heterogeneous discount factor economy with smaller unemployment insurance. In Section 7, the assumption of exogenous TFP is relaxed, and we present a version of the model in which TFP and thus output is partially demand-determined. In this version of the model, household heterogeneity has a potential impact not only on the size of the consumption recession, but also on the magnitude of the output decline, and by stabilizing individual consumption demand, unemployment insurance may act as a quantitatively important source of macroeconomic stabilization.

4. CALIBRATION OF THE BENCHMARK ECONOMY

In this section, we describe how we map our economy to the data. Since we want to address business cycles and transitions into and out of unemployment, we calibrate the model to *quarterly* data.

4.1 Technology and Aggregate Productivity Risk

Following Krusell and Smith (1998), we assume that output is produced according to a Cobb-Douglas production function

$$Y = ZK^{\alpha}N^{1-\alpha}. \tag{5}$$

We set the capital share to $\alpha = 36\%$ and assume a depreciation rate of $\delta = 2.5\%$ per quarter. For the aggregate technology process, we assume that aggregate productivity Z can take two values $Z \in \{Z_l, Z_h\}$, where we interpret Z_l as a potentially severe recession. The aggregate technology process is assumed to follow a first-order Markov chain with transitions

$$\pi = \begin{pmatrix} \rho_l & 1-\rho_l \\ 1-\rho_h & \rho_h \end{pmatrix}.$$

The stationary distribution associated with this Markov chain satisfies

$$\Pi_l = \frac{1-\rho_h}{2-\rho_l-\rho_h}$$
$$\Pi_h = \frac{1-\rho_l}{2-\rho_l-\rho_h}$$

With the normalization that $E(Z) = 1$, the aggregate productivity process is fully determined by the two persistence parameters ρ_l, ρ_h and the dispersion of aggregate productivity, as measured by Z_l/Z_h.

For the calibration of the aggregate productivity process, we think of a $Z = Z_l$ realization as a severe recession such as the Great Recession or the double-dip recession of the early 1980s (and a realization of $Z = Z_h$ as normal times). In this interpretation of the model, by choice of the parameters ρ_l, ρ_h, Z_l/Z_h we want the model to be consistent with the fraction of time periods spent in severe recessions, their expected length conditional on slipping into one, and the decline in GDP per capita associated with severe recessions.[z]

For this we note that with the productivity process set out above, the fraction of time spent in severe recessions is Π_l, whereas, conditional on falling into one, the expected length is given by

$$EL_l = 1 \times 1 - \rho_l + 2 \times \rho_l(1-\rho_l) + \ldots = \frac{1}{1-\rho_l}. \tag{6}$$

This suggests the following calibration strategy:

1. Choose ρ_l to match the average length of a severe recession EL_l. This is a measure of the persistence of recessions.
2. Given ρ_l, choose ρ_h to match the fraction of time the economy is in a severe recession, Π_l.
3. Choose $\frac{Z_l}{Z_h}$ to match the decline in GDP per capita in severe recessions relative to normal times.

In order to measure the empirical counterparts of these entities in the data, we need an operational definition of a severe recession. This definition could be based on GDP per

[z] This chapter shares the focus on rare but large economic crises with the body of work on rare disasters, see eg, Rietz (1988), Barro (2006), and Gourio (2013).

capita, total factor productivity, or unemployment rates, given the model assumption that the aggregate unemployment rate $\Pi_Z(y_u)$ is only a function of the aggregate state of the economy Z.

We chose the latter and define a severe recession to be one where the unemployment rate rises above 9% at least for one quarter and determine the length of the recession to be the period for which the unemployment rate remains above 7%. Using this definition over period from 1948.I to 2014.III we identify two severe recession periods: from 1980.II to 1986.II and from 2009.I to 2013.III. This delivers a frequency of severe recessions of $\Pi_l = 16.48\%$ with expected length of 22 quarters. The average unemployment rate in these severe recession periods is $u(Z_l) = 8.39\%$ and the average unemployment rate in normal times is $u(Z_h) = 5.33\%$. The implied Markov transition matrix that delivers this frequency and length of severe recessions has $\rho_l = 0.9545$ and $\rho_h = 0.9910$ and thus is given by

$$\pi = \begin{pmatrix} 0.9545 & 0.0455 \\ 0.0090 & 0.9910 \end{pmatrix}.$$

For the ratio $\frac{Z_l}{Z_h}$, we target a value of $\frac{Y_l}{Y_h} = 0.9298$, that is, a drop in GDP per capita of 7% relative to normal times.[aa] With average labor productivity if employed equal to 1 and if unemployed equal to zero, unemployment rates in normal and recession states equal to $u(Z_l) = 8.39\%$ and $u(Z_h) = 5.33\%$, and a capital share $\alpha = 0.36$, this requires $\frac{Z_l}{Z_h} = 0.9614$, which, together with the normalization

$$Z_l \Pi_l + Z_h \Pi_h = 1.$$

determines the levels of Z as $Z_l = 0.9676$, $Z_h = 1.0064$. Note that because of endogenous dynamics of the capital stock which falls significantly during the recession, the dispersion in total factor productivity is smaller than what would be needed to engineer a drop in output by 7% only through TFP and increased unemployment (which is the drop in output on impact, given that the capital stock is predetermined).[ab]

[aa] This is the decline in real GDP per capita during the two recession periods we identified, after GDP per capita is linearly detrended. The exact magnitude of the real GDP per capita decline is not crucial for our results, but it is important that severe recessions are deeper and (especially) more persistent than regular business cycle fluctuations.

[ab] In the short run,

$$\frac{Y_l}{Y_h} = \frac{Z_l}{Z_h}\left(\frac{1-u(Z_l)}{1-u(Z_h)}\right)^{0.64}$$

so that in order to generate a drop in output of 7% in the short run would require

$$\frac{Z_l}{Z_h} = \frac{0.9298}{\left(\frac{0.9161}{0.9467}\right)^{0.64}} = 0.9496.$$

.

4.2 Idiosyncratic Earnings Risk

Recall that households face two types of idiosyncratic risks: countercyclical unemployment risk described by the transition matrices $\pi(s'|s,Z',Z)$ and, conditional on being employed, acyclical earnings risk determined by $\pi(y'|y)$. We describe both components in turn.

4.2.1 Unemployment Risk

Idiosyncratic unemployment risk is completely determined by the four 2 by 2 transition matrices $\pi(s'|s,Z',Z)$ summarizing the probabilities of transiting in and out of unemployment for each (Z,Z') combination. Thus $\pi(s'|s,Z',Z)$ has the form

$$\begin{bmatrix} \pi_{u,u}^{Z,Z'} & \pi_{u,e}^{Z,Z'} \\ \pi_{e,u}^{Z,Z'} & \pi_{e,e}^{Z,Z'} \end{bmatrix}, \tag{7}$$

where, for example, $\pi_{e,u}^{Z,Z'}$ is the probability that an unemployed individual finds a job between one period and the next, when aggregate productivity transits from Z to Z'. Evidently each row of this matrix has to sum to 1. Note that, in addition, the restriction that the aggregate unemployment rate only depends on the aggregate state of the economy imposes one additional restriction on each of these 2 by 2 matrices, of the form

$$\Pi_{Z'}(u) = \pi_{u,u}^{Z,Z'} \times \Pi_Z(u) + \pi_{e,u}^{Z,Z'} \times (1 - \Pi_Z(u)). \tag{8}$$

Thus, conditional on targeted unemployment rates in recessions and expansions, (Π_l, Π_h) this equation imposes a joint restriction on $(\pi_{u,u}^{Z,Z'}, \pi_{e,u}^{Z,Z'})$ for each (Z,Z') pair. With these restrictions, the idiosyncratic transition matrices are uniquely pinned down by $\pi_{u,e}^{Z,Z'}$, ie, the job-finding rates.[ac]

We compute the job finding rate for a quarter as follows. We consider an individual that starts the quarter as unemployed and compute the probability that at the end of the quarter that individual is still unemployed. The possible ways that this can happen are (denoting as f_1, f_2, f_3 and as s_1, s_2, s_3 the job-finding and job-separation rates in months 1,2, and 3 of the quarter):

1. Does not find a job in month 1, 2, or 3, with probability $(1 - f_1) \times (1 - f_2) \times (1 - f_3)$.
2. Finds a job in month 1, loses it in month 2, does not find in month 3, with probability $f_1 \times s_2 \times (1 - f_3)$.
3. Finds a job in month 1, keeps it in month 2, loses it in month 3, with probability $f_1 \times (1 - s_2) \times s_3$.
4. Finds a job in month 2, loses it in month 3, with probability $(1 - f_1) \times f_2 \times s_3$.

[ac] One could alternatively use job-separation rates $\pi_{e,u}^{Z,Z'}$.

Thus the probability that someone that was unemployed at the beginning of the quarter is not unemployed at the end of the quarter is:

$$f = 1 - ((1 - f_1)(1 - f_2)(1 - f_3) + f_1 s_2 (1 - f_3) + f_1 (1 - s_2) s_3 + (1 - f_1) f_2 s_3) \tag{9}$$

We follow Shimer (2005) to measure the job-finding and separation rates from CPS data as averages for periods corresponding to specific Z, Z' transitions.[ad] Equating these with $\pi_{u,e}^{Z,Z'}$ delivers the following employment-unemployment transition matrices:

- Aggregate economy is and remains in a recession: $Z = Z_l . Z' = Z_l$

$$\begin{pmatrix} 0.3378 & 0.6622 \\ 0.0606 & 0.9394 \end{pmatrix} \tag{10}$$

- Aggregate economy is and remains in normal times: $Z = Z_h . Z' = Z_h$

$$\begin{pmatrix} 0.1890 & 0.8110 \\ 0.0457 & 0.9543 \end{pmatrix} \tag{11}$$

- Aggregate economy slips into recession: $Z = Z_h . Z' = Z_l$

$$\begin{pmatrix} 0.3382 & 0.6618 \\ 0.0696 & 0.9304 \end{pmatrix} \tag{12}$$

- Aggregate economy emerges from recession: $Z = Z_l . Z' = Z_h$

$$\begin{pmatrix} 0.2220 & 0.7780 \\ 0.0378 & 0.9622 \end{pmatrix} \tag{13}$$

We observe that the resulting matrices make intuitive sense. One possible (but quantitatively minor) exception is that the job-finding rate is higher if the economy remains in normal times than if it emerges from a recession. On the other hand, the lower job-finding rate is consistent with the experience during the Great Recession per our definition, as job-finding rates did not recover until well into 2014, whereas by our calibration the recession ended in 2013.

4.2.2 Earnings Risk Conditional on Employment

In addition to unemployment risk, we add to the model earnings risk, conditional on being employed. This allows us to obtain a more empirically plausible earnings distribution and makes earnings risk a more potent determinant of wealth dispersion (and thus reduces the importance of preference heterogeneity for this purpose). We assume that,

[ad] Let u_t = unemployment rate and u_t^S = short-term unemployment rate (people who are unemployed this month, but were not unemployed last month). Then we can define the monthly job-finding rate as $1 - (u_{t+1} - u_{t+1}^S)/u_t$ and the separation rate as $u_{t+1}^S/(1 - u_t)$. The series we use from the CPS are the unemployment level (UNEMPLOY), the short-term unemployment level (UNEMPLT5) and civilian employment (CE16OV). There was a change in CPS coding starting in February 1994 (inclusive), so UNEMPLT5 in every month starting with February 1994 is replaced by $UEMPL5 \times 1.1549$.

conditional on being employed, log-labor earnings of households follow a process with both transitory and persistent shocks.[ae] The process is specified as

$$\log(y') = p + \epsilon \tag{14}$$

$$p' = \phi p + \eta \tag{15}$$

with persistence ϕ and innovations of the persistent and transitory shocks (η, ϵ), respectively.[af] The associated variances of the shocks are denoted by $(\sigma_\eta^2, \sigma_\epsilon^2)$, and therefore the entire process is characterized by the parameters $(\phi, \sigma_\eta^2, \sigma_\epsilon^2)$. We estimate this process for household labor earnings after taxes (after first removing age, education and time effects) from *annual* PSID data and find estimates of $\phi, \sigma_\eta^2, \sigma_\epsilon^2$ equal to 0.9695, 0.0384 and 0.0522 respectively.[ag] Next we translate these estimates into a quarterly persistence and variance.[ah] We then use the Rouwenhorst procedure to discretize the persistent part of the process into a seven-state Markov chain.[ai] The *iid* shock only enters the computation of the expectation on the right-hand side of the Euler equation.[aj] We approximate the integral calculating the expectation using a Gauss–Hermite quadrature scheme with three nodes. Thus, we effectively approximate the continuous state space process by a discrete Markov chain with 7 × 3 = 21 states.[ak]

[ae] The formulation of log-earnings or log-income as a stochastic process with transitory and persistent (or fully permanent) shocks follows a large empirical literature in labor economics. See Meghir and Pistaferri (2004), Storesletten et al. (2004b), Guvenen (2009) and the many references discussed therein.

[af] Note that we assume that the variance and persistence of this process are independent of the state of the business cycle. Earnings risk in the data *is* countercyclical, as stressed by Storesletten et al. (2004a, 2007), and Guvenen et al. (2014); in our benchmark model earning risk is also countercyclical but only because of countercyclical unemployment risk.

[ag] For the exact definition of the labor earnings after taxes, sample selection criteria and estimation method, see Appendix A.

[ah] In order to ensure that quarterly log-earnings has the same persistence as annual log-earnings, we choose the persistence of the quarterly AR(1) to be $\phi = \hat{\phi}^{\frac{1}{4}}$. For the variances, we note that the main purpose of the earnings shocks is to help deliver a plausible cross-sectional distribution of labor income. Therefore we aim to maintain the same cross-sectional distribution of earnings at the quarterly frequency as we estimate at the annual frequency. Choosing a quarterly transitory variance equal to its annual counterpart and

$$\frac{\sigma_\eta^2}{1-\phi^2} = \frac{\hat{\sigma}_\eta^2}{1-\hat{\phi}^2}$$

achieves this goal.

[ai] See Kopecky and Suen (2010) for a detailed description and evaluation of the Rouwenhorst method.

[aj] This is because we use cash at hand and the persistent income state as state variables in the individual household dynamic programming problem.

[ak] For the computation of the distributional statistics we simulate a panel of households. In this simulation, realizations of the persistent shock remain on the grid, but the transitory shock is drawn from a normal distribution and thus is not restricted to fall on one of the quadrature points.

4.3 Preferences and the Life Cycle

In the benchmark economy, we assume that the period utility function over current consumption is given by a constant relative risk aversion utility function with parameter $\sigma = 1$. As described above, we study two versions of the model: the original Krusell–Smith (1998) economy in which households have identical time discount factors, and a model in which households, as in Carroll et al. (2015) have permanently different time discount factors (and die with positive probability, in order to ensure a bounded wealth distribution).

For the model with preference heterogeneity, we assume that households at the beginning of their life draw their permanent discount factor β from a uniform distribution[al] with support $[\overline{\beta} - \epsilon, \overline{\beta} + \epsilon]$ and choose $(\overline{\beta}, \epsilon)$ so that the model wealth distribution (with an unemployment insurance replacement rate of 50%) has a Gini coefficient for the working age population of 77% as in the data and a quarterly wealth-to-output ratio of 10.26 (as in Carroll et al., 2015) This requires $(\overline{\beta} = 0.9864, \epsilon = 0.0053)$ and implies that annual time discount factors in this economy range from $\beta = 0.9265$ to $\beta = 0.9672$. Finally, households in the working stage of their life cycle face a constant probability $1 - \theta$ of retiring, and retired households face a constant probability $1 - \nu$ of dying. For our quarterly model we choose $1 - \theta = 1/160$, implying an expected work life of 40 years, and $1 - \nu = 1/60$, with a resulting retirement phase of 15 years in expectation.

For the original Krusell–Smith economy, we choose the common quarterly discount factor $\beta = 0.9899$ to ensure that the capital-output ratio in this economy (again at quarterly frequency) equals that in the heterogeneous β economy. In this economy households neither retire nor die.

4.4 Government Unemployment Insurance Policy

The size of the social insurance (or unemployment insurance, more concretely) system is determined by the replacement rate ρ. For the benchmark economy that we assume $\rho = 50\%$ (see, eg, Gruber, 1994). We will also consider a lower value of $\rho = 10\%$, motivated by the observation that many households qualifying for unemployment insurance benefits fail to claim them (see, eg, Blank and Card, 1991 or Chodorow-Reich and Karabarbounis, 2016).

[al] In practice, we discretize this distribution and assume that each household draws one of five possible β's with equal probability; thus $B = \{\beta_1, \ldots \beta_5\}$ and $\Pi(\beta) = 1/5$. We also experimented with stochastic β's as in Krusell and Smith (1998) but found that the formulation we adopt enhances the model's ability to generate sufficiently many wealth-poor households. The results for the stochastic β economy generally lie in between those obtained in the original Krusell and Smith (1998) economy documented in detail in this chapter, and the results obtained in the model with permanent β heterogeneity, also documented in great detail below.

Finally, the payroll tax rate for social security is set to $\tau_{SS} = 15.3\%$. This choice implies an average (over the business cycle) and empirically plausible replacement rate of the social security system of approximately 40%. In the KS economy, in order to avoid numerical problems with zero consumption, we include a minimal unemployment insurance system with a replacement rate of $\rho = 1\%$.

5. EVALUATING THE BENCHMARK ECONOMY

5.1 The Joint Distribution of Earnings, Income, Wealth, and Consumption in the Benchmark Economy

In this section, we evaluate the extent to which our benchmark model is consistent with the main empirical facts characterizing the joint distribution of wealth, income, and consumption expenditures, as well as the changes in this distribution when the economy is subjected to a large negative aggregate shock.

5.1.1 Wealth Inequality in the Benchmark Economy

We have argued in the introduction that a model-implied cross-sectional wealth distribution that is consistent with the empirically observed concentration, and especially with a share of wealth of the bottom 40% of close to zero, is crucial when using the model as a laboratory for studying aggregate fluctuations. We now document that our benchmark economy has this property, whereas an economy akin to the one studied in Krusell and Smith's (1998) original work in which wealth inequality is entirely driven by idiosyncratic unemployment shocks and incomplete financial markets does not.[am]

Table 6 reports selected statics for the wealth distribution, those computed from the data (PSID and SCF) as well as those from two model economies, the original Krusell–Smith (1998) economy and our benchmark model with idiosyncratic income risk, incomplete markets, a rudimentary life cycle structure, unemployment insurance, and heterogeneous discount factors.[an] As indicated in the calibration section, through appropriate choice of the time discount factor(s), both economies have the same average (over the business cycle) capital-output ratio, and the benchmark economy displays a wealth Gini coefficient in line with the micro data from the PSID. All other moments of the empirical cross-sectional wealth distribution were not targeted in the calibration of the models.

[am] We retain *our* calibration of idiosyncratic unemployment risk, and thus the cross-sectional wealth distribution in our version of the Krusell–Smith economy differs from their original numbers, but not in a magnitude substantial enough to change any of the conclusions below.

[an] Recall that in the data, we restrict attention to working-age households. Consequently, when we report cross-sectional statistics from the benchmark model (which includes a retirement phase), we restrict attention to households in the working stages of their life.

Table 6 Net worth distributions: Data vs models

	Data		Models	
% Share held by:	**PSID, 06**	**SCF, 07**	**Bench**	**KS**
$Q1$	−0.9	−0.2	0.3	6.9
$Q2$	0.8	1.2	1.2	11.7
$Q3$	4.4	4.6	4.7	16.0
$Q4$	13.0	11.9	16.0	22.3
$Q5$	82.7	82.5	77.8	43.0
90–95	13.7	11.1	17.9	10.5
95–99	22.8	25.3	26.0	11.8
$T1\%$	30.9	33.5	14.2	5.0
Gini	0.77	0.78	0.77	0.35

From the table we note that, overall, the benchmark model fits the empirical wealth distribution in the data quite well (albeit not perfectly), especially at the bottom of the distribution. Specifically, it captures the fact that households constituting the bottom two quintiles of the wealth distribution hardly have any wealth, but also that the top wealth quintile holds approximately 80% of all net worth in the US economy. We also acknowledge that the benchmark model makes the wealth upper middle class (quintile 4 and also the bottom part of quintile 5) somewhat too wealthy. For example, households between the 90th and 99th percentiles of the net worth distribution account for about 36% of wealth in the data, but 44% in the model. Most problematically, the benchmark model still misses the wealth concentration at the *very top* of the distribution significantly. In the data the top 1% wealth holders account for over 30% of overall net worth in the economy, whereas the corresponding figure in the model is only 14.0%. A histogram of the model-implied wealth distribution can be found in Fig. 10.[ao]

Finally, Table 6 reproduces the well-known—since Krusell and Smith (1998)—result that transitory unemployment risk and incomplete financial markets alone are incapable of generating sufficient wealth dispersion. The problem relative to the data is two-fold: households at the top of the wealth distribution are not nearly wealthy enough, and, as we will argue, more importantly for the results to follow, households at the bottom of the distribution hold significantly too much wealth in the model. Relative to SCF or PSID micro data, in the model the bottom 40% own about 19% of net worth in the economy, whereas in the data that share is approximately 0. As a summary measure of wealth

[ao] Although this is clearly a shortcoming, note that in this range of wealth levels, the consumption function is essentially linear (as we will display below) and thus mechanically reshuffling wealth between the top 1% and the top 20% through top 1% would not alter aggregate consumption responses to shocks significantly. We will return to this point in Section 6.2.

inequality, whereas the wealth Gini in the data is well above 0.7, the original Krusell–Smith model delivers a number of only 0.35.

In the next section, we now decompose which model elements in the benchmark economy are responsible for generating a more realistic wealth distribution than in the original Krusell–Smith economy. We then turn to an evaluation of the benchmark model's success in reproducing the empirical *joint* distribution of earnings, income, consumption, and wealth in the data.

5.2 Inspecting the Mechanism I: What Accounts for Wealth Inequality in the Benchmark Economy?

A substantial literature, recently surveyed in De Nardi (2015), De Nardi et al. (2015), and Benhabib and Bisin (2016), explores alternative mechanisms for generating the empirically observed high wealth concentration in the data.[ap] These mechanisms include the use of very large but transient income realizations that the PSID misses out on (as in Castaneda et al., 2003; Kindermann and Krueger, 2015; or Brüggemann and Yoo, 2015), large uninsured or only partially insured medical expenditure shocks in old age (see eg, De Nardi et al., 2010 or Ameriks et al., 2015), the intergenerational transmission of wealth through accidental and intended bequests (as eg, in De Nardi, 2004), the interaction between wealth accumulation and entrepreneurship (see Quadrini, 2000; Cagetti and De Nardi, 2006; and Buera, 2009) or idiosyncratic shocks to investment opportunities or its returns, as in Benhabib et al. (2011).

In our benchmark model, we instead follow the sizeable literature that has explored the potential importance of empirically realistic, highly persistent earnings risk (conditional on employment) as well as preference heterogeneity in general, and cross-sectional dispersion in patience specifically, for generating an empirically plausible cross-sectional wealth distribution. Household heterogeneity in time discount factors had already been explored by the original Krusell and Smith (1998) paper, and has been further analyzed by Hendricks (2007) and Carroll et al. (2015); the latter also incorporates a stochastic earnings process in the analysis.

In the previous section, we argued that preference heterogeneity, when combined with idiosyncratic unemployment and earnings shocks as well as rudimentary life cycle elements[aq] and social insurance policies, generates a wealth distribution that resembles the

[ap] Gabaix et al. (2014) evaluate whether the existing theories discussed there are consistent with the secular rise in the share of income and wealth accruing to the top 1% of households, and argue that only theories embedding "superstar" phenomena are capable of reproducing the facts at the very top of these distributions.

[aq] The literature on quantitative studies of the cross-sectional wealth distributions in general equilibrium life-cycle economies with uninsurable idiosyncratic income risk starts with Huggett (1996).

Table 7 Net worth distributions and consumption decline: Different versions of the model

	Models[a]				
% Share:	**KS**	**+$\sigma(y)$**	**+Ret.**	**+$\sigma(\beta)$**	**+UI**
$Q1$	6.9	0.7	0.7	0.7	0.3
$Q2$	11.7	2.2	2.4	2.0	1.2
$Q3$	16.0	6.1	6.7	5.3	4.7
$Q4$	22.3	17.8	19.0	15.9	16.0
$Q5$	43.0	73.3	71.1	76.1	77.8
90–95	10.5	17.5	17.1	17.5	17.9
95–99	11.8	23.7	22.6	25.4	26.0
$T1\%$	5.0	11.2	10.7	13.9	14.2
Wealth Gini	0.350	0.699	0.703	0.745	0.767

[a]The KS model only has unemployment risk and incomplete markets, and thus the first column repeats information from Table 6. The column +$\sigma(y)$ adds idiosyncratic earnings shocks (transitory and permanent) while employed. The column +Ret. adds the basic life cycle structure (positive probability of retirement and positive probability of death, plus social security in retirement). The column +$\sigma(\beta)$ incorporates preference heterogeneity into the model, and finally the column +UI raises the replacement of the unemployment insurance system from 1% to 50%; the resulting model is therefore the benchmark model, with results already documented in Table 6. In all models, the (mean) discount factor is calibrated so that all versions have the same capital-output ratio.

data in 2006 well, both at the bottom and at the top of the distribution. In Table 7, we now show precisely which model elements are responsible for this finding.[ar]

The table (which partially repeats information from Table 6 to facilitate comparisons across different model economies) displays the share of net worth held by the five wealth quintiles, the wealth Gini, and more detailed information about the top of the net worth distribution, in the data and in a sequence of models, ranging from the original Krusell–Smith (1998) economy to our benchmark economy in the last column.

The table contains several important quantitative lessons. First, comparing the first and the second model columns, we see that the inclusion of highly persistent earnings risk, in addition to unemployment risk, increases wealth dispersion very significantly, relative to the economy with *only* unemployment risk. Consistent with a sizeable literature estimating stochastic labor earnings or income processes (see eg, Storesletten et al., 2004b) we find that the persistent component is indeed very persistent, with an annual autocorrelation (conditional on remaining employed) of 0.97. Thus, the economy contains a share of households with close to permanently low earnings, even in the absence of unemployment. These households, located predominantly in the lowest wealth quintile, have had no opportunity to accumulate significant wealth.[as] Consequently the share of

[ar] Castaneda et al. (1998) provide a decomposition similar in spirit, but focus on the evolution of the cross-sectional income distribution over the cycle.

[as] And if an unemployment insurance system with replacement rate of $\rho = 50\%$ is in place, as in the benchmark economy, they have no strong motive, either.

wealth held by the poorest households shrinks to fairly close to zero with idiosyncratic income risk, as observed in the data. At the same time, the top wealth quintile is populated with households with high earnings realizations for whom the risk of a persistent fall in earnings provides motivation to accumulate substantial wealth. As a result, the wealth Gini doubles in the economy with earnings risk, relative to the original Krusell–Smith unemployment-only model.

Second, adding a more explicit life-cycle structure does not change the wealth distribution (of the working-age population) much, but as we will see in the next section, will imply a more plausible *joint* wealth-consumption distribution, by adding a life-cycle savings for retirement motive to the precautionary saving motive. It also somewhat reduces wealth concentration at the top of the distribution, since earnings risk ceases with retirement and thus trims the precautionary motive of the wealth-rich.[at]

Third, as the examination of the very top of the wealth distribution in the first three columns of Table 7 reveals, income risk and life-cycle elements alone are insufficient to generate the very high wealth concentration observed in the data. This is where the discount factor heterogeneity in the benchmark model plays a crucial role. It creates a class of households that are patient and have a high propensity to save, and the fact that in addition to a precautionary saving motive, they also save for retirement (a phase they value highly because of their patience) ensures that they do not start to decumulate wealth even at high wealth levels. As Table 7 displays (comparing the last two columns), the model with both features (the life cycle and preference heterogeneity) is able to generate the wealth concentration at the top quintile of the distribution close to what is observed in US data (albeit not at the very top of the distribution).

Finally, inserting an unemployment insurance system into the model further reduces the wealth held by the bottom two quintiles of the distribution, since now losing a job with little net worth is not nearly as harmful. In Krueger et al. (2016), we argue that the size of the unemployment insurance system not only crucially shapes the bottom of the wealth distribution, but also has a strong impact on the welfare losses from severe recessions in the class of heterogeneous household macro models we study in this chapter.

5.2.1 Income and Consumption at Different Points of the Wealth Distribution

In this section, we evaluate the ability of the benchmark model to reproduce key features of the joint distribution of income, consumption, and wealth in the PSID data. To do so, Table 8 reports the share of earnings, disposable income, consumption expenditures, and the expenditure rates for the five quintiles of the wealth distribution, both for the data (as already contained in Table 2) and for the benchmark model.

[at] Our model imposes substantial structure on the link between idiosyncratic income shocks and consumption over the life cycle. In methodologically complementary work, Arellano et al. (2015) estimate a more flexible nonlinear empirical model of household earnings and consumption over the life cycle.

Table 8 Selected variables by net worth: Data vs models

	% Share of:						**% Expend. rate**			
	Earnings		**Disp. *Y***		**Expend.**		**Earnings**		**Disp. *Y***	
NW Q	**Data**	**Mod**	**Data**	**Mod**	**Data**	**Mod**	**Data**	**Mod**	**Data**	**Mod**
Q1	9.8	6.5	8.7	6.0	11.3	6.6	95.1	96.5	90.0	90.4
Q2	12.9	11.8	11.2	10.5	12.4	11.3	79.3	90.3	76.4	86.9
Q3	18.0	18.2	16.7	16.6	16.8	16.6	77.5	86.0	69.8	81.1
Q4	22.3	25.5	22.1	24.3	22.4	23.6	82.3	87.3	69.6	78.5
Q5	37.0	38.0	41.2	42.7	37.2	42.0	83.0	104.5	62.5	79.6
	Correlation with net worth									
	0.26	0.46	0.42	0.67	0.20	0.76				

On the positive side, first, the model is consistent with the significantly positive correlation between net worth on the one hand, and earnings, disposable income and consumption expenditures on the other. The shares of the latter three variables are all increasing with the net worth quintiles. Second, as in the data, disposable income (which includes capital income) displays a higher correlation with net worth than with labor earnings. Third, the model reproduces the crucial fact that the bottom two wealth quintiles, while accounting for essentially zero net worth, contribute a very significant share to aggregate consumption expenditures. In the data, that share is 23.7%, and in the model it is still highly significant at 17.9%. Since, as we will show below, this low-wealth group has the largest declines in their consumption, the fact that it accounts for a substantial part of aggregate consumption to start with is in turn crucial for the macro responses to an aggregate shock in the model. Fourth, turning to the consumption expenditure rates, the model is broadly consistent with the levels found in the data, and is broadly consistent with the empirical finding in the data that these rates decline with net worth. However, the wealth gradient is not quite as steep in the model as it is in the data, and in the model the top wealth quintile has expenditure rates that are higher than the forth quintile (very slightly so in relation to disposable income, much more strongly so in relation to labor earnings).

For this last finding, the inclusion of a retirement phase and thus a life cycle savings motive into the model is absolutely crucial. A pure infinite horizon version of the model, even with idiosyncratic income shocks and preference heterogeneity, displays expenditure rates that are significantly too high—averaging 100% across wealth quintiles—and implies expenditure rates that are U-shaped with respect to net worth. Absent the life-cycle savings motive, households accumulate wealth exclusively for the purpose of smoothing out negative income fluctuations, and thus individuals in the fourth and fifth wealth quintiles, having accumulated enough net worth for this purpose, display very high expenditure rates (in fact, significantly larger than 100%)—especially with respect to labor earnings. Preference heterogeneity mitigates this effect somewhat, but the resulting model still displays grossly counterfactual expenditure rates, whereas the version of the model with stochastic retirement brings the implications of the model much closer to their empirical counterpart, and is our primary justification for the presence of this model element.

We would also like to flag another dimension along which the model is not fully successful in capturing the empirical facts. First, although the model does generate consumption expenditure shares that are strongly increasing with wealth, not only are the wealth-poor too consumption-poor in the model (as already discussed above), but also the wealth rich (quintile 5) consume too much in the model (42% relative to 37.2% in the data). This is true even though the model captures the earnings and income share of this group of households quite well. This problem of the model is summarized by the fact that the correlation between net worth and consumption expenditures is positive in the model, as it is in the data, but is much larger than it is in the PSID.

We conclude this section with the overall assessment that the benchmark model captures well many qualitative features of the cross-sectional joint distribution of net worth, earnings, income and consumption expenditures, but fails to quantitatively match the joint distribution of net worth and expenditures, with the wealth poor consuming too little, and the wealth rich consuming too much, relative to the data.

5.3 The Dynamics of Income, Consumption, and Wealth in Normal Times and in a Recession

The previous section studied the joint distribution of the key economic variables at a given point in time (2006 in the data, a period after a long sequence of normal macroeconomic performance in the model). We now put the model to a more ambitious (and to our knowledge novel) test and assess whether the *dynamics* of wealth, income and consumption implied by the model can match those observed in the data. We ask this question both for a period of macroeconomic stability (in Section 5.3.1), and then, in Section 5.3.2, for a period characterized by a severe macroeconomic crisis. Note that none of the empirical moments along this dimension were targeted in the calibration of the model.

5.3.1 Normal Times: 2004–06

In the data we are somewhat limited in our choices by the sparse time series dimension of the PSID (for which comprehensive consumption data are available). We take *normal times* in the data to be the period from 2004 to 2006; we map this period into the model by studying an episode of eight quarters of good productivity, $Z = Z_h$, which in turn followed a long sequence of good aggregate shocks so that aggregates and distributions have settled down prior to this episode.

Table 9 Annualized changes in selected variables by net worth in normal times (2004-06): Data vs model

	Net worth (%)		Disp. *Y* (%)		Expend (%)		Exp. Rate (pp)	
NW Q	Data	Model	Data	Model	Data	Model	Data	Model
Q1	NaN	44	7.4	7.2	7.1	6.7	−0.2	−0.4
Q2	122	33	6.7	3.1	7.2	3.6	0.3	0.5
Q3	33	20	5.1	1.6	9	2.5	2.3	0.8
Q4	17	9	5	0.5	5.9	1.7	0.5	1.2
Q5	12	3	1.8	−1.0	2.7	0.5	0.5	1.4
All	16	5	4.1	0.7	5.6	1.8	0.9	0.7

Table 9 reports the statistics for the data (and thus repeats the information from Table 3) together with the model.[au] Recall from the description of Table 9 that for a given variable x (wealth, income, and consumption) and each wealth quintile we compute the quintile average for x in 2004 and the average x for the *same* households[av] in 2006 and then report the annualized percentage difference between the two figures. For the expenditure rates, which are already in percentage units, we compute the annualized percentage point differences between 2004 and 2006.

For net worth, the model captures the fact that in good economic times, wealth-poor households accumulate wealth at a faster rate than wealth-poor households. The percentage increase in wealth for all groups is lower in the model than in the data. We should note that in the data, the 2004–06 period was one of rapid appreciation of house prices and financial asset valuations, whereas in our model the relative price of wealth (capital) is constant at one, and thus an increase in net worth during normal times in the model has to come from net capital accumulation of households.[aw]

In terms of earnings (not reported) and disposable income, the model displays the substantial mean reversion built into the estimates of the idiosyncratic unemployment and earnings process, with income of the lowest wealth quintile rising fast (7.2%) and income of the highest wealth group actually falling (by 1%) even though aggregate incomes do not. This is because low wealth households tend to be low labor earnings and thus low income households with income. As we saw earlier, this is qualitatively consistent with the data, but quantitatively the model implies differences in income growth between the top and the bottom of the wealth distribution that are too large. In other words, the

[au] Since in Tables 9 and 10 the statistics for earnings and disposable income are quite similar, we only report those for disposable income.

[av] These households would typically not be in the same wealth quintile in 2006 as they were in 2004.

[aw] In a model without retirement and thus without life-cycle saving, generating positive changes in net worth for *all* wealth quintiles is of course very difficult; justifying again the inclusion of a basic life cycle element into the economy.

Table 10 Annualized changes in selected variables by net worth in a severe recession: Data vs model

	Net worth (%)		**Disp. *Y* (%)**		**Expend. (%)**		**Exp. rate (pp)**	
NW Q	**Data**	**Model**	**Data**	**Model**	**Data**	**Model**	**Data**	**Model**
Q1	NaN	24	6.7	4.9	0.6	4.5	−4.2	−0.4
Q2	24	15	4.1	0.3	2.0	1.2	−1.3	0.8
Q3	4	8	1.8	−2.4	0.8	0.0	−1.1	2.2
Q4	2	4	1.7	−4.0	−1.7	−1.5	−2.0	3.2
Q5	−5	−1	−1.2	−6.4	−3.7	−3.5	−1.4	4.6
All	−3	1	1.2	−3.7	−1.3	−0.8	−1.6	2.0

model implies slightly too much downward and upward mobility in incomes when households are ranked by wealth.[ax]

Finally, for changes in consumption expenditures, Table 10 reveals that during normal times, as in the data (and as for disposable income), consumption growth is strongest at the low end of the wealth distribution. The wealth gradient of the consumption growth rates (again, as for disposable income), is somewhat steeper in the model. As in the data, the expansion of consumption for households in the lowest (in 2004) wealth quintile falls short of their income growth and thus the expenditure rate of this group falls during normal times. The opposite is true for the wealthiest group of households in the population: as in the data, the expenditure rate of this group expands as the macro economy remains in normal times. The reason for this differential behavior in expenditure rates between the wealth-poor and the wealth-rich is intuitive from the perspective of the model: low wealth households have had, on average, unfortunate earnings realizations and their wealth is below their target wealth. Therefore, these households cut their expenditure to re-build their wealth buffers. The opposite logic applies to households at the top of the wealth distribution. This implication of the model matches the data, although quantitatively, the difference in changes in expenditure rates between the top and the bottom wealth quintiles is a bit larger in the model than in the data.

5.3.2 A Great Recession

After documenting the dynamics of wealth, income, and consumption (ordered by wealth) in normal times, Table 10 displays the same *model* statistics during a period in which the macro economy undergoes a large recession, induced by a transition of aggregate TFP from $Z = Z_h$ to $Z = Z_l$.[ay] To facilitate comparisons between the two tables, we

[ax] Ranking households by earnings or income would make this statement even stronger.

[ay] In the model the Great Recession hits in Q.I, 2009, consistent with our calibration. In that quarter, Z switches from $Z = Z_h$ to $Z = Z_l$ and remains there until Q.III, 2013. The statistics are based on comparing the average of the four 2010 quarters to the average of the four 2008 quarters. In the data, as discussed in Section 2, we consider the period from 2006 to 2010 because of the timing of the income and consumption data. Note that in the data, changes are all annualized.

Table 11 Difference in annualized growth rates between recession period and normal times: Data and Model

	Net worth (%)		Disp. *Y* (%)		Expend. (%)		Exp. rate (pp)	
NW Q	Data	Model	Data	Model	Data	Model	Data	Model
Q1	NaN	−20	−0.7	−2.3	−6.5	−2.2	−4.0	0.0
Q2	−98	−18	−2.6	−2.8	−5.2	−2.4	−1.6	0.3
Q3	−29	−12	−3.3	−4.0	−9.0	−2.7	−3.4	1.4
Q4	−15	−5	−3.3	−4.5	−7.4	−2.8	−2.5	2.0
Q5	−17	−4	−3.0	−5.4	−6.2	−2.9	−1.9	3.2
All	−19	−4	−2.9	−4.4	−6.9	−2.6	−2.5	1.3

display the difference in the growth rates between the recession period and normal times in Table 11.

Again, first focusing on net worth, the key endogenous state variable in our model that underlies the dynamics of all other economic variables, we observe that as in normal times (as in the data), the growth rate of net worth is declining in the level of net worth. And as in the data, the Great Recession significantly slows down the pace of wealth accumulation across all quintiles, and turns it negative for the wealthiest households, although the reduction predicted by the model is smaller than in the data. In the model, the wealth of the top net worth quintile declines by 1%, relative to the 3% growth in normal times. For the same quintile, annual wealth growth in the data slows down from 12% to −5% over a two-year period. As discussed above, in the data a large part of this reduction in wealth at the top of the distribution is likely the consequence of asset *price* movements which are, by construction, absent in the one-asset model studied here.[az]

The two other empirical facts we have documented in Section 2.3 were that income declines in the recession hit the top wealth quintiles more than the bottom quintiles, and that households in the bottom quintiles cut expenditure rates more than households in the top quintiles. Comparing disposable income growth rates in Tables 9 and 10, we observe that the first fact is captured well by the model, at least qualitatively. In the model, the decline in the income growth rate is 2.3 percentage points for the lowest wealth quintile, but 5.4 percentage points for the highest wealth quintile (and the decline is monotonically increasing in wealth in between these two extreme wealth quintiles). In the data, the wealth-poorest 20% of the working-age population see their income growth rate slow down by 0.7 percentage point, whereas for the wealthiest households, income growth slows down by 3.0 percentage points.

In contrast, the performance of the model with respect to the changes in consumption rates is more mixed. In the model, in the recession households all increase consumption

[az] Huo and Rios-Rull (2016) and Kaplan et al. (2016a) investigate the role of price movements in housing in explaining aggregate consumption dynamics in the Great Recession.

by more, or cut consumption by less, than disposable income, resulting in a rise in consumption rates, with the increase in consumption rates being smallest at the low end of the wealth distribution. In the data, all groups instead cut their consumption rates, the more so the less wealthy they are. Thus, although the model is consistent with the relative movement (in the recession vis-á-vis normal times) in consumption rates across wealth levels, with the wealth-poor decreasing consumption rates the most—in the data—or increasing them the least—in the model, the latter overstates consumption growth in the recession and thus underpredicts the decline in expenditure rates evident in the data.

In the model, when the recession hits and thus incomes decline (or grow less) relative to normal times, households have strong incentives to use their wealth to smooth consumption. This is especially true for those falling into unemployment. On the other hand, since the recession is long-lasting and comes with elevated unemployment risk, the motive to engage in precautionary saving against future unemployment spells increases, especially among those with little wealth coming into the recession. For high wealth households, the first motive dominates and the consumption rates of these households increase in the recession, whereas for low-wealth households both motives roughly balance out, leaving consumption rates roughly unchanged across the two time periods. We will show below that in an economy with less generous unemployment insurance, the precautionary savings motive becomes more potent, especially at the low end of the wealth distribution, and low-wealth households indeed cut their consumption rates during recessions, as is the case in the data.

We conclude this section by briefly summarizing the strengths and shortcomings of our baseline model when confronted with the PSID earnings, income, consumption, and wealth data. The model succeeds in replicating the observed cross-sectional wealth distribution (except at the very top) and does well in capturing the salient features of the joint distribution of wealth, income, and expenditures. It also replicates the relative movements of expenditure rates by wealth as the economy falls into a recession. However, it fails to predict the *decline* in consumption expenditure rates during recessions and fails to capture the large movements in wealth we see in the data during the years 2006–10, since it abstracts from asset price movements.

In the next section, we use the benchmark model and some of its variants to quantify the extent to which wealth inequality is important in determining the magnitude of aggregate consumption movements in response to a Great Recession type business cycle shock in TFP.

6. CROSS-SECTIONAL HOUSEHOLD HETEROGENEITY AND THE AGGREGATE DYNAMICS OF CONSUMPTION AND INVESTMENT IN A SEVERE CRISIS

In this section, we argue that the cross-sectional distribution of households across individual characteristics (primarily in wealth and impatience) is a crucial determinant of the aggregate consumption and investment response to a negative business cycle shock. In addition, we show that in the presence of such significant household heterogeneity, the generosity of social insurance policies strongly affects the dynamics of macroeconomic aggregates.

Our focus on the impact of household heterogeneity in wealth for the aggregate consumption dynamics during large recession is shared with a number of recent studies, including Guerrieri and Lorenzoni (2012), Glover et al. (2014), Heathcote and Perri (2015) as well as Berger and Vavra (2015).

When exploring the role that social insurance policies can play in shaping the aggregate consumption (and, in the next section, output) response to adverse business cycle shocks in economies with household heterogeneity we build, on the work by Krusell and Smith (2006), which also focuses on income insurance programs, and more concretely, unemployment insurance.[ba] Our work is also related to McKay and Reis (2016), who conduct a comprehensive study of automatic stabilization programs on business cycle dynamics, to Heathcote (2005), Kaplan and Violante (2014), and Jappelli and Pistaferri (2014), who study the role of discretionary changes in income taxation on aggregate consumption, and Brinca et al. (2016), who investigate the magnitude of aggregate fiscal multipliers in this class of heterogeneous agent models.

6.1 Benchmark Results

We consider two thought experiments, both of which take as an initial condition the wealth distribution after a long sequence of good shocks so that the cross-sectional distribution has settled down. Then a severe recession hits. In the first thought experiment, productivity returns to the normal state $Z = Z_h$ after one quarter (and remains there forever after). Although this thought experiment is not a good depiction of the actual Great Recession because of the short duration of the downturn, it displays the mechanics of the model recession most clearly.[bb] In the second thought experiment, we plot the responses of the economy to a Great Recession of typical length (according to our calibration) that lasts for 5.5 years (22 quarters). In both cases we trace out the impulse response functions (henceforth IRF) for the key macroeconomic aggregates. The main focus of interest is on the extent to which the aggregate consumption and investment responses differ across two economies that differ fundamentally in their extent of household heterogeneity.

To make our main point, we perform both experiments for two model economies: the original Krusell–Smith economy without preference heterogeneity, life-cycle structure, and only modest unemployment insurance, and our benchmark model that includes

[ba] As we do, Auclert (2014), Auclert and Rognlie (2016), and Kekre (2015) also stress the importance of the heterogeneity in the marginal propensity to consume across households for the dynamics of aggregate demand and the impact of redistributive policies. Wong (2015) stresses the heterogeneity in age across households for the transmission of monetary policy shocks to aggregate consumption.

[bb] Of course, households form expectations and make decisions based on the persistent Markov chain for Z driving the model even in this thought experiment.

these features and therefore, as documented above, provides a model wealth distribution that matches its empirical counterpart very well. We will also show that the aggregate consumption and investment behavior over the business cycle in the KS economy approximates an economy with representative agents (RA) very well (as already noted in the original Krusell and Smith (1998) paper), and thus as far as macroeconomic aggregates are concerned, the KS and the RA economy can be treated as quantitatively equivalent.

In Fig. 2, we plot the model impulse response to a onetime negative technology shock in which Z switches to Z_l after a long spell of good realizations Z_h. The upper left panel plots the time series of TFP Z fed into the model, and the remaining sub-plots show the model-implied dynamics of aggregate consumption, investment, and output induced by the Great Recession type TFP shock. By construction the time paths of exogenous TFP Z are identical in both economies in the short run; for output they are identical on impact and virtually identical over time. Since TFP and labor supply are exogenous in both

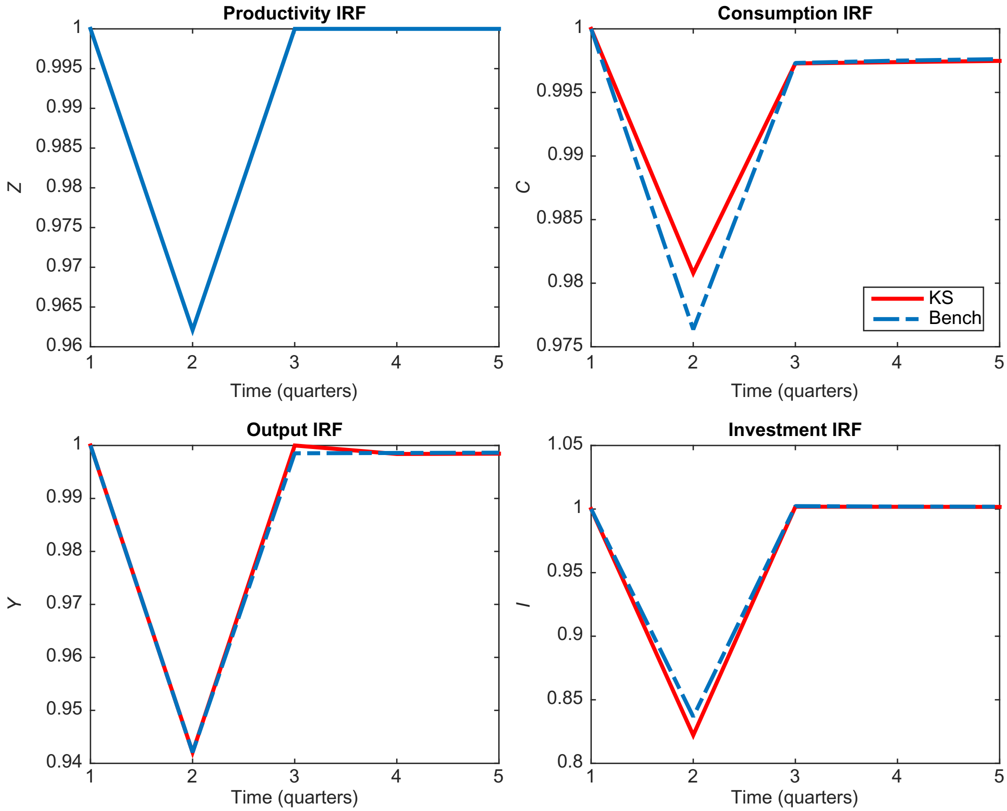


Fig. 2 Impulse response to aggregate technology shock in two economies: One time technology shock.

economies and follow the same time path, capital is predetermined on impact, and the one time shock is not sufficient to trigger a substantially different dynamics of the capital stock, the time path of output is virtually identical in both economies. Thus, the key distinction between both economies is the extent to which a very similar decline and recovery in output is reflected in lower aggregate consumption rather than aggregate investment.

The key observation we want to highlight is that the aggregate consumption (and thus investment) response to the negative productivity shock differs substantially between the two economies. In the benchmark model, consumption falls by 2.4% in response to a technology shock that induces a decline in output by 6% on impact. The same fall in output triggers a decline of only 1.9% in the original Krusell–Smith (labeled as KS) economy. Thus the impact of the recession on aggregate consumption increases by 0.5% percentage points more in the economy with empirically plausible wealth heterogeneity. Given that output is exogenous in the short run, and is used for consumption and investment only in this closed economy, the investment impulse response necessarily shows the reverse pattern: the decline in investment is much weaker in the high wealth inequality economy. This in turn triggers a less significant decline and more rapid recovery of the macro economy once the recession has ended. However, given that new investment is only a small fraction of the capital stock, these differential effects on capital, and thus output, are quantitatively minor, at least in the case in which the recession is short-lived.[bc]

Note that for all practical purposes, in what follows the KS economy displays aggregate consumption-investment dynamics that are very close to those in a representative agent (RA) economy. Fig. 3 shows this fact by displaying impulse responses to a one-period recession shock in the KS and RA economies. Although not identical, the impulse responses are quantitatively very close. For example, the aggregate consumption decline in the RA economy amounts to 1.78%, relative to a fall in aggregate consumption of 1.9% in the KS economy.

In Fig. 4, we display the dynamics of macroeconomic aggregates in a prolonged and severe recession, with a length of 22 quarters, under our operational definition of a severe recession. It demonstrates that in a Great Recession lasting several years, the differences in capital and output dynamics across the low-wealth inequality KS economy and the high inequality benchmark are now more noticeable, especially toward the end of the recession. As a result, the recovery after TFP has turned back up again is substantially stronger in the benchmark economy, by approximately 1 percentage point for capital and 0.3 percentage point for output in the period in which the recession ends.

[bc] In Section C.4, we argue that the fact that the wealth distribution is quantitatively important for the current aggregate consumption response to a TFP shock does not imply that higher moments of the wealth distribution are needed to accurately forecast *future* wages and interest rates.

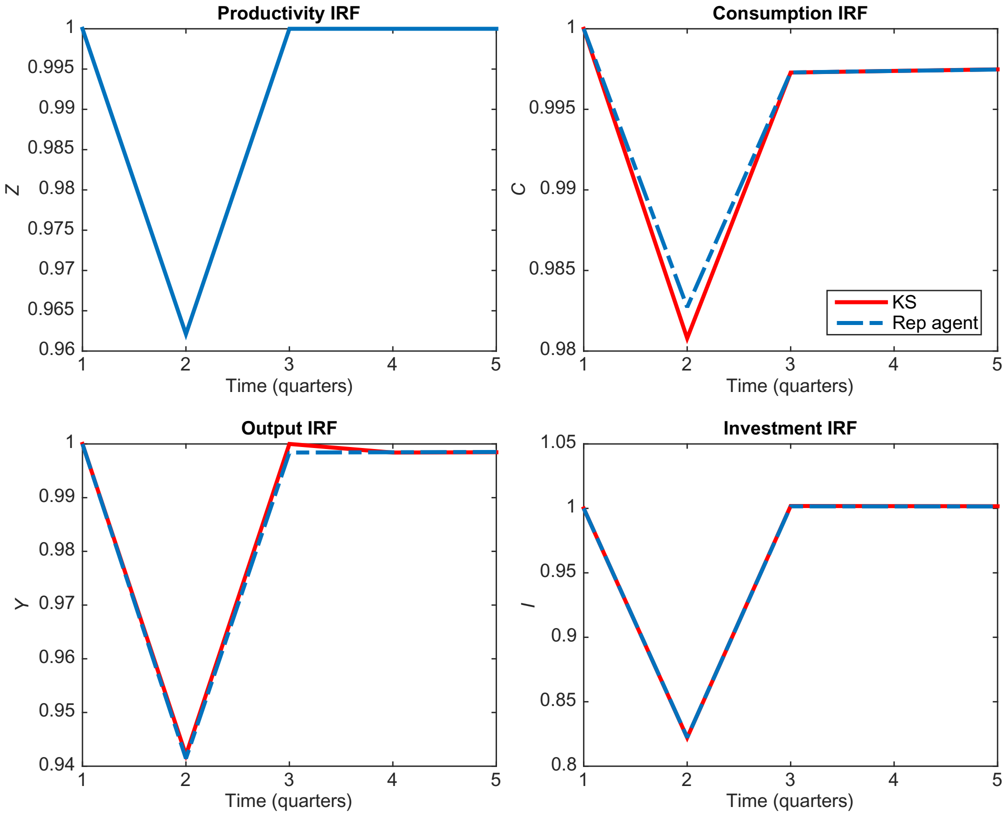


Fig. 3 Impulse response functions (IRF) to aggregate technology shock in KS and RA economies.

Since the KS economy and the benchmark differ along several model dimensions, in the next section we break down the reasons for the differential aggregate consumption response, again focusing on the interaction between the aggregate movement in consumption in a Great Recession and the cross-sectional wealth distribution prior to it.

6.2 Inspecting the Mechanism II: What Accounts for the Size of the Aggregate Consumption Recession

The key finding from the last section is that the aggregate consumption recession in our benchmark economy with preference and realistic wealth heterogeneity is more than twice as deep as it is in the corresponding RA economy (which in turn displays aggregate time series that are very close to those in the original KS economy). In this section, we dissect the reasons behind this finding. To start, in Fig. 5, we display the consumption functions and wealth distributions for both the KS and the benchmark economy. The left panel shows the consumption functions (plotted against individual wealth on the *x*-axis) in the original KS economy for three combinations of idiosyncratic employment

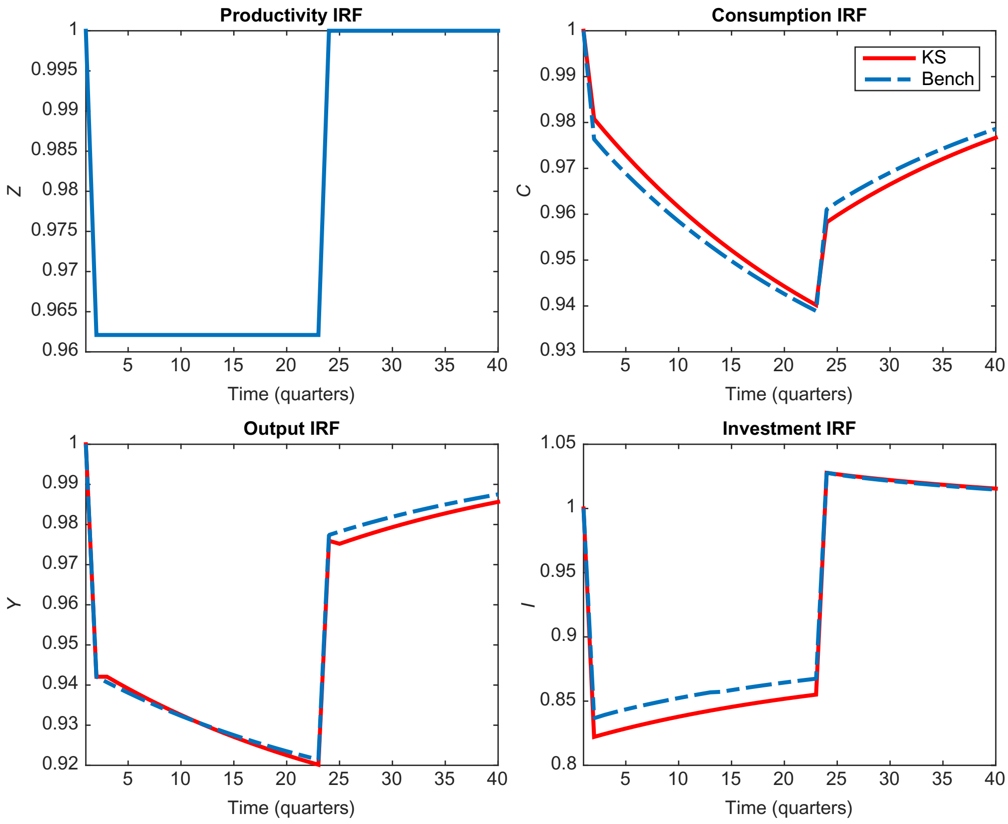


Fig. 4 Impulse response to aggregate technology shock in two economies: "Typical" severe recession technology shock.

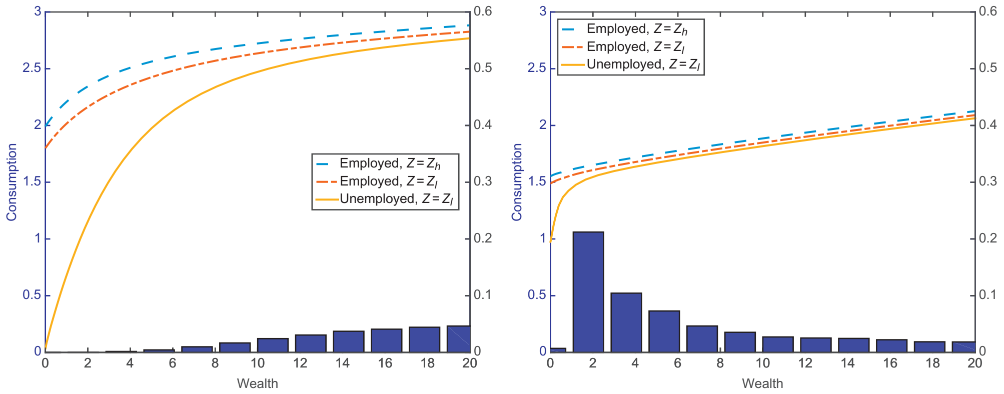


Fig. 5 Consumption function and wealth distribution: Krusell–Smith (left panel) and benchmark (right panel).

and aggregate productivity states. For a given wealth level, the vertical difference between the consumption functions for the employed in aggregate state $Z = Z_h$ (blue dashed line) and the employed in aggregate state $Z = Z_l$ (red dot-dashed line) gives the consumption drop in the Great Recession, conditional on not losing a job. In the same way, the vertical distance between the blue-dashed consumption function and the orange solid consumption function (for the unemployed in the recession) gives the consumption decline for those households that lose their jobs in a recession. The figure also contains the pre-recession wealth distribution, displayed as a histogram, with the mass of a particular wealth bin being measured on the right y-axis.[bd] The right panel displays the same information, but for our benchmark economy, for working-age households with median earnings state y and mean discount factor $\overline{\beta}$.

The first observation we make is that, for a given level of wealth, the drop in individual consumption as the KS economy falls into a Great Recession is substantially *larger* than in our benchmark economy.[be] This is especially true for households with little wealth that lose their jobs at the onset of the recession, because of the virtual absence of unemployment insurance.

The observation of larger individual consumption declines in the KS economy would suggest that the aggregate consumption recession is actually larger than it is in the benchmark economy, in contrast to the result documented in the previous section. However, as Fig. 5 (and Table 6) display clearly, the cross-sectional wealth distribution places almost no mass on households with very little net worth, exactly the households with the largest consumption declines. In contrast, the benchmark model with realistic wealth inequality places substantial probability mass at zero or close to zero wealth where the individual consumption losses are significant, especially (but not only) for newly unemployed households.[bf] Note that average net worth is the same in both economies: we truncate the plots at net worth twenty times average income in order to make the individual consumption declines at the low end more clearly visible, but the benchmark economy has a fat right-tailed wealth distribution that is well approximated by a Pareto distribution (as in the data, see eg, Benhabib and Bisin, 2016), whereas the original KS economy displays a wealth distribution whose right tail more closely resembles that of a log-normal distribution. Thus, both distributions have the same mean even though, as clearly visible from the figure, the benchmark economy has substantially more mass of households at low levels of net worth.

As we will see in Section 6.3, public social insurance programs will affect both the determinants of the aggregate consumption dynamics—the consumption response to

[bd] The aggregate capital stock associated with these plots is the prerecession capital stock; note that both economies, by virtue of the calibration, have the same average (over the cycle) capital stock.

[be] Fig. 5 displays the consumption functions in the benchmark economy for individuals with median (y, β), but the same statement applies, qualitatively, to the consumption functions for households with other (y, β) characteristics. Recall that there is no (y, β) heterogeneity in the original KS economy.

[bf] The wealth distribution in the right panel of Fig. 5 is for the entire working-age population, rather than conditioning on the specific (y, β) types for which the consumption functions are displayed.

aggregate shocks for a given wealth level—and the wealth distribution itself. Both components are crucial when determining the overall impact of unemployment insurance policies on the macro economy over the business cycle. Before turning to this point, we first further explore the precise reasons behind the significant differences in aggregate and distributional characteristics between the original KS economy and our benchmark, thereby pinpointing precisely which model elements (and their interaction) are responsible for the differences in aggregate consumption dynamics across different economies.

Recall that relative to the KS model, our benchmark includes idiosyncratic earnings shocks, a rudimentary life cycle structure with social security system, permanent preference heterogeneity as well as a more generous unemployment insurance system.

In Table 12, we repeat the information from Table 7 on the wealth distribution in different versions of the model, but now we also document the magnitude of the aggregate consumption response on impact in a Great Recession. Fig. 6 displays the associated impulse responses. From the table and figure we observe that the introduction of persistent idiosyncratic income risk on top of unemployment risk significantly amplifies the aggregate consumption response above that of the original KS model. In fact, the magnitude of the aggregate consumption response is larger than that obtained in the benchmark (the second to last column in the table). This is perhaps not surprising given our arguments thus far, as this version of the model generates significantly larger wealth inequality—and importantly—the two lowest wealth quintiles that hold very little net worth.[bg]

Table 12 Net worth distributions and consumption decline: Different versions of the model

	Models*					
% Share:	KS	$+\sigma(y)$	+Ret.	$+\sigma(\beta)$	+UI	KS + Top 1%
$Q1$	6.9	0.7	0.7	0.7	0.3	5.0
$Q2$	11.7	2.2	2.4	2.0	1.2	8.6
$Q3$	16.0	6.1	6.7	5.3	4.7	11.9
$Q4$	22.3	17.8	19.0	15.9	16.0	16.5
$Q5$	43.0	73.3	71.1	76.1	77.8	57.9
90–95	10.5	17.5	17.1	17.5	17.9	7.4
95–99	11.8	23.7	22.6	25.4	26.0	8.8
$T1\%$	5.0	11.2	10.7	13.9	14.2	30.4
Wealth Gini	0.350	0.699	0.703	0.745	0.767	0.525
ΔC	−1.9%	−2.5%	−2.6%	−2.9%	−2.4%	−2.0%

*The KS model only has unemployment risk and incomplete markets, and thus the first column repeats information from table 6. The column $+\sigma(y)$ adds idiosyncratic earnings shocks (transitory and permanent) while employed. The column +Ret. adds the basic life cycle structure (positive probability of retirement and positive probability of death, plus social security in retirement). The column $+\sigma(\beta)$ incorporates preference heterogeneity into the model, and finally the column +UI raises the replacement of the unemployment insurance system from 1% to 50%; the resulting model is therefore the benchmark model, with results already documented in table 6. In all models, the (mean) discount factor is calibrated so that all versions have the same capital-output ratio.

[bg] Note, however, that this mechanism is insufficient to generate the very high wealth concentration, as the examination of the wealth share very top of the wealth distribution reveals.

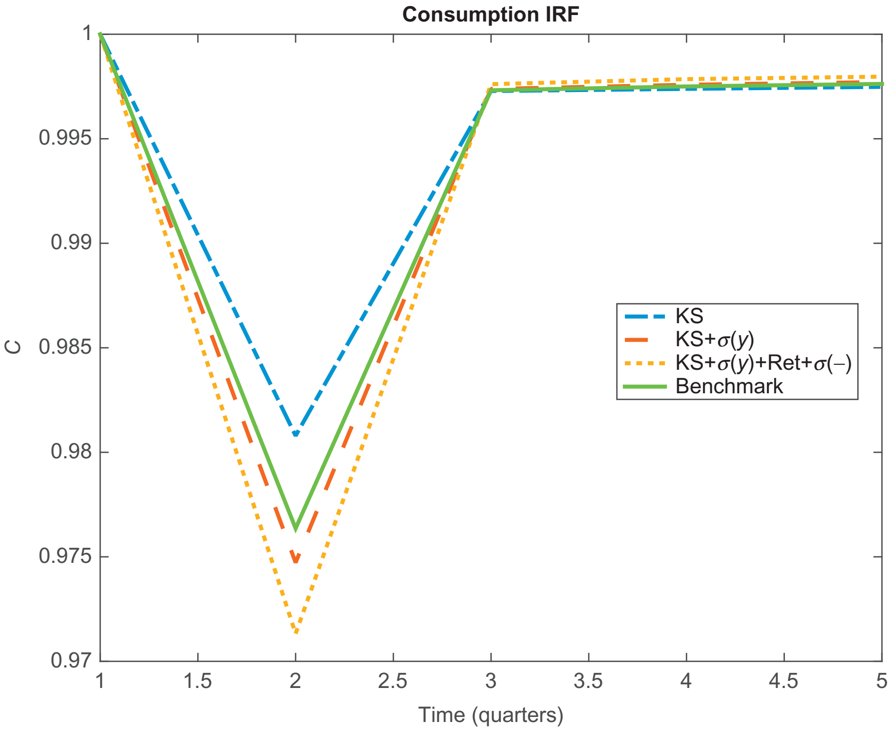


Fig. 6 Consumption recessions in various versions of the model.

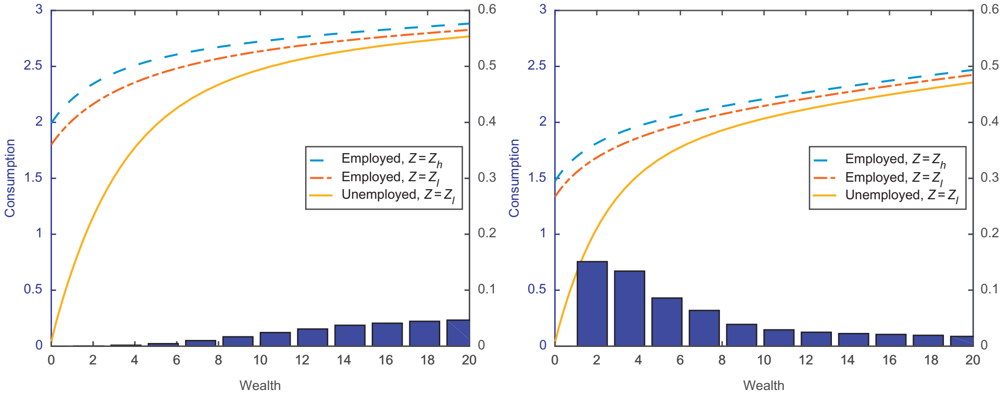


Fig. 7 Consumption function and wealth distribution: KS (left panel) and KS w/income risk (right panel).

Fig. 7 compares the consumption functions and equilibrium wealth distributions in the KS economy and the KS economy with just persistent earnings shocks added. In the latter, the policy functions are displayed for the median y realization. Whereas the consumption policy functions look broadly similar in both economies, the mass of households with low wealth and thus a large consumption response to the recession shock increases very substantially relative to the original KS economy. In this variant, the wealth distribution at the *bottom* looks already quite similar to the benchmark economy, although the absence of significant unemployment insurance implies that the mass of households at exactly zero wealth is negligible. On the other hand, because of the absence of unemployment insurance, the consumption drop of the wealth-poor for a given wealth level is comparable in magnitude to that in the original KS economy.

Fig. 8, which displays the consumption functions and wealth distributions for two different types households in the $KS + \sigma(\gamma)$ economy, clarifies the interaction between earnings inequality and wealth inequality. Households with low current (and very persistent) income realizations are highly concentrated at the low end of the wealth distribution. But even among households with contemporaneous median income, there is significantly more mass in the wealth region where consumption falls substantially upon unemployment.

Moving to the third column of Table 12, we see that although the introduction of life-cycle elements is crucial for delivering joint income-consumption distributions, their impact on the dynamics of aggregate consumption in the recession is limited. In contrast, adding preference heterogeneity to the model helps to amplify the consumption drop. Crucially, now the economy is populated by a share of highly impatient households at the bottom of the wealth distribution. In normal times, unemployment risk is low and these households consume at a high rate because of their impatience, ending up with little or no wealth. When the economy falls into the recession, idiosyncratic unemployment risk goes up significantly for the "foreseeable future" from the point of view of impatient households. Faced with the elevated chance of becoming unemployed,

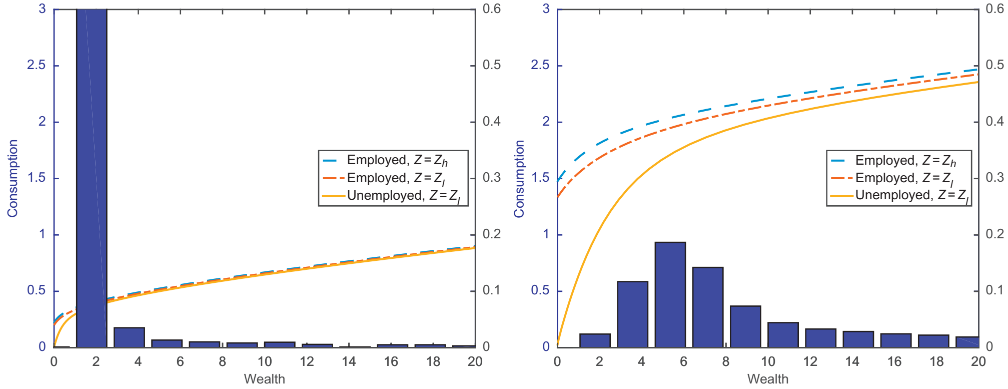


Fig. 8 Consumption function and wealth distribution: KS low income (left panel) and KS median income (right panel).

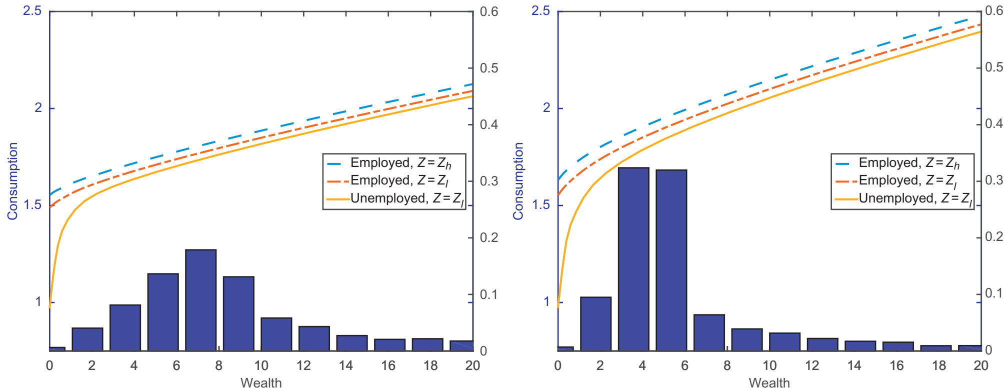


Fig. 9 Consumption function and wealth distribution: Patient households (left panel) and impatient households (right panel).

impatient households who have not yet lost their jobs and have currently medium to high income realizations start to save more for precautionary reasons.[bh]

For more patient employed households, the increase in precautionary saving and resulting drop in consumption at the onset of the recession is not quite as severe. These households were already saving a larger fraction of their income even in good times, since their patience makes them more focused on the long horizon. Because the persistent idiosyncratic income component is more persistent than the recession, patient households with high current income expect to have high income even when exiting the recession, so the short-run possibility of increased unemployment is not as big of a concern to them.

Fig. 9 displays the consumption policy functions for patient and impatient households, as well as the wealth distribution among these households. The key observation is that consumption falls more pronouncedly for impatient households when the aggregate state turns bad, even conditional on *not* losing a job. Also, not unexpectedly, among impatient households wealth levels tend to be lower, as the group-specific wealth distributions underneath the consumption functions in Fig. 9 show. As a broad summary measure of this differential effect, the contribution to the aggregate decline in consumption is more than twice as large for the most impatient group of households than for the most patient group, even though that they constitute equal shares of the population.

In the aggregate, the decline in aggregate consumption in the economy with income and preference heterogeneity amounts to 2.9%, and is thus a full 1 percentage point larger than in the KS economy, and 1.11% larger than in the representative agent economy. Both dimensions of heterogeneity are quantitatively important for the magnitude of the aggregate fluctuations, and so is their interaction, as the previous discussion of the importance of the impatient, employed with high income has indicated.

[bh] The small share of impatient, low-wealth households that do in fact lose their jobs at the onset of the recession behave like hand-to-mouth consumers instead, cutting their consumption one for one with income, and consume whatever little wealth they might have at the beginning of the recession.

Finally, the second to last column of Table 12 raises the unemployment insurance replacement rate to our benchmark value of 50%. As we discuss and quantify in the next part of the chapter, Section 6.3, this change in the generosity of social insurance has a two-fold impact on the economy: for a given wealth level it softens the decline in household consumption in the recession, but it also shifts the wealth distribution toward wealth levels that imply a large decline in consumption and thus make the recession more costly in welfare terms. The first effect reduces the aggregate consumption response to the Great Recession shock, the second magnifies it. As Table 12 shows, the net effect is a reduction of aggregate consumption volatility (with a decline of 2.4%), bringing the implications of the benchmark economy closer to that of the RA and KS economies with absent or limited wealth heterogeneity.

To summarize the main lessons from this section, the key aspects of the benchmark model that make its implied consumption dynamics different from its RA counterpart in a quantitatively meaningful way are (a) an equilibrium wealth distribution that makes the wealth-poor poor enough and has them cut consumption more significantly than the average household when the recession hits; and (b) that these wealth-poor households make up a significant share of aggregate consumption. These requirements are achieved through highly persistent income shocks that generate a set of households that are born wealth-poor and never accumulate much wealth, and are compounded by the presence of impatient households that do not want to accumulate much wealth. If these households do not have access to generous unemployment insurance, their consumption falls a lot more than that of the representative household in a recession, either because they have in fact lost their jobs (and the incidence of job loss is higher in recessions), or because they have not lost their job, but have cut consumption to hedge against a now more likely job loss in the future.

Preference heterogeneity produces not only impatient households with the characteristics discussed thus far, but also patient households that find it optimal to accumulate large amounts of wealth, thereby contributing significantly to wealth inequality. However, it is the lack of wealth at the bottom, as opposed to significant concentration at the very top, that is crucial for explaining aggregate consumption dynamics. To make this point sharply, we consider a version of the model that is identical to the original KS model but adds limited preference heterogeneity. Specifically, it constructs a model in which 99% of the population has a lower time discount factor β_l than the remaining 1% of the population. The two discount factors are chosen to match the capital-output ratio in the benchmark economy (which essentially pins down β_l) and the share of wealth held by the top 1%–30%–as in the PSID data (whereas in the benchmark economy, we match the capital-output ratio and the wealth Gini). This pins down the time discount factor β_h of the remaining 1% of the population.

The purpose of this economy is to evaluate the importance of the wealth concentration at the very top of the distribution for the aggregate consumption decline in a Great Recession (and to demonstrate that it is straightforward, with appropriate preference heterogeneity in time discount factors, to generate a wealth distribution as concentrated at the top as it is in the data). The wealth distribution and aggregate consumption decline from this version of the model are reported in the last column of Table 12. Since consumption functions are approximately linear for households with above-median wealth, and the individual consumption *drop* in a recession is roughly invariant to net worth at that level, it does not matter much *for aggregate consumption dynamics* if the top of the wealth distribution is populated by 1% of astronomically wealthy households, or by 20% of merely super rich households. Consequently, the consumption response is roughly the same in this variant of the model and in the original KS economy (and the RA economy for that matter).

6.2.1 The Importance of Precautionary Saving vs "Hand-to-Mouth" Consumers

Given the importance we assigned to households with *little net worth* in our discussion above, in this section we briefly ask whether a model with a *fixed* fraction of households κ that always have zero wealth and thus simply consume their income in every period has the same implications for the consumption dynamics as our benchmark model.[bi]

We have resolved our model under the assumption that the bottom $\kappa = 40\%$ of the wealth distribution in model period $t-1$ just consumes their earnings and unemployment benefits (if applicable) from period t on, whereas the remainder of the distribution (in period $t-1$) continues to follow the intertemporally optimal decision rules from the benchmark economy.

The drop of consumption in a one-period Great Recession now amounts to 2%, relative to the decline in the benchmark economy of 2.4%. The drop is larger in the benchmark economy since households at the bottom of the wealth distribution on average (and especially those not currently unemployed) find it optimal to *reduce* consumption rates for precautionary reasons: the Great Recession is expected to last a long time, and those not yet affected by a job loss try to build a buffer to hedge against the increased risk of being laid off in the future.[bj] This precautionary saving motive in the face of increased idiosyncratic risk in recessions, also discussed lucidly in a recent paper by McKay (2015), is absent among households that follow a mechanical hand-to-mouth consumption rule and is

[bi] This question is interesting from a modeling perspective since a model in which a fixed fraction of hand-to-mouth households and the remaining fraction employs permanent income consumption and savings functions (which are linear in wealth with identical marginal propensities to consume out of wealth, given our model) would give rise to easy aggregation.

[bj] In the versions of the model studied here, labor supply is exogenous (but its productivity fluctuating over the cycle), and thus saving is the only possible household response to hedge against higher idiosyncratic risk. In models with endogenous labor supply choice, such as the ones studied in Chang and Kim (2007) and Athreya et al. (2015), households have another margin of adjustment and thus the impact of elevated risk on precautionary saving will be smaller. For a model that combines household precautionary saving and *frictional* labor markets, see Krusell et al. (2010).

responsible for the deeper recession in the benchmark economy.[bk] We will return to this point in the next section, where we study the impact of the generosity of unemployment insurance on our results, and will show that with less generous unemployment insurance benefits, the additional precautionary savings motive from elevated unemployment risk is more potent, and the divergence between the class of models studied here and hand-to-mouth consumer models is even more significant.

It is important to note that in our formulation the share of households that behave as hand-to-mouth consumers is exogenous. In recent work Kaplan and Violante (2014) and Bayer et al. (2015) construct models with wealthy hand-to-mouth consumers where a share of households endogenously choose to behave like hand to mouth consumers despite having non-trivial net worth. However, since their net worth is primarily in the form of assets that are costly to liquidate (think of owner-occupied real estate and tax-favored retirement accounts), the consumption behavior of this group of households approximates that of the hand-to-mouth consumers modeled here, especially for income shocks of moderate magnitude.

6.3 The Impact of Social Insurance Policies

In this section, we ask how the presence of public social insurance programs affects the response of the macro economy to aggregate shocks in a world with household heterogeneity.[bl] We focus specifically on the effects of government-provided, and tax-financed unemployment insurance. We will argue that the impact of this policy is two-fold: it changes the consumption-savings response of a household with a *given* wealth level to income shocks, and it changes the cross-sectional wealth distribution in society, at least in the medium to long run. In order to decompose the overall impact of social insurance into these two effects, we consider two thought experiments. In the first, we simply compare the dynamics of macroeconomic aggregates of the benchmark economy with that of an identical economy that has a lower unemployment insurance replacement rate of $\rho = 10\%$. We interpret the latter economy as providing basic social insurance (as embedded in basic welfare programs), or alternatively, as a world where a significant share of households do not claim unemployment benefits despite being entitled to it.[bm] This thought

[bk] Obviously, the magnitude of this effect depends on the share of hand-to-mouth consumers κ. In the limit, as $\kappa = 0$ we are back in the benchmark economy. For $\kappa = 20\%$ the fall in aggregate consumption is 2.1%, about halfway between the RA economy and the benchmark.

[bl] The purpose of this analysis is purely positive in nature, and limited in scope by the assumption that transitions between employment and unemployment are exogenous and thus policy-invariant. See Hagedorn et al. (2013) and Hagedorn et al. (2015) for an analysis of the effects of unemployment benefit extensions on vacancy creation and employment.

[bm] We prefer to model a replacement rate of $\rho = 10\%$ rather than $\rho = 1\%$ as in the original Krusell–Smith economy studied in the previous section, since we think $\rho = 10\%$ is a more empirically relevant case. The resulting macro effects will lie right in between that of the benchmark economy, and the economy with a replacement rate of $\rho = 1\%$ displayed in the forth column (the $\sigma(\beta)$ economy) of Table 12.

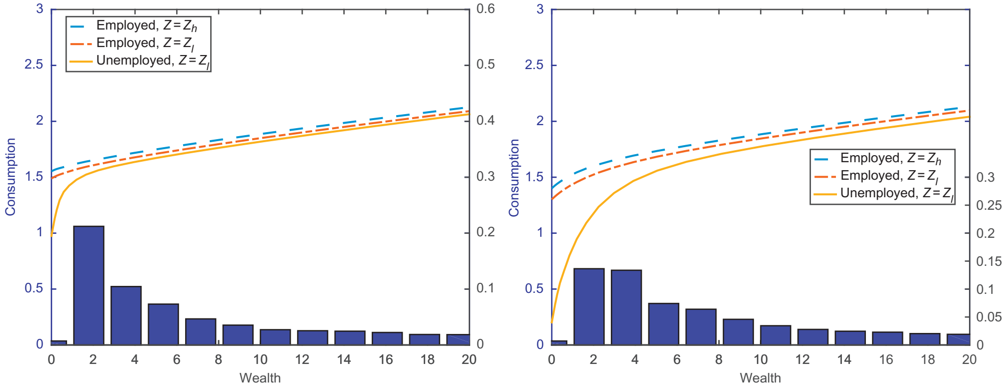


Fig. 10 Consumption function and wealth distribution: Benchmark (left panel) and low UI (right panel).

experiment will encompass the effect of unemployment insurance both on individual consumption behavior as well as on the equilibrium wealth distribution. To isolate the former effect, we will also consider an economy with low unemployment insurance, but entering the recession with the *same pre-recession wealth distribution* as in the benchmark economy.[bn]

In the left panel of Fig. 10, we plot, against wealth, the consumption functions (for the unemployed in the low and the employed in the high aggregate shock, with the mean discount factor) as well as the wealth histogram in the benchmark economy (with a replacement rate of 50%). This was the right panel of Fig. 5. The right panel of Fig. 10 does the same for an economy with an unemployment insurance system of only 10%. We chose to display the consumption function for the employed in an expansion and the unemployed in a recession because this helps us to best to understand what drives the aggregate consumption impulse response below.[bo]

We want to highlight three observations. First, in the high unemployment insurance economy, households with low wealth consume much more than in the economy with small unemployment insurance. Second, and relatedly, the decline in consumption for low-wealth households from experiencing a recession with job loss is much more severe in the low-benefit economy. Third, the size of the social insurance system however, by affecting the extent to which households engage in precautionary saving, is a crucial determinant of the equilibrium wealth distribution. In the benchmark economy (as in the data), a sizeable mass of households has little or no wealth, whereas in the no-benefit economy this share of the population declines notably. Specifically. average

[bn] One can interpret this thought experiment as a surprise permanent removal (or a surprise failure of extension) of unemployment benefits exactly in the period in which the recession hits.

[bo] Setting $\rho = 0$ would create the problem of zero consumption in some of the decomposition analyses we conduct later on.

assets increase by 0.5% relative to the benchmark economy, and only 0.9% of the population holds exactly zero assets, relative to 3.1% in the benchmark economy.

The difference in the consumption decline in a recession across the two economies can then be decomposed into the differential consumption response of households, integrated with respect to the *same* cross-sectional wealth distribution (which is a counterfactual distribution for one of the two economies), and the effect on the consumption response stemming from a policy-induced difference in the wealth distribution coming into the recession. As it turns out, both effects (the change in the consumption functions and the change in the wealth distribution) are quantitatively large, but partially offset each other.

In order to isolate the first effect, we now plot, in Fig. 11, the recession impulse response for the benchmark economy and the economy with low unemployment insurance, but starting at the *same pre-recession wealth distribution* as in the benchmark economy. Under this fixed wealth distribution scenario, the consumption response in both cases is

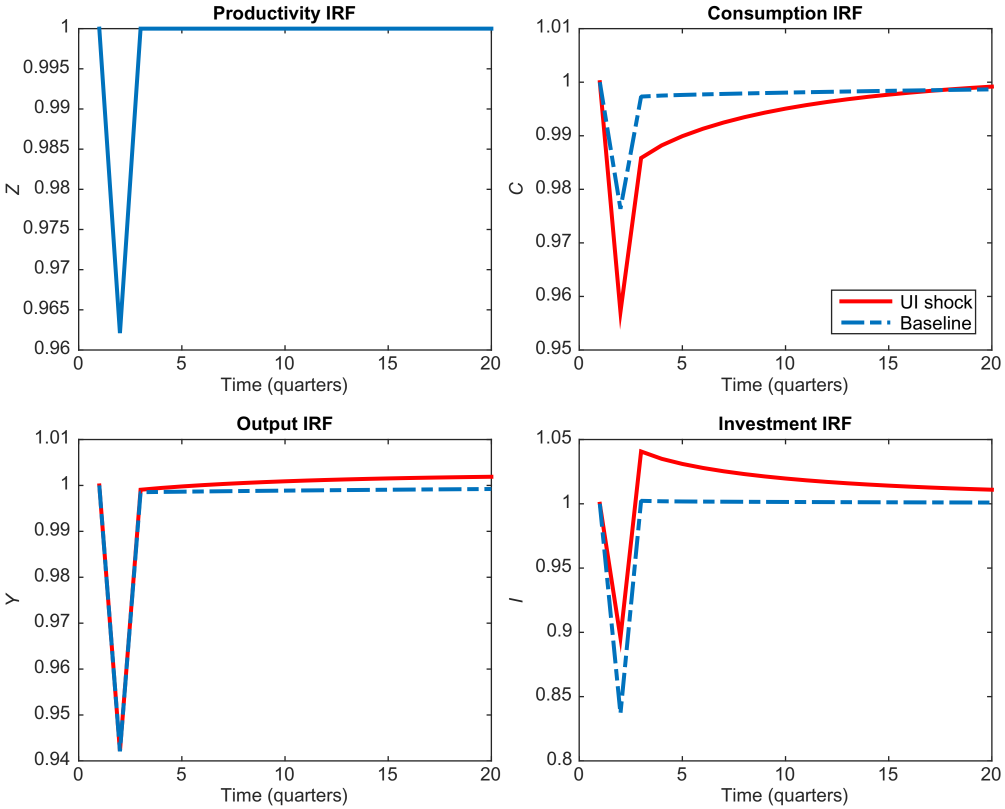


Fig. 11 Impulse response to aggregate technology shock with and without generous unemployment insurance, fixed wealth distribution: Onetime technology shock.

given by the difference in the consumption functions (in both panels) integrated with the wealth distribution of the high UE insurance economy. We find that consumption declines much more substantially in the economy with a low replacement rate, by 4.6%, relative to 2.4% in the benchmark economy. This is, of course, exactly what the consumption functions in Fig. 10 predict.

To further quantify what drives this differential magnitude in the consumption response, in Table 13, we display the fall in consumption for four groups in the population that differ in their transitions between idiosyncratic employment states as the aggregate economy slips into a recession. The share of households undergoing a specific transition is exogenous and the same across both economies, and is given in the second column of the table. Most (88.1%) households retain their jobs even though the aggregate economy turns bad. In contrast, the fraction of households making the transition from employment to unemployment is only 6.6% (and 3.5% of households make the reverse transition), but based on the consumption functions we expect them to display the largest decline in individual consumption.

The aggregate consumption decline documented in the last row of Table 13 corresponds to the impulse responses of Fig. 11. The rows above give the share of the consumption decline accounted for by each of the four groups, so that the sum of the rows adds up to 100%. Similarly, Table 14 summarizes the percentage consumption decline of each of the four groups and gives, in the second column, the prerecession population shares of each of these four groups.

Table 13 Consumption response by group in three economies: Share of total decline

Transitions	Pop. share	ρ = 50%, $\Phi^{\rho=0.5}$	ρ = 10%, $\Phi^{\rho=0.5}$	ρ = 10%, $\Phi^{\rho=0.1}$
$s=e, s'=e$	88.1%	79.8%	72.8%	71.6%
$s=e, s'=u$	6.6%	13.8%	18.5%	21.8%
$s=u, s'=e$	3.5%	2.5%	2.9%	0.3%
$s=u, s'=u$	1.8%	3.8%	5.8%	6.3%
Total decline	100%	−2.4%	−4.6%	−2.7%

Table 14 Consumption response by group in three economies: Consumption growth rates of different groups

Transitions	Pop. share	ρ = 50%, $\Phi^{\rho=0.5}$	ρ = 10%, $\Phi^{\rho=0.5}$	ρ = 10%, $\Phi^{\rho=0.1}$
$s=e, s'=e$	88.1%	−1.5%	−2.3%	−1.5%
$s=e, s'=u$	6.6%	−3.5%	−7.6%	−6.1%
$s=u, s'=e$	3.5%	−1.2%	−2.3%	−0.0%
$s=u, s'=u$	1.8%	−3.5%	−8.8%	−6.8%
Total decline	100%	−2.4%	−4.6%	−2.7%

From both tables we observe that, even though the share of households that become newly unemployed (6.6% of the population, $s=e, s'=u$) and remain unemployed (1.8% of the population, $s=u, s'=u$) is relatively small, these groups account for a disproportionately large fraction of the overall consumption collapse in both the economy with generous, and in the economy with modest unemployment insurance.[bp] See columns 3 and 4 of Table 13 (which are based on the same prerecession wealth distribution).

These two groups of households make up 8.4% of the population, but in the benchmark economy (column 3, Table 13) account for 17.6% of the consumption drop. Carrying out the same decomposition for the economy with a small unemployment insurance system (column 4, Table 13) we observe that the total drop in consumption is about twice as large now, as already displayed in the impulse response plot. Now the (newly and existing) unemployed have significantly larger percentage consumption drops (see the fourth column of Table 14) and the share of the (now larger) consumption drop rises to 24.3%. Of course, the more pronounced consumption drop of the unemployed in a low UI benefit environment (and holding the wealth distribution fixed) is exactly what one would expect, and is already apparent in the policy functions of Fig. 10.

Table 14 contains a second important observation that we wish to stress. Looking at the magnitude of the consumption drops of households that have *not yet* lost their jobs as the economy falls into the recession (households with the idiosyncratic state transitions $s=u, s'=e$ and $s=e, s'=e$), we observe that these households, which constitute the vast majority of the population, also cut their consumption much more significantly in the (surprise) low-benefit economy, again comparing columns 3 and 4 of Table 14. This is true even though these groups in both economies start with the same wealth distribution (by construction of the thought experiment) and experience the same income loss coming from a modest decline in aggregate wages. The lower UI benefits do not have an immediate impact on the earnings of these households, since they are currently employed even though the macro economy is doing poorly. The larger cuts in consumption of these groups instead emerge because future unemployment risk has gone up for these households as the economy falls into the highly persistent recession, and the potential future income losses from unemployment are larger in the economy with low unemployment insurance. Employed households, especially those with little new worth to start with, respond by elevating their saving and cutting their consumption rates, and since employed households make up 91.6% of the population, the extra fall in consumption of about 1 percentage point (in the economy with low UI, relative to the economy with high UI) is an important contributor to the overall larger decline of aggregate consumption in the low UI economy.

[bp] For a recent empirical study on the link between unemployment and consumption expenditures, see Ganong and Noel (2015), who find reductions in consumption expenditures that are quantitatively similar to the ones our model with low unemployment insurance predicts.

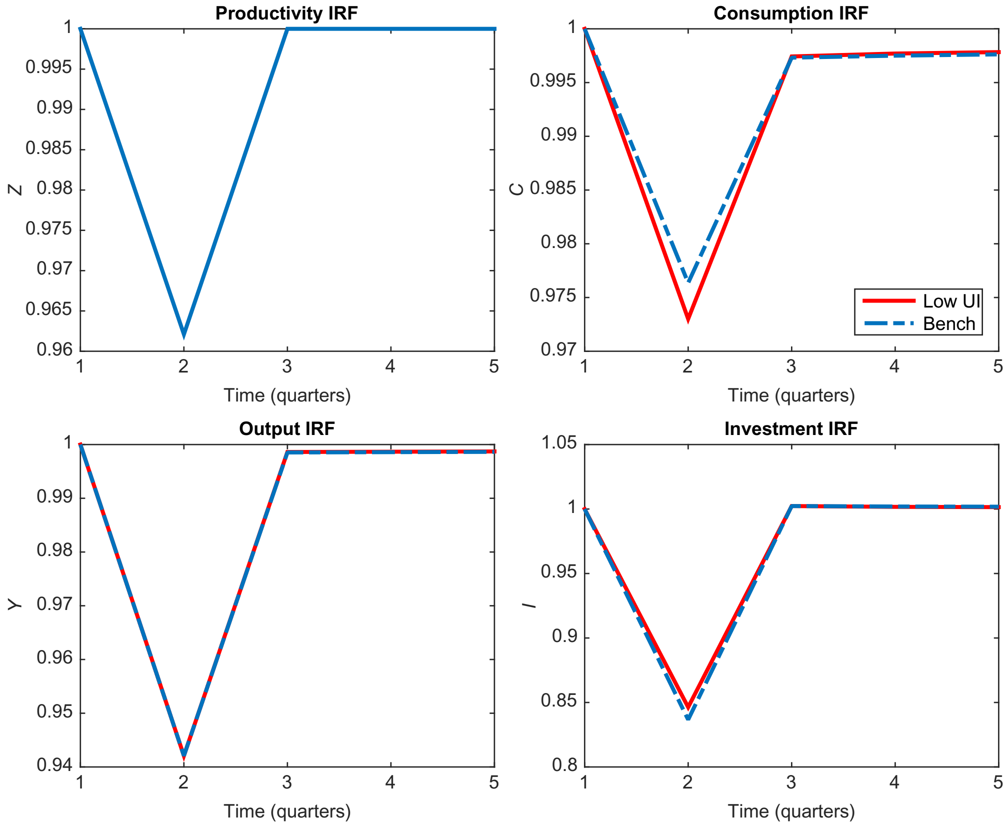


Fig. 12 Impulse response to aggregate technology shock with and without generous unemployment insurance: Onetime technology shock.

Finally, we document what happens if the wealth distribution is determined endogenously and responds to the absence of an unemployment insurance system. Fig. 12 displays the impulse responses for the benchmark economy (again) and the no-benefits economy with a prerecession wealth distribution that emerges in *that economy* after a long period of economic prosperity.[bq] Column 5 of Tables 13 and 14 breaks down the consumption response by subgroups. Overall we observe that the endogenous shift in the wealth distribution to the right that is due to the less generous unemployment insurance partially offsets the larger individual consumption declines in the no-benefits economy for a given wealth level.

To see this more precisely, compare the third and fifth columns of Table 14. The aggregate consumption decline in the economy with little unemployment insurance is somewhat larger than in the benchmark economy (by 0.3 percentage point). But very

[bq] That wealth distribution was displayed in the right panel of Fig. 10.

notably, in this economy the unemployed (both newly and already existing ones) account for a substantially larger share of the reduction in consumption, even though this group understands the possibility of a Great Recession and has access to self-insurance opportunities to prepare for it. This is primarily because the employed, now fully aware of the fact that unemployment benefits will be low if they happen to become unemployed in the recession, enter the recession with larger wealth levels and do not cut their consumption as much as when they were surprised by the expiration of their benefits (compare columns 4 and 5 in Table 14 for the employed, $s' = e$). Thus, all of the larger magnitude of the aggregate consumption decline with low UI benefits is driven by the small group of unemployed (compare columns 3 and 5 of Table 14). The end effect is an aggregate consumption decline of 2.7% that is somewhat larger, but broadly consistent with that in the benchmark economy even though *individual consumption* responses to the crisis differ markedly across the two economies for the unemployed.

6.3.1 Revisiting the Importance of "Hand-to-Mouth" Consumers

In the absence of a generous unemployment insurance system, not only is the decline in aggregate consumption larger, as the previous section has argued, but the wealth-poor, not yet unemployed households have a greater incentive to save for now more likely unemployment spells. As such, our economy with low replacement rate responds to aggregate shocks more strongly, relative to an economy with hand-to-mouth consumers, than the benchmark economy with $\rho = 50\%$. Recall that with $\rho = 50\%$ the aggregate consumption decline was 2.4%, relative to a fall of 2% in an economy with 40% hand-to-mouth consumers. With $\rho = 10\%$, the fall amounts to 2.7% in our economy and 2.1% in the hand-to-mouth consumer economy, and thus the divergence between the two models becomes stronger, on account of the elevated importance of the precautionary savings behavior of the wealth-poor, which is absent in models with exogenously given fixed shares of hand-to-mouth consumers. The recent papers by Ravn and Sterk (2013), McKay (2015), and Den Haan et al. (2016) are important examples that have stressed the importance of precautionary savings in the face of increased idiosyncratic risk for the dynamics of macro aggregates.[br]

7. INEQUALITY AND AGGREGATE ECONOMIC ACTIVITY

In the model studied so far, the wealth distribution did potentially have an important impact on the dynamics of aggregate consumption and investment, but—by construction—only a fairly negligible effect on aggregate economic activity. Output depends on capital, labor input, and aggregate TFP, and in the previous model the latter two are exogenously given.

[br] In related work Harmenberg and Oberg (2016) analyze the dynamics of consumption expenditures on durables in the presence of time-varying income risk.

The capital stock is predetermined in the short run, and even in the medium run only responds to net investment, which is a small fraction of the overall capital stock. So the output response to a negative productivity shock is exogenous on impact and, to a first approximation, exogenous (to the wealth distribution and to social insurance policies) even in the medium run. That is why in the previous section we focused on the distribution of the output decline between aggregate consumption and investment.

In the models discussed so far, aggregate demand played no independent role in shaping business cycle dynamics and, by construction, government demand management is ineffective. We now present a version of the model in which the output response to a negative shock is endogenous even in the short run, and thus potentially depends on the wealth distribution in the economy as well as policies that shape this distribution. The model retains the focus on real, as opposed to nominal, factors.[bs]

The aggregate production function continues to be given by

$$Y = Z^{*} F(K, N)$$

with $Z^{*} = ZC^{\omega}$ and $\omega > 0$,
but now consider a world in which $\omega > 0$ and thus TFP $Z^{*} = ZC^{\omega}$ endogenously responds to the level of aggregate demand. A decline in aggregate consumption triggered by a fall in Z and an ensuing reduction of aggregate wages and household incomes endogenously reduces TFP and thus output further. This model with aggregate demand externalities is in the spirit of Bai et al. (2012), Huo and Rios-Rull (2013), and Kaplan and Menzio (2014), who provide micro foundations for the aggregate productivity process we are assuming here.[bt]

Since in this model a reduction in aggregate consumption C (say, induced by a negative Z shock) feeds back into lower TFP and thus lower output, government "demand management" might be called for, even in the absence of incomplete insurance markets against idiosyncratic risk. A social insurance program that stabilizes consumption demand of those adversely affected by idiosyncratic shocks in a crisis might be desirable not just from a distributional and insurance perspective, but also from an aggregate point of view. In the model with consumption externalities, in addition to providing consumption insurance it increases productivity and accelerates the recovery.[bu]

[bs] In this chapter we abstract completely from nominal frictions that make output partially demand-determined. Representative papers that contain a lucid discussion of the demand- and supply-side determinants of aggregate *output* fluctuations in heterogeneous agent New Keynesian models are Gornemann et al. (2012), Challe et al. (2015), and Kaplan et al. (2016b).

[bt] We are certainly not claiming that our and their formulations are isomorphic on the aggregate level; rather, their work provides the fully micro-founded motivation for the reduced form approach we are taking in this section.

[bu] We think of this model as the simplest structure embedding a channel through which redistribution affects output directly and in the short run.

We now first discuss the calibration of the extended model before documenting how the presence of the demand externality affects our benchmark results.

7.1 Calibration Strategy

We retain all model parameters governing the idiosyncratic shock processes (s, y), but recalibrate the *exogenous* part of aggregate productivity Z. In addition we need to specify the strength of the externality ω. Our basic approach is to use direct observations on TFP to calibrate the exogenous process Z and then choose the magnitude of the externality ω such that the demand externality model displays the same volatility of output as the benchmark model which (as the reader might recall) was calibrated to match the severity of the two severe recession episodes we identified in the data.[bv]

7.1.1 Exogenous TFP Process Z

For comparability with the benchmark results, we retain the transition matrix $\pi(Z'|Z)$ but recalibrate the states (Z_l, Z_h) of the process. To do so, we HP-filter the Fernald (2012) data for total factor productivity, identify as severe recessions the empirical episodes with high unemployment as in the benchmark analysis, and then compute average TFP (average percentage deviations relative to the HP-trend) in the severe recession periods, identified from unemployment data, as well as in normal times. This delivers

$$\frac{Z_l}{Z_h} = \frac{1-1.84\%}{1+0.36\%} = 0.9781.$$

Thus, the newly calibrated exogenous TFP process is significantly less volatile than in the benchmark economy, where the corresponding dispersion of TFP was given by $\frac{Z_l}{Z_h} = 0.9614$.

7.1.2 Size of the Spillover ω

Given the exogenous TFP process, we now choose ω such that the externality economy has exactly the same output volatility as the benchmark economy. This requires $\omega = 0.30$.

7.2 Results

7.2.1 Aggregate Dynamics

In Fig. 13, we display the dynamics of a typical Great Recession (22 quarters of low TFP) in both the baseline economy and the demand externality economy (labeled

[bv] An alternative approach would have been to retain the original calibration of the Z process, choose a variety of ω values, and document how much amplification, relative to the benchmark model, the externality generates. The drawback of this strategy is that output is counterfactually volatile in these thought experiments unless $\omega = 0$.

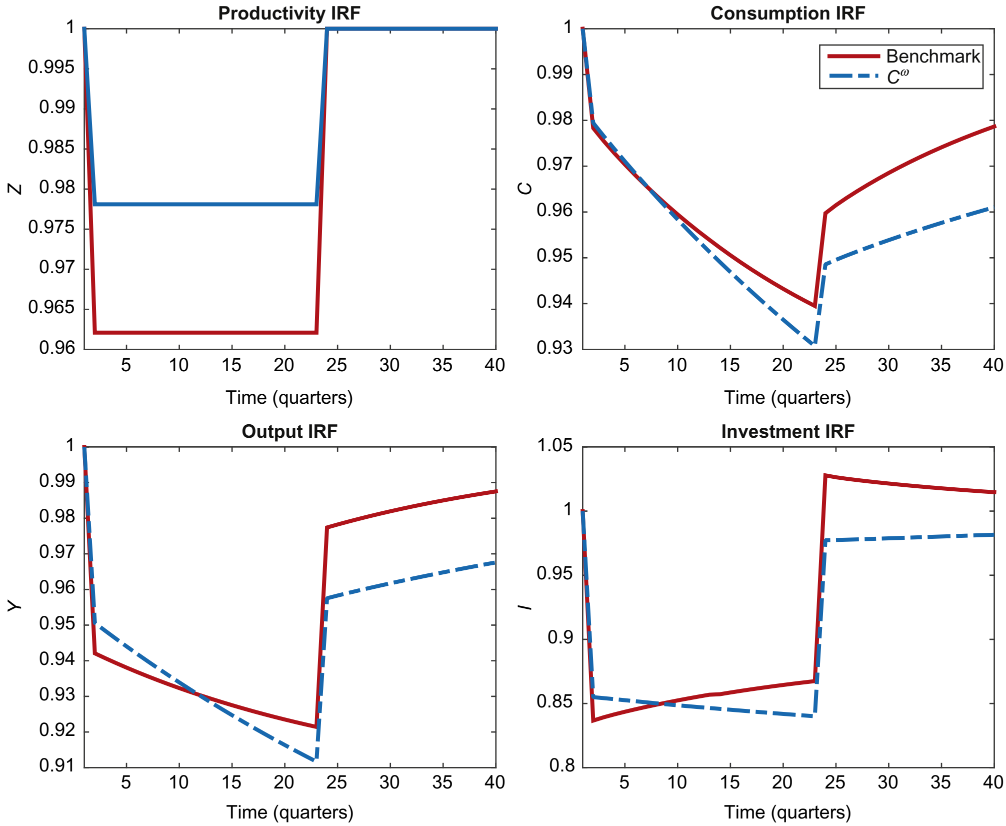


Fig. 13 Impulse response to aggregate technology shock: Comparison between benchmark and demand externality economy.

C^{ω}).[bw] The upper left panel shows that, as determined in the calibration section, a significantly smaller exogenous shock (2.2% as opposed to a 3.9% fall in TFP) is needed in the externality economy to generate a decline in output (and thus consumption and investment) of a given size. The impulse response functions are qualitatively similar in both economies, but with important quantitative differences.

First, the average decline in output in a Great Recession is the same across both economies since this is how $\frac{Z_l}{Z_h}$ was calibrated in the externality economy. However, since aggregate consumption declines during the course of a Great Recession and aggregate consumption demand impacts productivity, the decline in output is more pronounced and the recovery slower in the externality economy. Thus, the consumption externality

[bw] The figure for a one-quarter Great Recession is qualitatively similar, but less useful in highlighting the differences between both economies.

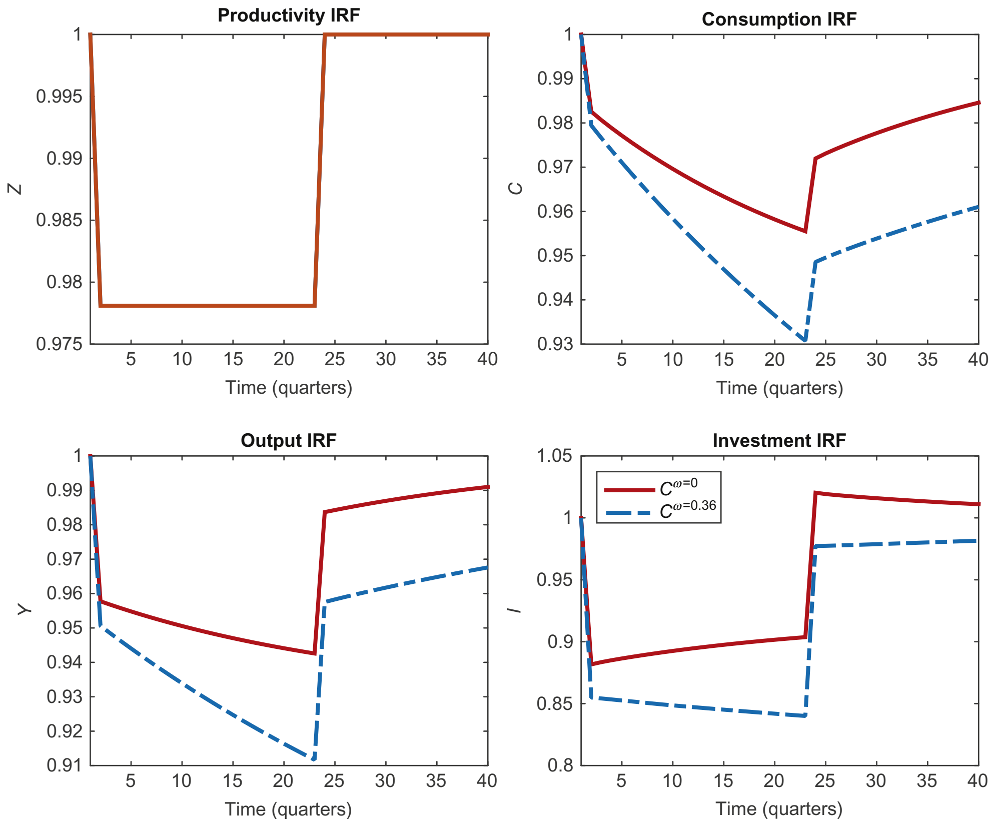


Fig. 14 Impulse response to identical aggregate technology shock: Comparison between economies with and without demand externality.

adds endogenous persistence to the model, over and above the channel already present through endogenous capital accumulation.

Of course, the demand externality mechanism also adds endogenous volatility to the model, but the fact that, via calibration, both models have the same output volatility obscures this fact. In Fig. 14, we display the magnitude of this amplification by comparing the impulse responses in two economies with the *same* exogenous TFP process (the one recalibrated for the demand externality economy), but with varying degrees of the externality ($\omega = 0$ and $\omega = 0.30$).

In contrast to Fig. 13, now the differences in the dynamics of the time series are purely driven by the presence of the demand externality. The amplification of the exogenous shock is economically important: the initial fall in output, consumption and investment is substantially larger (5.0%, 2.1% and 14.5% versus 4.2%, 1.7% and 11.8%, respectively). In addition, and consistent with Fig. 13, these larger output and consumption losses are more persistent in the economy with negative feedback effects from aggregate demand on productivity and thus production.

7.2.2 On the Importance of the Wealth Distribution When Output Is Partially Demand-Determined

In principle, the previous results measuring the importance of aggregate consumption demand for output fluctuations did not require household heterogeneity at all. However, in the previous part of the chapter, we argued that the wealth distribution is a crucial determinant of aggregate consumption fluctuations, so it stands to reason the same is true with *output* fluctuations in economies where GDP is demand-determined. In Fig. 15, we verify this point by displaying the aggregate impulse responses to a Great Recession in both the externality economy with plausible wealth heterogeneity and a version of the original Krusell–Smith economy, but also including the demand externality. The underlying exogenous TFP process is identical in both economies (and the same as in

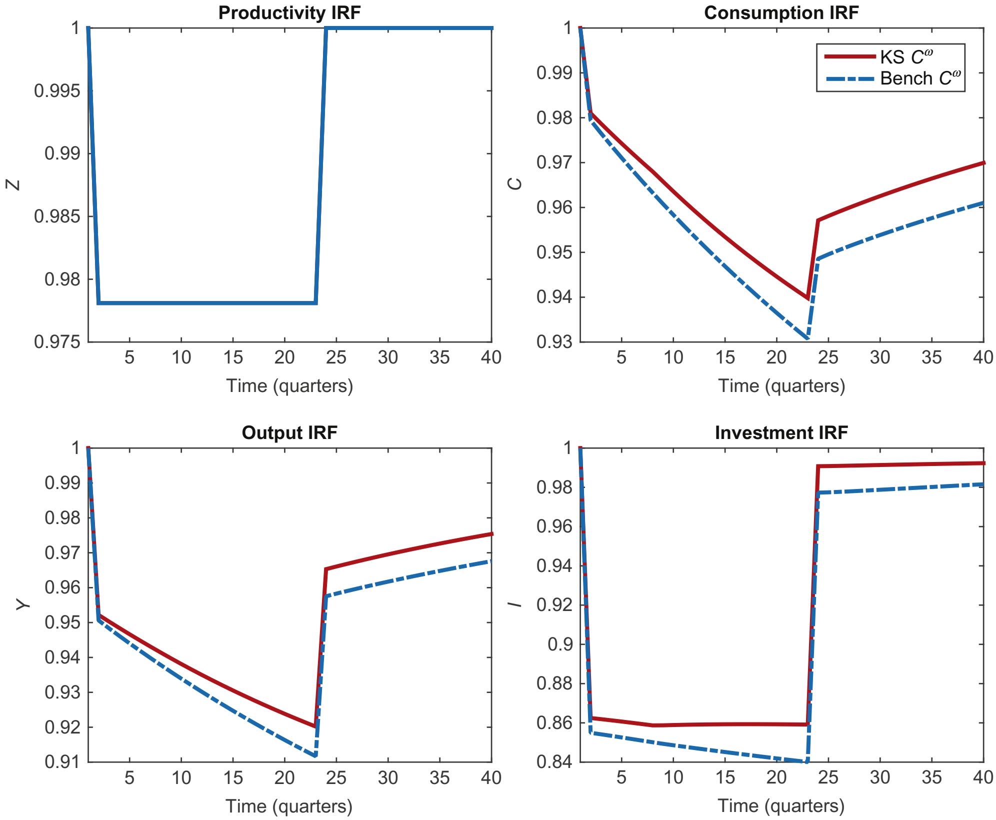


Fig. 15 Impulse response to identical aggregate technology shock: Comparison between economies with high and low wealth inequality.

Table 15 Consumption and output declines in four economies

Economy	$\Delta_1 C$	$\Delta_1 Y$	$\Delta_{22} C$	$\Delta_{22} Y$
KS, $\omega = 0$	−1.9%	−5.8%	−6.0%	−8.0%
Bench., $\omega = 0$	−2.4%	−5.8%	−6.1%	−7.8%
KS, $\omega = 0.3$	−1.9%	−4.8%	−6.0%	−8.0%
Bench., $\omega = 0.3$	−2.1%	−5.0%	−6.9%	−8.8%

Fig. 14), and to display the differences between the models most clearly, we display the dynamics of the macro economy through a 22-quarter Great Recession.

As the figure clearly indicates, in the economy with realistic wealth inequality, the output recession is significantly greater, with output losses of 5.0% on impact and 8.8% at the end of the recession, compared with declines of 4.8% and 8.0% in the original KS economy (but with demand externality).[bx] In Table 15, we summarize the consumption and output declines (on impact, and at the end of a Great Recession) for both the original KS and the benchmark economy, both with and without consumption externality.[by] It reconfirms the main message of Fig. 15: larger wealth dispersion, and especially lower wealth at the bottom of the wealth distribution, amplifies aggregate consumption recessions, as well as aggregate output recessions if the level of production is partially demand-determined. In the latter case, lower output in turn feeds back into an even more severe consumption recession. The magnitude of the differences is quantitatively significant, amounting to an additional drop of aggregate (and thus per capita) consumption of 0.9% at the end of the recession, because of larger wealth inequality induced by more realistic household heterogeneity (again comparing the benchmark model with the original KS economy).

7.2.3 On the Interaction of Social Insurance and Wealth Inequality with Demand Externalities

In Section 6.3, we demonstrated that the presence of social insurance policies has a strong impact on the aggregate consumption response to an adverse aggregate shock *for a given wealth distribution*, but also alters the long-run wealth distribution in the economy. With output partially demand-determined, these policies indirectly impact aggregate productivity and thus output. As the previous figures suggested, the effects are particularly important in the medium run due to the added persistence in the demand externality economy.

[bx] As in the economy without externality, the KS version of the model provides a very good approximation, as far as macroeconomic aggregates are concerned, for the corresponding representative agent economy.

[by] It is important to note that the results with $\omega = 0$ and $\omega = 0.3$ are not directly comparable, since in the economy with demand externality we feed in smaller TFP fluctuations, as described in the calibration section.

In Fig. 11 we documented that, holding the wealth distribution fixed, the size of the social insurance system matters greatly for the aggregate consumption (and thus investment) response to an aggregate productivity shock. Fig. 16 repeats the same thought experiment (an impulse response to a TFP shock in economies with $\rho = 50\%$ and $\rho = 10\%$ with the same prerecession wealth distribution), but now for the consumption externality model.

The key observations from Fig. 16 are that now, in the consumption externality model, the size of the unemployment insurance system affects not only the magnitude of the aggregate consumption decline on impact, but also aggregate output, and the latter effect is quite persistent.

This can perhaps be more clearly seen in Fig. 17, which displays the *difference* in the impulse response functions for output and consumption between economies with $\rho = 50\%$ and $\rho = 10\%$, for both the benchmark model and the demand externality model. The presence of sizeable unemployment insurance stabilizes aggregate

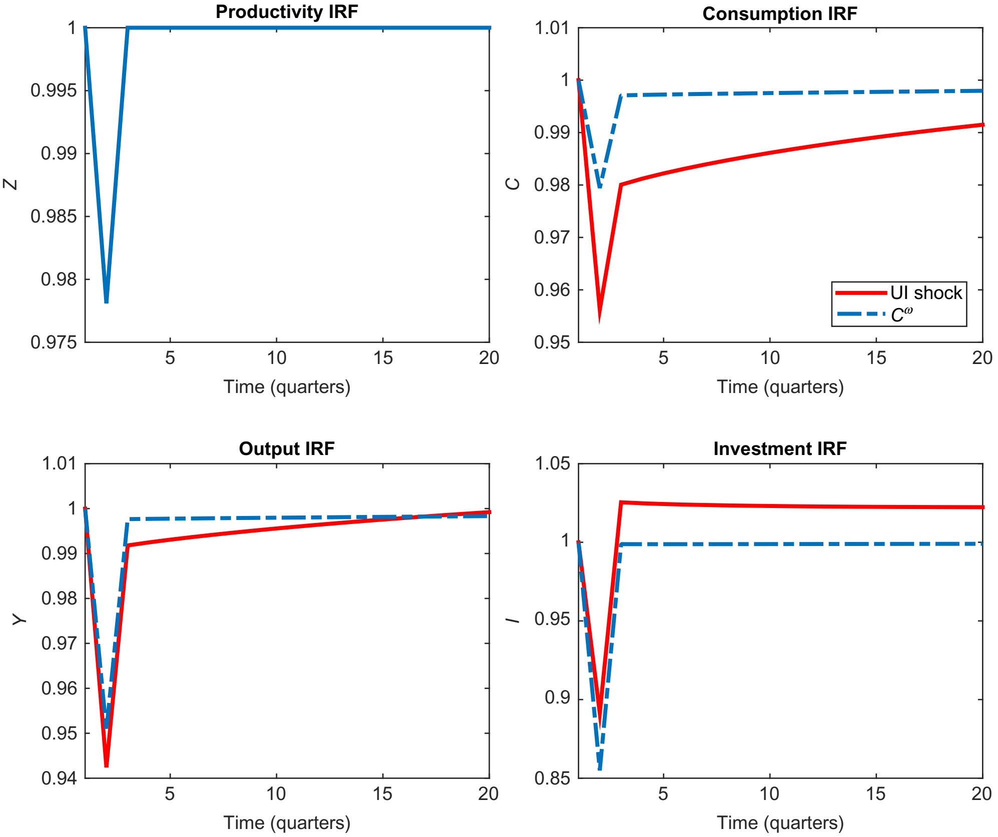


Fig. 16 Impulse response to aggregate technology shock with and without generous unemployment insurance in consumption externality model, fixed wealth distribution.

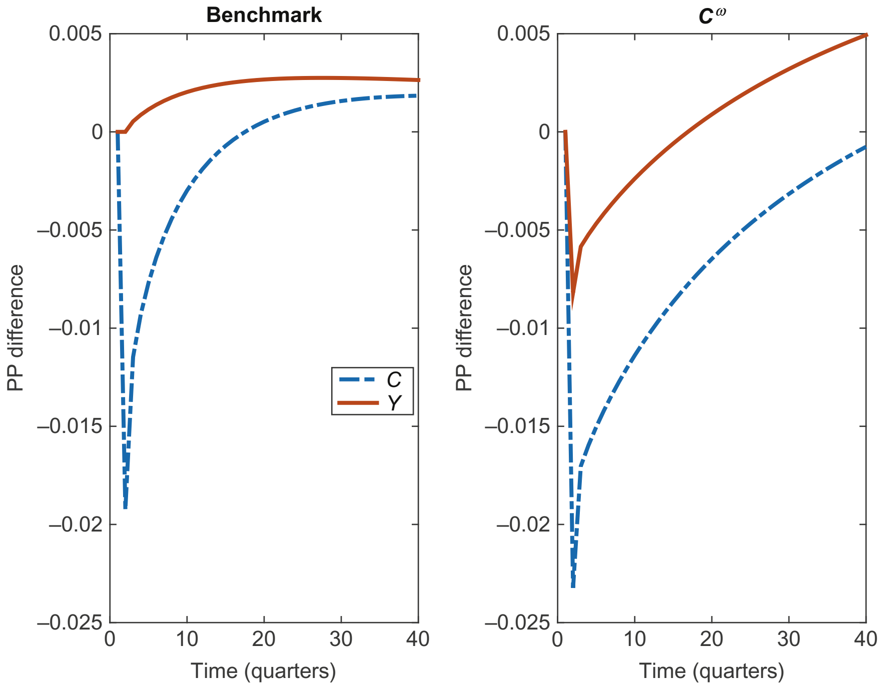


Fig. 17 Difference in IRF between $\rho = 50\%$ and $\rho = 10\%$, with and without consumption externality.

consumption more in the externality economy (the UI-induced reduction in the fall of C is 2.3% on impact and 1.3% after ten quarters of the initial shock in the externality economy, relative to 1.9% and 0.5% in the benchmark economy).

In addition, whereas in the benchmark economy more generous social insurance has no impact on output in the short run (by construction) and a moderately negative impact in the medium run (since investment recovers more slowly in the presence of more generous UI), with partially demand-determined output, UI stabilizes output significantly (close to 1% on impact, with the effect fading away only after 20 quarters—despite the fact that the shock itself only lasts for one quarter in this thought experiment.

Finally, we want to make a perhaps somewhat unexpected observation that turns out to be important for the calculation of the welfare losses of Great Recessions that we pursue in Krueger et al. (2016).[bz] The surprise removal of unemployment benefits leaves

[bz] In that paper, we contribute to the very large literature that studies the normative consequences of social insurance policies (such as unemployment insurance, social security and progressive income taxation) in quantitative heterogeneous household models. See Domeij and Heathcote (2004), Caucutt et al. (2006), Conesa et al. (2009), Peterman (2013), Heathcote et al. (2014), Mitman and Rabinovich (2015), Bakis et al. (2015), Karabarbounis (2015), Krebs et al. (2015), and Krueger and Ludwig (2016) for recent representative contributions to this literature.

households—especially those at the low end of the wealth distribution—with suboptimally small assets. These households start to save massively, especially in light of the elevated unemployment risk. Thus, in the medium run, wealth (the capital stock) and therefore aggregate consumption starts to rise. And since total factor productivity is linked to aggregate consumption demand (and since the capital stock in the economy increases), aggregate wages and output rise strongly in the medium run in the externality economy with low unemployment insurance benefits.[ca] As long as households are sufficiently patient[cb] and have *not* lost their job in the recession, the stronger recovery of the macro economy with low unemployment benefits might make these households prefer less generous unemployment insurance, despite the fact that unemployment insurance benefits act as effective aggregate demand stabilizers in the short run (again as Fig. 17 clarifies).

This last finding, discussed in much greater length in Krueger et al. (2016), leads us back to the main overall theme of this chapter: we have demonstrated that the extent of household heterogeneity with respect to income, wealth and preferences, in a canonical heterogeneous household business cycle model, determines the aggregate consumption and output dynamics over the business cycle in a quantitatively significant way. It gives social insurance policies that shape the income, consumption and wealth distributions a potentially important role in aggregate consumption and output stabilization and has (as we show in our companion work) welfare implications that vary strongly across households with different characteristics. Modeling microeconomic heterogeneity explicitly in the analysis of Great is therefore potentially quantitatively important, even if the object of research interest is purely aggregate in nature.

8. CONCLUSION

In this chapter, we used PSID data on earnings, income, consumption, and wealth as well as different versions of a canonical business cycle model with household earnings and wealth heterogeneity to study the conditions under which the cross-sectional wealth distribution shapes the business cycle dynamics of aggregate output, consumption and investment in a quantitatively meaningful way. We argued that the low end of the wealth distribution is crucial for the answer to this question. We studied mechanisms that helped to generate close to 40% of households without significantly positive net worth, including highly persistent earnings shocks, preference heterogeneity and publicly provided social insurance programs. We showed that the decline in consumption of this group of wealth-poor households at the onset of the recession generates a significantly larger

[ca] Mitman and Rabinovich (2014) argue, reversely, that the *extension* of unemployment benefits goes a long way towards explaining recent slow recoveries in US data.

[cb] Recall that the population is heterogeneous with respect to the time discount factor.

aggregate consumption drop than in a representative-household version of the neoclassical growth model. The same is true for output if it is partially demand-determined. We argued that the key mechanism underlying this result is increased precautionary savings against elevated unemployment risk, and we investigated the extent to which social insurance programs impact the strength of this channel.

Our work suggests that there are at least three important research directions that could yield new insights on the role of heterogeneity for macro outcomes.[cc]

The first is the introduction of additional dimensions of household heterogeneity, so that the model can better capture the *joint* distribution of wealth, income and expenditure we observe in the data. A more accurate mapping between the model and household micro data might change our quantitative conclusions regarding the impact of household heterogeneity on macro dynamics.

The second dimension is the introduction of a richer model of the labor market, with elastic labor supply and other frictions impacting equilibrium hours and unemployment. Doing so would allow us to better understand the link between changes in aggregate consumption expenditures and changes in aggregate output, which in this chapter we have modeled in a very reduced form way.

The final direction for promising work is the explicit introduction of aggregate shocks to the net worth of households (which one may call financial shocks). The micro data on the dynamics of household wealth have shown that during the Great Recession large changes in the net worth of households occurred, and the current model with only one asset does not capture these changes. Introducing a mechanism that can generate these fluctuations in the price of different assets could modify the mechanisms leading from the micro wealth distribution to aggregate consumption and output described in this chapter.

More generally, the emergence of new rich household and firm-level data sets, coupled with continuous theoretical and computational advances in the solutions of macro models with micro heterogeneity, as well as renewed scientific and popular interest in distributional questions, make the research field of quantitative heterogeneous agent macroeconomics an exciting area for future inquiry.

APPENDICES

A Data and Estimation Appendix

A.1 Aggregates in PSID and BEA

The series for disposable income from the BEA is Disposable Personal Income minus Medicare and Medicaid transfers, which are not reported in the PSID. The disposable income series from the PSID is constructed by adding, for each household and from

[cc] We fully acknowledge that exciting work in all these dimensions is already under way.

all members, wage and salary income, income from business and farm, income from assets (including the rental equivalent for the main residence for home owners), and all money transfers minus taxes (computed using the NBER TAXSIM calculator).

The series for consumption expenditures (from both the BEA and the PSID) includes the following expenditures categories: cars and other vehicles purchases, food (at home and away), clothing and apparel, housing (including rent and imputed rental services for owners), household equipment, utilities, transportation expenses (such as public transportation and gasoline), and recreation and accommodation services. In the PSID, imputed rental services from owners are computed using the value of the main residence times an interest rate of 4%. Total consumption expenditures are reported for a two-year period because of the timing of reporting in the PSID. In the PSID, some expenditures categories (food, utilities) are reported for the year of the interview, while others are reported for the year preceding the interview, so total expenditures span a two-year period. The measure of total consumption from the BEA is constructed by aggregating the different categories using PSID timing; so, for example, total expenditures in 2004–05 include car purchases from 2004 and food expenditures from 2005. We have excluded health services because PSID only reports out-of-pocket expenditures and insurance premia. All PSID observations are aggregated using sample weights. Table A.1 reports the 2004 levels of the per capita variables plotted in Fig. 1, alongside, for comparison purposes, the level of food expenditures from both sources and the total household personal consumption expenditures from the BEA.

Table A.1 suggests that the levels from the PSID and the BEA are not too far off, although there are differences. In particular, the aggregated PSID data are different from the aggregates from the BEA for two reasons. Comparing lines 2–3 across columns, we see that for a given category, the average from the PSID is different (typically lower) than that reported by the BEA. This discrepancy between aggregate and aggregate survey data has been widely documented before. The second reason is that some categories are not included in our PSID aggregate, either because they are mismeasured in the PSID (eg, Health expenditures) or because they are not reported by the PSID (eg, expenditures in financial services). One might wonder whether these omitted categories matter for the aggregate pattern of expenditures. Fig. A.1 reports the growth rate of total household personal consumption expenditures from the BEA, along with the growth rate for the BEA consumption expenditures that are included in the PSID

Table A.1 Per capita levels in 2004: BEA vs PSID

	BEA	PSID
1. Disposable income	$24120	$21364
2. Personal consumption (PSID aggregate)	$18705	$15889
3. Food expenditures	$3592	$2707
4. Personal consumption (total)	$27642	–

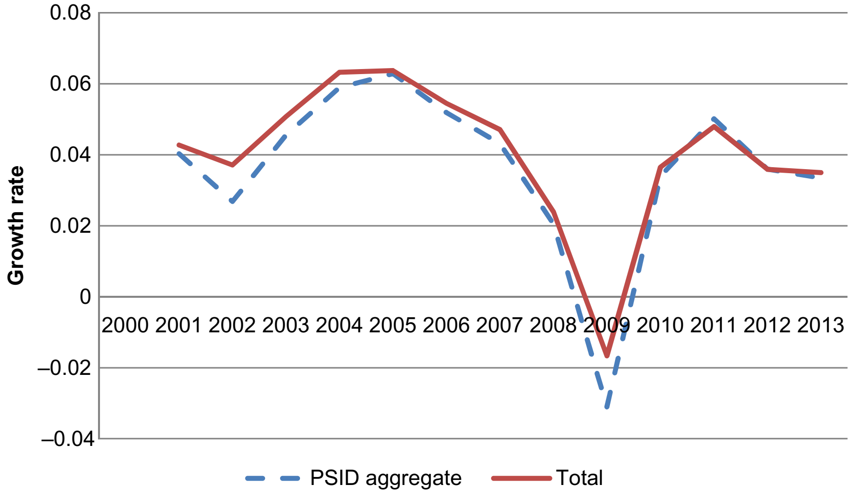


Fig. A.1 BEA consumption growth for two different aggregates.

aggregate defined above. Table A.1 suggests that categories included in the PSID aggregate cover only about 65% of the total consumption expenditures; Fig. A.1, however, shows that the cyclical pattern of total expenditures is similar to the one in the PSID aggregate, suggesting that the missing consumption categories in the PSID aggregate should not make a big difference for our results.

A.2 Standard Errors and Additional Tables

Table A.2 Annualized changes in variables across PSID net worth (2004–06 vs 2006–10) with standard errors[a]

	Net worth[b]				Disp. *Y* (%)		Cons. Exp.(%)		Exp. Rate (pp)	
	(1)		(2)		(3)	(4)	(5)	(6)	(7)	(8)
	04–06		06–10		04–06	06–10	04–06	06–10	04–06	06–10
All	**15.7** **(4.4)**	**44.6** **(12.4)**	**−3.0** **(1.6)**	**−10.2** **(6.4)**	**4.1** **(1.5)**	**1.2** **(0.3)**	**5.6** **(1.0)**	**−1.3** **(0.5)**	**0.9** **(0.9)**	**−1.6** **(0.3)**
Q1	NA	12.9 (1.5)	NA	6.6 (1.5)	7.4 (1.0)	6.7 (0.8)	7.1 (1.2)	0.6 (0.7)	−0.2 (0.9)	−4.2 (0.7)
Q2	121.9 (38.3)	19.5 (5.9)	24.4 (5.2)	3.7 (0.8)	6.7 (1.0)	4.1 (0.6)	7.2 (1.4)	2.0 (0.6)	0.3 (1.0)	−1.3 (0.4)
Q3	32.9 (3.7)	23.6 (3.1)	4.3 (1.5)	3.3 (1.1)	5.1 (0.7)	1.8 (0.4)	9.0 (4.1)	0.0 (0.7)	2.3 (2.6)	−1.1 (0.4)
Q4	17.0 (2.1)	34.7 (4.4)	1.7 (1.7)	3.8 (3.7)	5.0 (0.6)	1.7 (0.4)	5.9 (1.8)	−1.5 (0.5)	0.5 (1.1)	−2.0 (0.3)
Q5	11.6 (5.5)	132.2 (63.3)	−4.9 (1.7)	−68.4 (31.5)	1.8 (3.2)	−1.2 (0.6)	2.7 (1.7)	−3.5 (1.1)	0.5 (1.7)	−1.4 (0.8)

[a] Standard errors (in parentheses) are computed using bootstrapping with 50 sample replications.

[b] The first figure is the percentage change (growth rate), the second is the change in 000's of dollars. Standard errors for those figures are also in 000's of dollars.

Table A.3 Annualized changes in variables across PSID net worth (2006–08)

	Net worth[a]		Disp. *Y* (%)	Cons. Exp. (%)	Exp. Rate (pp)
All	**−5.1**	**−17.3**	**2.5**	**−3.3**	**−3.6**
Q1	NA	7.7	8.6	−0.7	−7.0
Q2	131.3	19.0	7.7	2.9	−3.1
Q3	18.5	13.8	3.4	−3.4	−4.2
Q4	10.4	23.0	3.2	−1.6	−3.0
Q5	−10.8	−150	−1.1	−7.3	−3.7

[a] The first figure is the percentage change (growth rate), the second is the change in 000's of dollars.

Table A.4 Annualized changes in variables across PSID net worth (2008–10)

	Net worth[a]		Disp. *Y* (%)	Cons. Exp. (%)	Exp. Rate (pp)
All	**0.5**	**1.3**	**−0.2**	**1.3**	**0.9**
Q1	NA	14.7	5.4	1.8	−2.4
Q2	101.5	5.6	0.6	3.4	2.0
Q3	24.2	11.6	0.7	1.4	0.4
Q4	12.7	20.4	0.2	2.8	1.5
Q5	−4.2	−44.6	−2.6	−0.8	1.0

[a] The first figure is the percentage change (growth rate), the second is the change in 000's of dollars.

A.3 Estimation of Earnings Process for Employed Households

To estimate the income process for employed households, we use annual household data from the PSID from 1970 to 1997. (These are all the years the PSID survey was conducted annually and for which we can construct comparable data.) We select all households with a head between ages 25 to 60. For each household, we compute total household labor income as the sum of the labor income of the head, the labor income of the spouse, income from farm and business, plus transfers. We then compute tax liabilities for each household using the TAXSIM (ver. 9) tax calculator and subtract it from household labor income to construct household disposable labor income. We then deflate disposable income using the CPI and divide it by the number of members in the household to obtain a measure of per capita real disposable household income. We then exclude the household/years observations where the head of the household is unemployed and where the wage (computed as the head's labor income divided by the head's total hours worked) is below half the minimum wage for that year. On this sample, we regress the log of per capita real disposable income on age dummies, education dummies, interaction of age and education dummies, and year dummies. Before proceeding with estimation we exclude all household income sequences that are shorter than five years. This leaves us with our final sample of 3878 household/years sequences, of an average length of 13.1 years. On these data, we compute the first differences and then the autocovariance matrix of the first differences. We then estimate the stochastic process specified in the text using generalized method of moments, targeting the covariance matrix. The weighting matrix is the identity matrix. Many thanks to Chris Tonetti for providing the Matlab routines that perform the estimation.

B Theoretical Appendix

B.1 Explicit Statement of Aggregate Law of Motion for Distribution

Since the extent of heterogeneity and the choice problem of young and old households differ significantly, it is easiest to separate the cross-sectional probability measure Φ into two components (Φ_W, Φ_R) and note that the measures integrate to Π_W and Π_R, respectively. First define the Markov transition function, conditional on staying in the young age group $j = W$ as

$$Q_{W,(Z,\Phi,Z')}((s,y,a,\beta),(\mathcal{S},\mathcal{Y},\mathcal{A},\mathcal{B})) = \sum_{s'\in\mathcal{S},y'\in\mathcal{Y}} \begin{cases} \pi(s'|s,Z',Z)\pi(y'|y): & d'_W(s,y,a,\beta;Z,\Phi)\in\mathcal{A},\beta\in\mathcal{B} \\ 0 & else \end{cases}$$

and for the old, retired age group, as

$$Q_{R,(Z,\Phi,Z')}((a,\beta),(\mathcal{A},\mathcal{B})) = \begin{cases} 1: & d'_R(a,\beta;Z,\Phi)\in\mathcal{A},\beta\in\mathcal{B} \\ 0 & else \end{cases}$$

For each Borel sets $(\mathcal{S},\mathcal{Y},\mathcal{A},\mathcal{B}) \in P(\mathcal{S}) \times P(\mathcal{Y}) \times B(\mathcal{A}) \times P(\mathcal{B})$, the cross-sectional probability measures of the young and old tomorrow are then given by[cd]

$$\begin{aligned} H_W(Z,\Phi,Z')(\mathcal{S},\mathcal{Y},\mathcal{A},\mathcal{B}) = \theta \int Q_{W,(Z,\Phi,Z')}((s,y,a,\beta),(\mathcal{S},\mathcal{Y},\mathcal{A},\mathcal{B}))d\Phi_W \\ +(1-\nu)\mathbf{1}_{\{0\in\mathcal{A}\}}\sum_{s'\in\mathcal{S}}\Pi_Z(s')\sum_{y'\in\mathcal{Y}}\Pi(y')\sum_{\beta'\in\mathcal{B}}\Pi(\beta') \end{aligned}$$

and

$$\begin{aligned} H_R(Z,\Phi,Z')(\mathcal{A},\mathcal{B}) = \nu \int Q_{R,(Z,\Phi,Z')}((a,\beta),(\mathcal{A},\mathcal{B}))d\Phi_R \\ +(1-\theta)\int Q_{W,(Z,\Phi,Z')}((s,y,a,\beta),(S,Y,\mathcal{A},\mathcal{B}))d\Phi_W. \end{aligned}$$

[cd] These expressions capture the assumption that in each period, a measure $1 - nu$ of newborn households enter the economy as workers, with zero assets and with idiosyncratic productivities and discount factors drawn from the stationary distributions, and that a fraction $1 - \theta$ of working households retire, and that the retirement probability is independent of all other characteristics.

C Computational Appendix

The computational strategy follows the framework developed initially in Krusell and Smith (1998), which was further adapted by Storesletten et al. (2007) and Gomes and Michaelides (2008). In particular, we employ the computational strategy outlined in Maliar et al. (2010), focusing on the nonstochastic simulation algorithm first introduced by Young (2010).

C.1 The Individual Problem

We approximate the true aggregate state (S=(Z, Φ)) by $\hat{S}$, whose specific form depends on which version of the model we solve, which is detailed explicitly later. Thus, the household state is determined by $(s, y, a, \beta; \hat{S})$ in working life and $(a, \beta; \hat{S})$ when retired.

The solution method from Maliar et al. (2010) is an Euler equation algorithm that takes into account occasionally binding borrowing constraints. The problem to be solved is as follows:

$$\begin{aligned}
&\textbf{Retired:}\\
c_R(a,\beta;\hat{S})^{-\sigma} - \lambda &= \nu\beta\mathbb{E}[(1-\delta+r'(\hat{S}'))c_R'(d_R',\beta;\hat{S}')^{-\sigma}]\\
d_R'(a,\beta;\hat{S}) + c_R(a,\beta;\hat{S}) &= b_{SS}(\hat{S}) + (1+r(\hat{S})-\delta)a/\nu\\
d_R'(a,\beta;\hat{S}) &\geq 0\\
\lambda \geq 0, &\quad \lambda d_R'(a,\beta;\hat{S}) = 0
\end{aligned}$$

$$\begin{aligned}
&\textbf{Working:}\\
c_W(s,y,a,\beta;\hat{S})^{-\sigma} - \lambda &= \theta\beta\mathbb{E}[(1-\delta+r'(\hat{S}'))c_W'(s',y',d_W',\beta;\hat{S}')^{-\sigma}]\\
&\quad +(1-\theta)\beta\mathbb{E}[(1-\delta+r'(\hat{S}'))c_R'(d_W',\beta;\hat{S}')^{-\sigma}]\\
d_W'(s,y,a,\beta;\hat{S}) + c(s,y,a,\beta;\hat{S}) &= (1-\tau(Z;\rho))w(\hat{S})y[1-(1-\rho)1_{s=u}] + (1+r(\hat{S})-\delta)a\\
d_W'(s,y,a,\beta;\hat{S}) &\geq 0\\
\lambda \geq 0, &\quad \lambda d_W'(s,y,a,\beta;\hat{S}) = 0,
\end{aligned}$$

where λ is the Lagrange multiplier on the borrowing constraint.

We eliminate consumption via the budget constraint and then guess a policy rule for $d_W'(s,y,a,\beta;\hat{S})$ and $d_R'(a,\beta;\hat{S})$. We then substitute the policy rule to compute $d_W''(s',y',d_W',\beta;\hat{S}')$, $d_R''(d_W',\beta;\hat{S}')$ and $d_R''(d_R',\beta;\hat{S}')$, and use the Euler equation to back out the implied policy rule for d'. If the implied policy rule is the same as the conjectured policy rule, we have computed the optimal policy; if not, we update the guess and repeat.

C.2 The Simulation Algorithm

In order to simulate the model, we pick a grid on $\mathcal{A}$ and fix a distribution of workers $\Phi_0 \in S \times Y \times A \times B$ space. We fix a long time series for the realization of the aggregate shock, Z. Using the realization Z_t and Φ_t, we can compute $\hat{S}_t$ and then apply the policy rules from the individual problem and the Markov transition matrices associated with s and y to compute Φ_{t+1} by interpolating onto the grid points in $\mathcal{A}$.

C.3 Approximating the Aggregate Law of Motion

C.3.1 KS and Benchmark Economies

For the KS and benchmark economies, we approximate the true aggregate state with $\hat{S}_t = (Z_, \bar{K}_t)$ where $\bar{K}_t$ is the average capital in the economy. Agents need to forecast the evolution of the capital stock. We conjecture that the law of motion in capital depends only on the Z and $\bar{K}$:

$$\log(\bar{K}_{t+1}) = a_0(Z_t) + a_1(Z_t)\log(\bar{K}_t)$$

We conjecture coefficients a_0 and a_1, solve the household problem, and simulate the economy. Then, using the realized sequence of $\hat{S}$, we perform the previous regression and check whether the implied coefficients are the same as the conjectured ones. If they are, we have found the law of motion; if not, we update our guess and repeat.

For the KS economy, the computed law of motion is as follows:

$$\begin{aligned} \log(\bar{K}_{t+1}) &= 0.1239 + 0.9652\log(\bar{K}_t) \quad \text{if} \quad Z_t = Z_l \\ \log(\bar{K}_{t+1}) &= 0.1334 + 0.9638\log(\bar{K}_t) \quad \text{if} \quad Z_t = Z_h. \end{aligned}$$

The R^2 for both regressions are in excess of 0.999999. Note, however, that Den Haan (2010) points out that despite having large R^2 values, the accuracy of the solution can still be poor, and suggests simulation of the capital stock under the policy rule and comparing it with the capital stock that is calculated by aggregating across the distribution. We do this for 3000 time periods. The average error between the implied law of motion from the forecast equations and the computed law of motion is 0.02%, with a maximum error of 0.10%.

For the benchmark economy, the computed law of motion is as follows:

$$\begin{aligned} \log(\bar{K}_{t+1}) &= 0.0924 + 0.9716\log(\bar{K}_t) \quad \text{if} \quad Z_t = Z_l \\ \log(\bar{K}_{t+1}) &= 0.0929 + 0.9723\log(\bar{K}_t) \quad \text{if} \quad Z_t = Z_h \end{aligned}$$

The R^2 for both regressions are in excess of 0.99999. Similar to the previous computation, we check the accuracy of the law of motion. We find that the average error between the implied law of motion and the actual capital stock computed from the distribution is 0.01%, with a maximum error of 0.07%.

C.3.2 Consumption Externality Economy

In the economy with the aggregate consumption externality, we add contemporaneous consumption as a state variable in our approximation of the true aggregate state, $\hat{S} = (Z, \bar{K}, C)$. We therefore need an additional law of motion for how aggregate consumption evolves. We conjecture the same form of law of motion for the average capital stock; however, we allow the evolution of aggregate consumption to depend on both the average capital stock and aggregate consumption:

$$\log(\bar{K}_{t+1}) = a_0(Z_t) + a_1(Z_t)\log(\bar{K}_t)$$
$$\log(C_{t+1}) = b_0(Z_t, Z_{t+1}) + b_1(Z_t, Z_{t+1})\log(\bar{K}_t) + b_2(Z_t, Z_{t+1})\log(C_t).$$

Note that because capital is predetermined in the current period, the forces rule for capital depends only on contemporaneous variables. Because aggregate consumption is an equilibrium outcome in the next period, we allow for the forecast to depend on the subsequent period's realization of the Z shock. Thus, there are four sets of coefficients to be estimated for the law of motion for consumption. The computed forecast equations are as follows:

$$\begin{aligned} \log(\bar{K}_{t+1}) &= 0.0872 + 0.9736\log(\bar{K}_t) \quad \text{if} \quad Z_t = Z_l \\ \log(\bar{K}_{t+1}) &= 0.0626 + 0.9816\log(\bar{K}_t) \quad \text{if} \quad Z_t = Z_h \end{aligned}$$

and

$$\begin{aligned} \log(C_{t+1}) &= -0.0205 + 0.0023\log(\bar{K}_t) + 0.9675\log(C_t) \quad \text{if} \quad (Z, Z') = (Z_l, Z_l) \\ \log(C_{t+1}) &= -0.5061 + 0.2882\log(\bar{K}_t) + 0.5297\log(C_t) \quad \text{if} \quad (Z, Z') = (Z_l, Z_h) \\ \log(C_{t+1}) &= -0.3560 + 0.1893\log(\bar{K}_t) + 0.6626\log(C_t) \quad \text{if} \quad (Z, Z') = (Z_h, Z_l) \\ \log(C_{t+1}) &= -0.0506 + 0.0360\log(\bar{K}_t) + 0.9295\log(C_t) \quad \text{if} \quad (Z, Z') = (Z_h, Z_h) \end{aligned}$$

with R^2 in excess of 0.9999, 0.9999999, 0.9999, 0.9999, 0.99999, 0.99999, respectively. As before, we check the accuracy of the two laws of motion. We find that the average error between the implied law of motion and the actual capital stock computed from the distribution is 0.02%, with a maximum error of 0.30%, and for the path of aggregate consumption the mean error is 0.02% with a maximum error of 0.24%. Although the externality economy has slightly larger forecast errors, the fit of the predicted aggregates is still excellent.

C.4 Digression: Why Quasi-Aggregation?

One of the implications of the results in the main text is that the wealth distribution (and especially the fraction of the population with little or no wealth) is quantitatively important for the macroeconomic consumption and investment response to an aggregate technology shock. This, however, does not imply that Krusell and Smith's (1998) original quasi-aggregation result fails.[ce] Recall that this result states that only the mean of the current wealth distribution (as well as the current aggregate shock Z) is required to accurately predict the future capital stock and therefore future interest rates and wages.

[ce] In fact, our computational method that follows theirs rather closely relies on quasi-aggregation continuing to hold.

The previous experiment compared consumption and investment dynamics in *two economies* that differed substantially in their wealth distributions. For a given economy, if the wealth distribution does not move significantly in response to aggregate shocks, then it would be irrelevant for predicting future aggregates and prices. However, in the high-wealth-inequality economy, the wealth distribution *does* move over the cycle. For example, the share of households at the borrowing constraint displays a coefficient of variation of 7%. However, what is really crucial for quasi-aggregation to occur is whether the movement, over the cycle, in the key features of the wealth distribution is explained well by movements in Z and K, the state variables in the forecast equations of households. We find that it is, even in the high-wealth-inequality economy.

For example, if we regress the fraction of people at the borrowing constraint tomorrow on Z in simulated data, we obtain an R^2 of around 0.8. Therefore, the vast majority of the variation in households at the borrowing limit is very well predicted by the aggregate state variables (Z, K). This finding is robust to alternative definitions of constrained households (households exactly at wealth 0, households who save less than 1%, less than 10%, or less than 25% of the quarterly wage) and alternative moments of the wealth distribution. It is this finding that makes quasi-aggregation hold, despite the strong impact of the wealth distribution on the aggregate consumption and investment response to aggregate technology shocks.

C.5 Recovering the Value Function

As we solve the model by exploiting the Euler equation, if one were to perform welfare calculations (as in Krueger et al. (2016) one needs to recover the value functions as a function of the idiosyncratic and aggregate states. To calculate them, we use policy function iteration. We make an initial guess for the value function, v^0, then calculate v^1 by solving the recursive household decision problem (we need not perform the maximization, since we have already computed the optimal policy function). We approximate the value function with a cubic spline interpolation in assets, as well as in aggregate capital (and for the demand externality model, we also aggregate consumption). If v^1 is sufficiently close to v^0 (in the sup-norm sense), we stop; otherwise, we proceed to compute v^2 taking v^1 as the given value function. We proceed until convergence. For the economies with retirement, we first recover the value function for retired households, v_R, and then proceed to recover the value function for working-age households, v_W.

ACKNOWLEDGMENTS

We wish to thank the editors John Taylor and Harald Uhlig, discussants of the chapter, Carlos Thomas, Claudio Michelacci, Tony Smith, and Chris Tonetti, as well as seminar participants at several conferences and institutions for many useful comments and Simone Civale for outstanding research assistance. D.K. gratefully acknowledges financial support from the NSF under grants SES 1123547 and SES 1326781.

REFERENCES

Aguiar, M., Bils, M., 2015. Has consumption inequality mirrored income inequality? Am. Econ. Rev. 105, 2725–2756.

Atkinson, A., Bourguignon, F. (Eds.), 2015. Handbook of Income Distribution, vol. 2. North-Holland, Amsterdam.

Aiyagari, S., 1994. Uninsured idiosyncratic risk and aggregate saving. Q. J. Econ. 109, 659–684.

Ameriks, J., Briggs, J., Caplin, A., Shapiro, M., Tonetti, C., 2015. Long-term care utility and late in life saving. NBER working paper no. 20973.

Arellano, M., Blundell, R., Bonhomme, S., 2015. Household earnings and consumption: a nonlinear framework. Working paper.

Athreya, K., Owens, A., Schwartzman, F., 2015. Does redistribution increase output? The centrality of labor supply. Working paper.

Attanasio, O., 1999. Consumption. In: Taylor, J., Woodford, M. (Eds.), Handbook of Macroeconomics. Elsevier Science, Amsterdam.

Attanasio, O., Weber, G., 2010. Consumption and saving: models of intertemporal allocation and their implications for public policy. J. Econ. Lit. 48, 693–751.

Auclert, A., 2014. Monetary policy and the redistribution channel. Working paper.

Auclert, A., Rognlie, M., 2016. Inequality and aggregate demand. Working paper.

Bachmallnn, R., Caballero, R., Engel, E., 2013. Aggregate implications of lumpy investment: new evidence and a DSGE model. Am. Econ. J. Macroecon. 5, 29–67.

Bai, Y., Rios Rull, J.V., Storesletten, K., 2012. Demand shocks as productivity shocks. Working paper.

Bakis, O., Kaymak, B., Poschke, M., 2015. Transitional dynamics and the optimal progressivity of income redistribution. Rev. Econ. Dyn. 18, 679–693.

Barro, R., 2006. Rare disasters and asset markets in the twentieth century. Q. J. Econ. 121, 823–866.

Bayer, C., Lütticke, R., Pham-Dao, L., Tjaden, V., 2015. Precautionary savings, illiquid assets, and the aggregate consequences of shocks to household income risk. Working paper.

Benabou, R., 2002. Tax and education policy in a heterogeneous-agent economy: what levels of redistribution maximize growth and efficiency? Econometrica 70, 481–517.

Benhabib, J., Bisin, A., 2016. Skewed wealth distributions: theory and empirics. NBER working paper no. 21924.

Benhabib, J., Bisin, A., Zhu, S., 2011. The distribution of wealth and fiscal policy in economies with finitely lived agents. Econometrica 79, 123–157.

Berger, D., Vavra, J., 2015. Consumption dynamics during recessions. Econometrica 83, 101–154.

Bewley, T., 1986. Contributions to mathematical economics in honor of gerard debreu. In: Hildenbrand, W., Mas-Colell, A. (Eds.), Stationary Monetary Equilibrium with a Continuum of Independently Fluctuating Consumers. North-Holland, Amsterdam, pp. 79–102.

Blank, R., Card, D., 1991. Recent trends in insured and uninsured unemployment: is there an explanation? Q. J. Econ. 106, 1157–1190.

Blundell, R., Pistaferri, L., Preston, I., 2008. Consumption inequality and partial insurance. Am. Econ. Rev. 98, 1887–1921.

Brinca, P., Holter, H., Krusell, P., Malafry, L., 2016. Fiscal multipliers in the 21st century. J. Monet. Econ. 77, 53–69.

Brüggemann, B., Yoo, J., 2015. Aggregate and distributional effects of increasing taxes on top income earners. Working paper.

Buera, F., 2009. A dynamic model of entrepreneurship with borrowing constraints: theory and evidence. Ann. Finance 5, 443–464.

Burkhauser, R.V., Larrimore, J., Simon, K., 2012. A 'second opinion' on the economic health of the American middle class. Natl. Tax J. 65, 7–32.

Cagetti, M., De Nardi, M., 2006. Entrepreneurship, frictions, and wealth. J. Polit. Econ. 114, 835–870.

Carroll, C., 1992. The buffer-stock theory of saving: some macroeconomic evidence. Brook. Pap. Econ. Act. 1992, 61–135.

Carroll, C., 1997. Buffer-stock saving and the life cycle/permanent income hypothesis. Q. J. Econ. 112, 1–55.

Carroll, C., Slacalek, J., Tokuoka, K., White, M., 2015. The distribution of wealth and the marginal propensity to consume. Working paper.

Castaneda, A., Diaz-Gimenez, J., Rios-Rull, J.V., 1998. Exploring the income distribution business cycle dynamics. J. Monet. Econ. 42, 93–130.

Castaneda, A., Diaz-Gimenez, J., Rios-Rull, J.V., 2003. Accounting for the U.S. earnings and wealth inequality. J. Polit. Econ. 111, 818–857.

Caucutt, E., Imrohoroglu, S., Kumar, K., 2006. Does the progressivity of income taxes matter for human capital and growth? J. Public Econ. Theory 8, 95–118.

Challe, E., Matheron, J., Ragot, X., Rubio-Ramirez, J., 2015. Precautionary saving and aggregate demand. Working paper.

Chang, Y., Kim, S.B., 2007. Heterogeneity and aggregation: implications for labor-market fluctuations. Am. Econ. Rev. 97, 1939–1956.

Chatterjee, S., Corbae, D., Nakajima, M., Rios-Rull, J.V., 2007. A quantitative theory of unsecured consumer credit with risk of default. Econometrica 75, 1525–1590.

Chodorow-Reich, G., Karabarbounis, L., 2016. The cyclicality of the opportunity cost of employment. J. Polit. Econ. (forthcoming).

Cocco, J., Gomes, F., Maenhout, P., 2005. Consumption and portfolio choice over the life-cycle. Rev. Financ. Stud. 18, 491–533.

Conesa, J., Kitao, S., Krueger, D., 2009. Taxing capital? Not a bad idea after all!. Am. Econ. Rev. 99, 25–48.

De Nardi, M., 2004. Wealth inequality and intergenerational links. Rev. Econ. Stud. 71, 743–768.

De Nardi, M., 2015. Quantitative models of wealth inequality: a survey. NBER working paper no. 21106.

De Nardi, M., French, E., Jones, J., 2010. Why do the elderly save? The role of medical expenses. J. Polit. Econ. 118, 39–75.

De Nardi, M., Fella, G., Yang, F., 2015. Piketty's book and macro models of wealth inequality. NBER working paper no. 21730.

Deaton, A., 1991. Saving and liquidity constraints. Econometrica 59, 1221–1248.

Deaton, A., 1992. Understanding Consumption. Oxford University Press, New York.

Den Haan, W., 2010. Assessing the accuracy of the aggregate law of motion in models with heterogeneous agents. J. Econ. Dyn. Control 34, 79–99.

Den Haan, W., Rendahl, P., Riegler, M., 2016. Unemployment (fears), precautionary savings, and aggregate demand. Working paper.

Domeij, D., Heathcote, J., 2004. On the distributional effects of reducing capital taxes. Int. Econ. Rev. 45, 523–554.

Fernald, J., 2012. Productivity and potential output before, during, and after the great recession. Federal Reserve Bank of San Francisco, Working paper 2012–18.

Fisher, J., Johnson, D., Smeeding, T., Thompson, J., 2015. Inequality in 3D: income, consumption and wealth. Presentation at 2015 NBER Summer Institute.

Gabaix, X., Lasry, J., Lions, P., Moll, B., 2014. The dynamics of inequality. Working paper.

Ganong, P., Noel, P., 2015. How does unemployment affect consumer spending? Working paper.

Glover, A., Heathcote, J., Krueger, D., Rios-Rull, J.V., 2014. Inter-generational redistribution in the great recession. Working paper.

Gomes, F., Michaelides, A., 2008. Asset pricing with limited risk sharing and heterogeneous agents. Rev. Financ. Stud. 21, 415–448.

Gornemann, N., Kuester, K., Nakajima, M., 2012. Monetary policy with heterogeneous agents. Working paper.

Gourio, F., 2013. Credit risk and disaster risk. Am. Econ. J. Macroecon. 5, 1–34.

Gruber, J., 1994. The consumption smoothing benefits of unemployment insurance. NBER working paper no. 4750.

Guerrieri, V., Lorenzoni, G., 2012. Credit crises, precautionary savings, and the liquidity trap. Working paper.

Guvenen, F., 2009. An empirical investigation of labor income processes. Rev. Econ. Dyn. 12, 58–79.

Guvenen, F., 2011. Macroeconomics with heterogeneity: a practical guide. Econ. Q. 97, 255–326.

Guvenen, F., Ozkan, S., Song, J., 2014. The nature of countercyclical income risk. J. Polit. Econ. 122, 621–660.

Hagedorn, M., Karahan, F., Manovskii, I., Mitman, K., 2013. Unemployment benefits and unemployment in the great recession: the role of macro effects. NBER working paper no. 19499.

Hagedorn, M., Manovskii, I., Mitman, K., 2015. The impact of unemployment benefit extensions on employment: the 2014 employment miracle? NBER working paper no. 20884.

Harmenberg, K., Oberg, E., 2016. Durable expenditure dynamics under time-varying income risk. Working paper.

Heathcote, J., 2005. Fiscal policy with heterogeneous agents and incomplete markets. Rev. Econ. Stud. 72, 161–188.

Heathcote, J., Perri, F., 2015. Wealth and volatility. NBER working paper no. 20994.

Heathcote, J., Storesletten, K., Violante, G., 2009. Quantitative macroeconomics with heterogeneous households. Annu. Rev. Econ. 1, 319–354.

Heathcote, J., Perri, F., Violante, G.L., 2010. Unequal we stand: an empirical analysis of economic inequality in the United States, 1967–2006. Rev. Econ. Dyn. 13, 15–51.

Heathcote, J., Storesletten, S., Violante, G., 2014. Optimal tax progressivity: an analytical framework. Working paper.

Hendricks, L., 2007. How important is discount rate heterogeneity for wealth inequality? J. Econ. Dyn. Control. 31, 3042–3068.

Herkenhoff, K., 2015. The impact of consumer credit access on unemployment. Working paper.

Huggett, M., 1993. The risk-free rate in heterogeneous-agent incomplete-insurance economies. J. Econ. Dyn. Control. 17, 953–969.

Huggett, M., 1996. Wealth distribution in life-cycle economies. J. Monet. Econ. 38, 469–494.

Huggett, M., 1997. The one-sector growth model with idiosyncratic shocks: steady states and dynamics. J. Monet. Econ. 39, 385–403.

Huo, Z., Rios-Rull, J.V., 2013. Paradox of thrift recessions. Working paper.

Huo, Z., Rios-Rull, J.V., 2016. Balance sheet recessions. Working paper.

Imrohoroglu, A., 1989. Cost of business cycles with indivisibilities and liquidity constraints. J. Polit. Econ. 97, 1364–1383.

Jappelli, T., Pistaferri, L., 2014. Fiscal policy and MPC heterogeneity. Am. Econ. J. Macroecon. 6, 107–136.

Kaplan, G., Menzio, G., 2014. Shopping externalities and self-fulfilling unemployment fluctuations. J. Pol. Econ. 123, 771–825.

Kaplan, G., Violante, G., 2014. A model of the consumption response to fiscal stimulus payments. Econometrica 82, 1199–1239.

Kaplan, G., Mitman, K., Violante, G., 2016a. Consumption and house prices in the great recession: model meets evidence. Working paper.

Kaplan, G., Moll, B., Violante, G., 2016b. Monetary policy according to hank. NBER working paper No.21897.

Karabarbounis, M., 2015. A road map for efficiently taxing heterogeneous agents. American Economic Journal: Macroeconomics 8, 182–214.

Kekre, R., 2015. Unemployment insurance in macroeconomic stabilization. Working paper.

Khan, A., Thomas, J., 2008. Idiosyncratic shocks and the role of nonconvexities in plant and aggregate investment dynamics. Econometrica 76, 395–436.

Kindermann, F., Krueger, D., 2015. High marginal tax rates on the top 1%? lessons from a life cycle model with idiosyncratic income risk, *NBER Working Paper 20601*.

Kopecky, K., Suen, R., 2010. Finite state Markov-chain approximations to highly persistent processes. Rev. Econ. Dyn. 13, 701–714.

Krebs, T., Kuhn, M., Wright, M., 2015. Human capital risk, contract enforcement and the macroeconomy. Am. Econ. Rev. 105, 3223–3272.

Krueger, D., Ludwig, A., 2016. On the optimal provision of social insurance: progressive taxation versus education subsidies in general equilibrium. Journal of Monetary Economics 77, 72–98.

Krueger, D., Perri, F., 2006. Does income inequality lead to consumption inequality? Evidence and theory. Rev. Econ. Stud. 73, 163–193.

Krueger, D., Perri, F., 2011. How do households respond to income shocks? Working paper.

Krueger, D., Perri, F., Pistaferri, L., Violante, G., 2010. Cross-sectional facts for macroeconomists. Rev. Econ. Dyn. 13, 1–14.

Krueger, D., Mitman, K., Perri, F., 2016. On the distribution of the welfare losses of large recessions. Proc. Econ. Soc. World Congress (forthcoming).

Krusell, P., Smith, A., 1997. Income and wealth heterogeneity, portfolio choice, and equilibrium asset returns. Macroecon. Dyn. 1, 387–422.

Krusell, P., Smith, A., 1998. Income and wealth heterogeneity in the macroeconomy. J. Polit. Econ. 106, 867–896.

Krusell, P., Smith, T., 2006. Quantitative macroeconomic models with heterogeneous agents. In: Advances in Economics and Econometrics: Theory and Applications. Ninth World Congress.

Krusell, P., Mukoyama, T., Sahin, A., 2010. Labour-market matching with precautionary savings and aggregate fluctuations. Rev. Econ. Stud. 77, 1477–1507.

Kuhn, M., Rios-Rull, V., 2015. 2013 update on the U.S. earnings, income, and wealth distributional facts: a view from macroeconomic modelers. Working paper.

Kuznets, S., 1955. Economic growth and income inequality. Am. Econ. Rev. 45, 1–28.

Livshits, I., MacGee, J., Tertilt, M., 2007. Consumer bankruptcy: a fresh start. Am. Econ. Rev. 97, 402–418.

Maliar, L., Maliar, S., Valli, F., 2010. Solving the incomplete markets model with aggregate uncertainty using the Krusell-Smith algorithm. J. Econ. Dyn. Control. 34, 42–49.

McKay, A., 2015. Time-varying idiosyncratic risk and aggregate consumption dynamics. Working paper.

McKay, A., Reis, R., 2016. The role of automatic stabilizers in the U.S. business cycle. Econometrica 84, 141–194.

Meghir, C., Pistaferri, L., 2004. Income variance dynamics and heterogeneity. Econometrica 72, 1–32.

Mitman, K., Rabinovich, S., 2014. Do unemployment benefit extensions explain the emergence of jobless recoveries? Working paper.

Mitman, K., Rabinovich, S., 2015. Optimal unemployment insurance in an equilibrium business-cycle model. J. Monet. Econ. 71, 99–118.

Perri, F., Steinberg, J., 2012. Inequality and redistribution during the great recession. Federal Reserve Bank of Minneapolis, Economic Policy Paper.

Peterman, W., 2013. Determining the motives for a positive optimal tax on capital. J. Econ. Dyn. Control. 37, 265–295.

Piketty, T., 2014. Capital in the Twenty-First Century. Belknap Press, Cambridge, MA.

Piketty, T., Saez, E., 2003. Income inequality in the United States, 1913–1998. Q. J. Econ. 118, 1–41.

Quadrini, V., 2000. Entrepreneurship, saving, and social mobility. Rev. Econ. Dyn. 3, 1–40.

Quadrini, V., Rios-Rull, J.V., 2015. Inequality in macroeconomics. In: Atkinson, A.B., Bourguignon, F.J. (Eds.), Handbook of Income Distribution, vol. 2B. North Holland, Amsterdam, pp. 1229–1302.

Ravn, M., Sterk, V., 2013. Job uncertainty and deep recessions. Working paper.

Rietz, T., 1988. The equity risk premium: a solution. J. Monet. Econ. 22, 117–131.

Shimer, R., 2005. The cyclical behavior of equilibrium unemployment and vacancies. Am. Econ. Rev. 95, 25–49.

Skinner, J., 1987. A superior measure of consumption from the panel study of income dynamics. Econ. Lett. 23, 213–216.

Smith, M., Tonetti, C., 2014. A bayesian approach to imputing a consumption-income panel using the PSID and CEX. Working paper.

Storesletten, K., Telmer, C., Yaron, A., 2004a. Cyclical dynamics in idiosyncratic labor market risk. J. Polit. Econ. 112, 695–717.

Storesletten, K., Telmer, C., Yaron, A., 2004b. Consumption and risk sharing over the life cycle. J. Monet. Econ. 51, 609–633.

Storesletten, K., Telmer, C., Yaron, A., 2007. Asset pricing with idiosyncratic risk and overlapping generations. Rev. Econ. Dyn. 10, 519–548.

Wong, A., 2015. Population aging and the transmission of monetary policy to consumption. Working paper.

Young, E., 2010. Solving the incomplete markets model with aggregate uncertainty using the Krusell-Smith algorithm and non-stochastic simulations. J. Econ. Dyn. Control. 34, 36–41.

CHAPTER 12

Natural Experiments in Macroeconomics

N. Fuchs-Schündeln[*,†], **T.A. Hassan**[†,‡,§]

[*]Goethe University Frankfurt, Frankfurt, Germany
[†]CEPR, London, United Kingdom
[‡]University of Chicago, Chicago, IL, United States
[§]NBER, Cambridge, MA, United States

Contents

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.008

Abstract

A growing literature relies on natural experiments to establish causal effects in macroeconomics. In diverse applications, natural experiments have been used to verify underlying assumptions of conventional models, quantify specific model parameters, and identify mechanisms that have major effects on macroeconomic quantities but are absent from conventional models. We discuss and compare the use of natural experiments across these different applications and summarize what they have taught us about such diverse subjects as the validity of the Permanent Income Hypothesis, the size of the fiscal multiplier, and about the effects of institutions, social structure, and culture on economic growth. We also outline challenges for future work in each of these fields, give guidance for identifying useful natural experiments, and discuss the strengths and weaknesses of the approach.

Keywords

Permanent income hypothesis, Fiscal multiplier, Institutions, Social ties, Networks, Social structure, Civic capital, Trust, Multiple equilibria

JEL Classification Codes

C1, C9, E21, E62, H31, O11, O14, O43, O50

1. INTRODUCTION

Establishing causality is a major challenge in economics, especially in macroeconomics, where the direction of various important causal relationships is widely discussed, as illustrated, for example, by large-scale debates about the causal effects of monetary and fiscal policies. Most empirical applications of macroeconomic models focus on matching conditional correlations and improving the fit of models to a set of data moments. Despite substantial advances in this area in recent years, these conditional correlations often cannot identify causal chains. For example, New Keynesian models and real business cycle models can match similar sets of conditional correlations but have very different predictions about the causal effects of fiscal or monetary policies. This lack of identification of clear causal channels is especially troubling when it comes to providing policy advice.

In applied microeconomic fields, causality is often established by designing laboratory or field experiments. In these types of experiments, the researcher consciously influences the economic environment in a way that allows the establishment of causality. The most prevalent and clearest method in this spirit is to randomly allocate agents into a treatment group and a control group, and then analyze the effect of the treatment by directly comparing the relevant outcome variables between both groups, or the change in the outcome variables of both groups coinciding with the introduction of the treatment in a difference-in-differences approach. Field experiments randomize treatment in a real-world economic environment, whereas laboratory experiments do so in a controlled environment. Both methods are mostly unavailable to macroeconomists for fairly obvious reasons. Because macroeconomics deals with phenomena that affect the economy at large (eg, economic growth, unemployment, monetary policy, fiscal policy), any field interventions would be very expensive and would have far-reaching consequences because they cannot easily be targeted at a specific small group, making it unlikely that anyone would agree to carry them out. Bringing key features of the economic environment into the laboratory is also complicated in macroeconomics, where the interplay of different agents and markets often plays a key role (see Duffy, 2008 for a survey of laboratory experiments in macroeconomics).

Natural experiments are an alternative to field and laboratory experiments. For the purposes of our discussion, *we define natural experiments as historical episodes that provide observable, quasi-random variation in treatment subject to a plausible identifying assumption.* The "natural" in natural experiments indicates a researcher did not consciously design the episode to be analyzed, but can nevertheless use it to learn about causal relationships. The episode under consideration can be a policy intervention carried out by policy makers (eg, changes in the tax law), historical episodes that go beyond simple policy measures (eg, the fall of Communism), or a so-called "natural natural" experiment that arises from natural circumstances (rainfall, earth quakes, etc.). Maybe the most widely exploited natural experiment in the macroeconomics literature is the German separation in 1949 and subsequent reunification in 1989. This episode split a homogeneous population into two parts that lived under vastly different economic and political systems with minimal contact between them, only to be reunited 40 years later. Importantly, one can argue this split was exogenous to preferences, economic conditions, and other factors that would directly predict different economic outcomes after reunification. Thus, the assignment of an individual to East or West Germany at the date of separation can be considered random, as in a field experiment. At the same time, vast micro and macro data are available to analyze the episode. Fuchs-Schündeln and Schündeln (2005) first used this experiment to study the self-selection into occupations according to risk aversion and its effect on precautionary savings. Later applications have studied diverse subjects ranging from endogenous preferences for economic policies (Alesina and Fuchs-Schündeln, 2007) and the importance of market access (Redding and Sturm, 2008) to the economic impact of social ties (Burchardi and Hassan, 2013).

Whereas the main task of a researcher carrying out a laboratory or field experiment lies in designing it in a way that allows causal inference, the main task of a researcher analyzing a natural experiment lies in arguing that in fact the historical episode under consideration resembles an experiment, and in dealing with weaknesses of the ex-post experimental setup that one would have avoided a priori in a designed experiment.

To show the episode under consideration resembles an experiment, identifying valid treatment and control groups, that is, arguing the treatment is in fact randomly assigned, is crucial. Establishing such quasi-random treatment requires showing that two groups are comparable along all dimensions relevant for the outcome variable except the one involving the treatment. The methods used to do this are often adapted from the microeconometric literature on field and laboratory experiments.

The goal of this chapter is to acquaint the reader with the use of natural experiments in macroeconomics, summarize what we have learned from them so far, and distill what makes a successful application of a natural experiment to answer a macroeconomic question. We provide in the conclusion of this chapter a summary of common features that distinguish successful papers that rely on the use of natural experiments. Although every natural experiment is different and thus leads to different challenges, these features can serve as guidelines for future papers. Moreover, we discuss the embedding of natural experiments into structural models as a promising general avenue for future research, and point out limitations in the use of natural experiments.

Rather than attempt to cover all papers in macroeconomics that feature natural experiments (which would be a formidable task), we instead select three specific lines of enquiry that use natural experiments for three different purposes: to verify underlying model premises (verification), to quantify specific policy parameters (quantification), and to identify causal mechanisms that operate outside conventional models (identification).

The first line is the literature on the Permanent Income Hypothesis. In contrast to the simple Keynesian consumption theory, the Permanent Income Hypothesis assumes agents are rational and forward looking when making their consumption decisions. Therefore, in addition to current income and current assets, the expected value of future income plays a role in the optimal consumption choice today. This forward-looking behavior can be subjected to a simple test using a preannounced income change: the household should adjust consumption as soon as information about the future income change arrives. By contrast, consumption growth should be unaffected at the time of the implementation of the income change, given that the household knew about it in advance. In this literature, natural experiments serve to identify such preannounced income changes. A finding that households adjust their consumption at the time of implementation of the preannounced income change casts doubts on the fundamental assumption of most micro-founded macroeconomic models that agents are forward looking in their decision making.

The second line is the literature striving to quantify the fiscal multiplier. The fiscal multiplier is one of the most important policy parameters in the macroeconomics literature. Can the government stimulate the economy via government spending or tax policies? If yes, how large is the effect of a given fiscal policy on GDP per capita? The main challenge in the estimation of the fiscal multiplier lies in identifying changes in fiscal policies that are not motivated by business-cycle considerations. In this context, researchers specifically use natural experiments to identify such exogenous changes in government spending.

These first two lines of literature rely not only on natural experiments, but also on other approaches, for example, instrumental variables approaches in which the instruments are not historical episodes, or vector autoregression (VAR) models with exclusion restrictions. By contrast, the third line of the literature relies almost exclusively on natural experiments to identify the fundamental causes of growth. The goal of this literature is to identify mechanisms that are absent from standard macroeconomic models. What can explain the vastly different GDP per capita levels across poor and rich countries? Standard growth models point to human or physical capital accumulation or R&D investment as explanations, but these factors are proximate rather than fundamental causes of growth: why have some countries invested much more than others? The literature on the fundamental causes of growth identifies institutions, social structure, and culture as such fundamental causes. All three of these concepts are largely absent from conventional models of economic growth. Moreover, multiple equilibria can lead to different growth paths despite common initial conditions. Empirically analyzing the fundamental causes of growth is intimately linked to using natural experiments: the "historic episodes" are truly historic here in the sense that they typically come from the distant past and are used to establish causal links by providing quasi-random variation in institutions, social structure, or culture across countries, regions, or time.

Within each of the three lines of literature, we again do not attempt to survey the entire literature on the topic but instead focus on showing how different authors use natural experiments to address research questions arising within each of the three specific contexts, by verifying, quantifying, or identifying causal mechanisms. A common theme across almost all of these applications is that the econometric methods used are fairly simple applications of standard methods, such as OLS, instrumental variables, regression discontinuity, or fixed-effects estimators. Instead, the complexity of many of these papers lies in identifying the episode that generates quasi-random variation, and appropriately dealing with any flaws in nature's experimental design. In this sense, the most crucial ingredient of many papers using natural experiments is the appropriate statement and defense of an identifying assumption, which is the focus of our discussion.

This chapter has two target audiences: the first is researchers with a solid background in applied econometrics who are considering studying a natural experiment in any area of macroeconomics. We hope the juxtaposition of natural experiments used in different areas will generate ideas for intellectual arbitrage for this group. In each of the areas that we cover, we also attempt to point out the research frontier in terms of method and substance, and often explicitly point out important avenues for future research. The second target audience is researchers in mainstream macroeconomics. With this group in mind, we attempt to summarize what natural experiments have taught us about the Permanent

Income Hypothesis, the fiscal multiplier, and the fundamental causes of macroeconomic growth, in the hope that this summary will help direct future theoretical research.

A set of slides that develops the material covered in this chapter in two 90-minute lectures is available on the authors' websites.

2. VERIFICATION: THE PERMANENT INCOME HYPOTHESIS

Natural experiments can be used in macroeconomics to test the validity of major underlying model assumptions. This is done in the use of natural experiments to test the validity of the Permanent Income Hypothesis. The Permanent Income Hypothesis, as developed by Friedman (1957), contrasts with the simple Keynesian consumption theory, which postulates that consumption depends on current income only and is equal to a nonincreasing fraction of current income. To the present day, the Permanent Income Hypothesis is the major building block of modern consumption theory, for example, the life cycle theory, the precautionary savings theory, and also behavioral consumption models involving hyperbolic discounting. The most important insight of the Permanent Income Hypothesis is that individuals are rational and forward looking when making their consumption decisions over the life cycle.

According to the Permanent Income Hypothesis, individual i solves a utility maximization problem of the form

$$\max_{\left\{C_{i,t+j}\right\}_{j=0}^{\infty}} E_t \sum_{j=0}^{\infty} \beta^j u\left(C_{i,t+j}\right) \tag{1}$$

subject to the intertemporal budget constraint

$$\sum_{j=0}^{\infty}\left(\frac{1}{1+r}\right)^j C_{i,t+j} = A_{i,t} + \sum_{j=0}^{\infty}\left(\frac{1}{1+r}\right)^j Y_{i,t+j}, \tag{2}$$

where $C_{i,t}$ is consumption of individual i in period t, β is the discount factor, r is the interest rate, $A_{i,t}$ are initial assets in period t, $Y_{i,t}$ is income in period t, and E_t is the expectations operator conditional on information available at time t. For simplicity, let us assume $\beta(1 + r) = 1$. Also for simplicity, let's assume for now that the utility function takes the quadratic form, such that certainty equivalence holds:

$$u\left(C_{i,t+j}\right) = C_{i,t+j} - \frac{\alpha}{2} C_{i,t+j}^2. \tag{3}$$

This simple model has several powerful implications. Most importantly, consumption is not a function only of current income. Instead, it also depends on current assets and expected future income, and is in fact equal to permanent income. Permanent income is defined as the annuity value of total net worth, which is the sum of current assets and the expected discounted net present value of all future income streams:

$$C_{i,t} = \frac{r}{1+r}\left[A_{i,t} + E_t\left(\sum_{j=0}^{\infty}\left(\frac{1}{1+r}\right)^j Y_{i,t+j}\right)\right]. \tag{4}$$

Because the expected discounted net present value of future income enters the optimal consumption decision of an individual, optimal consumption will change whenever new relevant information arrives. Conversely, any *anticipated* change in income will not affect optimal consumption. Consumption growth depends only on changes in the information set between periods t and $t+1$. Thus, we have

$$\Delta C_{i,t+1} = \frac{r}{1+r}\left[E_{t+1}\left(\sum_{j=0}^{\infty}\left(\frac{1}{1+r}\right)^j Y_{i,t+j+1}\right) - E_t\left(\sum_{j=0}^{\infty}\left(\frac{1}{1+r}\right)^j Y_{i,t+j+1}\right)\right] \tag{5}$$

and specifically

$$\Delta C_{i,t+1} = 0 \text{ if } E_{t+1} = E_t. \tag{6}$$

Eq. (6) holds independent of the form of the utility function used in (1), as long as the desired consumption path is flat. The predictions from Eqs. (5) and (6) can be tested by analyzing the reaction of consumption to anticipated and unanticipated income changes in the data. The empirical challenge lies in identifying in the data whether the individual anticipated any observed income change, and natural experiments are used to identify clearly unexpected or clearly anticipated income changes.

We start out describing the few papers analyzing the reaction of consumption to unexpected income shocks. The literature on the reaction of consumption to anticipated income changes is much larger, for reasons described below, and we will use this literature to gain more insights into the specifics of the use of natural experiments.

2.1 Reaction of Consumption to Unexpected Income Shocks

Only a few papers test whether consumption responds to unanticipated income shocks as predicted by Eq. (5). The reason is that the specific optimal reaction of consumption to an income shock depends among other things on the nature of the shock (whether it is temporary or permanent), on the age of the recipient (if we deviate from an infinite horizon assumption and instead employ a life-cycle setup), and on the functional form of the utility function, which in a more realistic setup might involve prudence from part of the household, such that households build a buffer stock of savings to partly self-insure against future income fluctuations in the absence of perfect insurance.

2.1.1 Unexpected Temporary Income Shocks

If we maintain the assumption of a quadratic utility function, and if an unexpected income change, that is, an income shock, is a strict one-time temporary income change, Eq. (5) reduces to

$$\Delta C_{i,t+1} = \frac{r}{1+r}[Y_{i,t+1} - E_t(Y_{i,t+1})]; \tag{7}$$

that is, the optimal consumption change is equal to the annuity value of the unexpected income change. Thus, as a generalization of this prediction, the optimal consumption change after a temporary income shock clearly should be small. One therefore needs large temporary income changes in the data in order to identify the response of consumption.

A very early paper testing this prediction is Kreinin (1961), whose analysis was later supported by further evidence by Landsberger (1966). Kreinin (1961) uses the 1957/58 Israeli Survey of Family Savings to analyze how Israeli households spent one-time restitution payments from Germany, which around 4% of urban Israeli households received during the year of the survey. He finds that Israeli households saved approximately 85% of the restitution payments, which on average amounted to close to one annual disposable income.[a] This behavior seems roughly in line with a small response of consumption to the temporary income change.

Imbens et al. (2001) and Kuhn et al. (2011) analyze the consumption of lottery winners. Lottery wins are historical episodes that clearly identify random large temporary income shocks, and can as such be seen as natural experiments. Kuhn et al. (2011) compare consumption of Dutch lottery winners and nonwinners.[b] The lottery wins in their episode amount to 12,500 Euros, which is equal to eight monthly average household incomes in the Netherlands. In line with the Permanent Income Hypothesis, Kuhn et al. (2011) find that nondurable consumption does not increase significantly after a lottery win, but durable expenditures increase somewhat. Imbens et al. (2001) analyze significantly larger lottery wins, which are reimbursed over 20 years, and find that the increase in savings after a win is in line with the life cycle hypothesis. The authors of both studies collect their own data by sending out questionnaires to lottery winners and a sample of nonwinners. The final sample sizes are then comparatively small, with 220 lottery winners in Kuhn et al. (2011), and 340 in Imbens et al. (2001).

Brueckner and Gradstein (2013) take a macroeconomic approach to analyze the response of consumption to unexpected temporary income shocks. Exploiting the fact that rainfall is a significant driver of annual aggregate output in sub-Saharan African countries, and that annual variations in rainfall are random and unexpected, they use rainfall as an instrument for aggregate output in a regression that analyzes the reaction of aggregate private consumption to aggregate output. They estimate a marginal propensity to consume out of temporary output shocks that is not significantly different from 0, with

[a] By contrast, Bodkin (1959) finds that windfall incomes of National Service Life Insurance dividends paid out to US veterans were largely consumed. However, these windfalls amounted to, on average, only around 5% of annual disposable income.

[b] They also analyze social effects in a partial population design.

a point estimate of 0.2. Thus, similar to the studies relying on micro data, they find evidence of significant consumption smoothing of temporary income shocks.

2.1.2 An Unexpected Permanent Income Shock: the Natural Experiment of German Reunification

Germany's separation and subsequent reunification constitute in many ways a perfect natural experiment. A country with a common history is split into two parts and the two populations live under very different economic and political systems for 40 years before being reunified. Importantly, it can confidently be argued that the separation of Germany was exogenous to the preferences of the underlying populations and the economic conditions in East and West at the time. The exact location of the border was largely determined by the position of the allied forces at the end of the war, which in turn was partly determined by the geographic location of the allies vis-a-vis Germany. To put it bluntly: if the Soviet Union would have been located to the West of Germany, some western part would have been socialist for 40 years. That the location of the East–West border can be considered random is best documented in the paper by Alesina and Fuchs-Schündeln (2007), who provide an overview of the economic and political situation in Germany before World War II, and show that no marked differences existed between East and West prior to separation. Based on this evidence, West Germans can be taken as a control group for East Germans, and economic conditions of East Germans at reunification, resulting from living under the socialist system of the former German Democratic Republic for 40 years, can be considered exogenous with respect to the new economic system after reunification, since German Reunification was an unexpected surprise event. This is a large-scale experiment, affecting close to 20 million people in East Germany in a multitude of dimensions.

German Reunification has thus been used in a number of studies in the last two decades to analyze different questions. The first paper using German Reunification as a natural experiment is Fuchs-Schündeln and Schündeln (2005), who analyze self-selection in occupational choice according to risk preferences and its effects on estimates of precautionary savings. Redding and Sturm (2008) study the role of market access, and Redding et al. (2011) and Ahlfeldt et al. (2015) focus on industrial location choices. Gebhardt (2013) uses German reunification as a natural experiment to analyze the effect of ownership on relationship specific investment in the housing market, and Bursztyn and Cantoni (2016) to investigate the effect of television advertisement on consumption. The studies by Alesina and Fuchs-Schündeln (2007), analyzing endogenous preferences for redistribution, and Burchardi and Hassan (2013), studying the effect of social ties on growth, are described below and also rely on the natural experiment of German reunification.

In the context of the Permanent Income Hypothesis, Fuchs-Schündeln (2008) exploits German Reunification as a large positive permanent income shock for East

Germans. This permanent income shock is embedded into a structural life cycle model of consumption. This is one of the very few papers which combine a structural model in macroeconomics with a natural experiment.[c] As in any structural model, this implies making assumptions about functional forms and calibrating the model carefully. Yet, it has the advantage that one can talk about the match between quantitative model implications and the data, and can distill the relative importance of different model components.

The life cycle model in Fuchs-Schündeln (2008) incorporates a retirement saving motive, a precautionary saving motive due to income risk and an exogenous liquidity constraint, and deterministically changing household size over the life cycle. West German life cycles play out in this model context from start to end, but East German households enter the new economic model environment in 1990 at a certain age. At this point in time, they are endowed with an exogenous wealth level, which is taken as the cohort-specific East–West wealth ratio in 1992 from the data. Importantly for the predictions of the model, the East–West wealth ratio at reunification was very low (lower than the East–West income ratio), which is especially true for older cohorts closer to retirement. From that point on, East Germans also live in this new economic model environment. Life-cycle income growth, income risk, and changing household sizes are calibrated separately for East and West Germans.

The calibrated model is able to qualitatively and quantitatively match three stylized features of East and West German saving rates after reunification: (i) East Germans have higher saving rates than West Germans; (ii) this East–West saving rate difference is increasing in age at reunification; that is, it is larger for older birth cohorts than for younger birth cohorts; and (iii) for every birth cohort, this difference is declining over time, with full convergence of saving rates within roughly 10 years. The higher East German saving rates after reunification are a result of their low initial wealth levels, which leave them unprepared for the new economic environment in terms of both precautionary and retirement savings. The East–West difference in saving rates is especially large for older cohorts, because older cohorts of East Germans are least prepared for the new environment: their wealth position relative to their West German counterparts is especially low, and they have less time left over their working life to accumulate more wealth through higher saving rates. The rapid convergence of East German saving rates toward West German levels is the stylized feature that allows for differentiation between the different components of the life cycle model. A precautionary savings motive is essential to replicate this feature, because precautionary savings imply that saving rates decrease as wealth levels approach the target level of wealth from below.

The demographic developments after reunification alone would actually predict rising East German saving rates for younger cohorts, running counter to the empirical evidence. Disentangling a precautionary saving motive from a demographic saving motive based on

[c] This approach is more common in other fields, see, eg, the paper by Ahlfeldt et al. (2015).

changing household size over the life cycle is difficult in a standard setting, since both saving motives predict a hump-shaped consumption path over the life cycle. In the context of the natural experiment of German reunification, however, both saving motives lead to opposite predictions for the saving behavior of East Germans relative to West Germans. The paper concludes that East Germans react according to the predictions of the life cycle model after the large shock of German Reunification, despite being confronted with entirely new economic conditions, and that a precautionary saving motive is essential for replicating the data. The first conclusion is in line with the conclusions of the other studies analyzing large temporary income shocks. The second conclusion is only possible in a structural model, pointing to the advantages of the approach used in this paper. Relying on a structural model, one can go beyond analyzing main model predictions to analyzing the importance of different specific model components.

2.2 Reaction of Consumption to Expected Income Changes

In this section, we describe the literature using natural experiments to test the prediction of the Permanent Income Hypothesis that consumption growth should be insensitive to preannounced income changes, as specified in Eq. (6). This is a very large literature: Table A.1 lists 25 published studies directly testing this prediction, and six further studies related to it in some way. We first focus on the methodological side by describing the use of natural experiments, then discussing in Section 2.2.1 different ways to support the random treatment assumption in these studies, and next analyzing how the presence of liquidity constraints modifies the predictions of the theory, and how the papers deal with liquidity constraints. Section 2.2.3 then turns away from the methodolgy to focus on the findings of the studies, and Section 2.2.4 tries to reconcile these sometimes contradictory findings by organizing them along two lines: the size of the income change and the repetitiveness of the episode under study.

The second implication of the Permanent Income Hypothesis—that an *anticipated* income change should not lead to a change in consumption—has the advantage of holding independently of the concrete setup of the problem. In particular, it holds also under functional forms of the utility function other than the quadratic one (eg, under constant relative risk aversion), independent of the age of the individual in a life-cycle setup, and independent of the permanency of the income change at hand. This prediction can be tested if the econometrician knows that an observed income change was anticipated; that is, $Y_{t+1} \neq Y_t$, but $E_{t+1} = E_t$. The null hypothesis would then be that $\Delta C_{t+1} = 0$ and can be tested against the alternative $\Delta C_{t+1} \neq 0$ in a simple reduced-form regression of the form

$$\Delta C_{i,t+1} = \alpha + \beta \Delta Y_{i,t+1}^{expected} + \gamma' \Delta X_{i,t+1} + \epsilon_{i,t+1}, \tag{8}$$

where X is a vector containing any characteristics that are relevant for consumption and might have changed over time, for example, age and household size. The identifying assumption is that the error term is uncorrelated with the expected income change, that is, $Cov[\Delta Y_{i,t+1}^{expected}, \epsilon_{i,t+1}] = 0$, meaning no unobserved variables are correlated with the expected income change and the consumption change. The Permanent Income Hypothesis states that $\beta = 0$. If the underlying assumption of rational expectations and forward looking behavior is violated, we would expect that $\beta \neq 0$, and specifically that $\beta > 0$ under the Keynesian consumption theory.

Running this regression is easy if an expected income change can be directly observed in the data, that is, if we know the underlying assumption $E_{t+1} = E_t$ holds. However, in general, whether any observed income change in the data was expected or unexpected by the individual is unclear. A common way to run this regression in the macro literature relying on aggregate consumption data involves the use of instruments. For example, Ludvigson and Michaelides (2001) regress quarterly consumption changes on quarterly income changes, instrumenting income changes with their own lags. Carroll and Summers (1991) run similar regressions on international data, again instrumenting with lags of income growth. However, at the micro level, to which the theory applies, finding a suitable instrument is much harder.

A more elegant and convincing way to run this regression on the micro level is to exploit a natural experiment. Natural experiments in this context are clear historical episodes in which we know that an income change occurred, and that it was preannounced and thus anticipated by the households. Typical income changes of this kind analyzed in the literature are associated with taxation (tax rebates, tax refunds, changes in tax laws, etc.), wages (wage payment schedule, wage changes, social security receipts), and committed consumption (college cost, mortgage payments, etc.). All these changes have in common that they are clearly announced some time in advance, and thus the recipient anticipates them. The Permanent Income Hypothesis predicts that households should adjust their consumption at announcement of the income change. The size of the optimal consumption adjustment at announcement depends among other things on the expectations about the exact nature of the income change and is therefore hard to gauge, as in the papers described in Section 2.1. By contrast, testing the prediction that consumption should not react when the preannounced income change actually happens is easy.

In a more general sense, one can think of the test for whether $\beta = 0$ in Eq. (8) as a general test of the validity of the rational expectation assumption in consumption decisions. We might not care from either a macro or micro point of view whether households adjust their consumption at the announcement or the implementation of an income change, because both typically happen within a short period of time in the natural experiments analyzed in the literature. However, for welfare purposes, whether households build rational expectations and are forward looking when deciding how much to consume and how much to save matters tremendously. For example, to save appropriately for retirement, households have to understand the income process over their life cycle early on and act accordingly.

2.2.1 Random Treatment: Determining an Appropriate Control Group

The estimation of Eq. (8) using a natural experiment to establish that an income change was anticipated still faces some challenges. Importantly, Eq. (8) can only be estimated consistently if the error term is uncorrelated with the preannounced income change; that is, $Cov[\Delta Y_{t+1}^{expected}, \epsilon_{t+1}] = 0$. Otherwise, the preannounced income change and the consumption change would be spuriously correlated due to omitted variables.

One important feature that could lead to correlation between the error term and the preannounced income change could be seasonality effects. For example, workers in many countries receive a 13th salary in the month of December, leading to a preannounced change in monthly income between November and December. At the same time, expenditures increase in December because of holiday shopping. This leads to a spurious correlation between the preannounced income change and the consumption change. The income change is endogenous because the 13th salary in December was established precisely because firms recognized the higher average household expenditure in December.

In the spirit of an experimental setup, a valid control group can overcome this problem. If the above-mentioned preannounced income change exhibits temporal variation, that is, if it does not occur in the same month for all households, then variation is present in the timing of the treatment, and one can include monthly dummies to account for seasonality in expenditures directly. The same applies if the preannounced income change happens in different months in different years, though in that case, one has to argue that expenditure seasonality should be the same year by year, for example by analyzing whether major events usually causing increases in expenditure, like public holidays or vacations, happen in the same months every year. Variation in the individual amount of the preannounced income change relative to permanent income could help, but only if one could reasonably argue that this variation is exogenous to any desired seasonality in expenditure.

In the ideal experiment, one group does not receive any preannounced income change, and another one does, and both should be comparable along all other observable and unobservable characteristics, including preferences that lead to consumption seasonality. In that case, one can think of the first group as the "control" group and of the second group as the "treatment" group. Here, the natural experiment is very close to a designed field or laboratory experiment: two groups exist, one of which is quasi-randomly treated and the other one not, and the behavior of both groups is compared. The analysis of consumption changes then corresponds to a difference-in-differences setup. Whereas laboratory or field experiments would be designed to make the assignment into the treatment group explicitly random, the main challenge of a natural experiment is to convincingly argue the randomness of the assignment and thus the appropriateness of the control group. Arguing this point is generally easiest if both groups receive the same treatment, but at randomly different points in time. This distinguishes natural experiments from field or laboratory experiments, which typically leave a control group untreated.[d]

In this section, we describe different methods to determine randomness in treatment. In passing, we also discuss some findings of the papers, which are, however, the focus of Section 2.2.3.

2.2.1.1 Clearly Established Randomness in Treatment

A set of studies that are particularly successful in establishing randomness in the treatment assignment are the papers by Johnson et al. (2006) and Agarwal et al. (2007), who exploit the 2001 Federal Income Tax Rebates as a natural experiment, and the studies by Parker et al. (2013) and others, who analyze the 2008 Economic Stimulus Payments as a natural experiment.[e] In both cases, households received one-time tax rebates in the form of checks sent to them. The media extensively discussed the rebates in advance, and as such, households should have known about them. In addition, for the 2001 Bush tax rebates, households received an individual letter several months in advance stating the specific amount of the rebate.[f] Although the amount received varied little between households, mostly driven by household size and thus not exogenous, nice and clearly exogenous variation exists in the timing of the payments: because sending out all rebate checks on the same day was logistically impossible, the IRS spread out the payments over 10 weeks in 2001 and 9 weeks in 2008, and determined the exact date on which each household would receive the check by the second to last digit of the Social Security Number of the main tax payer, which is randomly assigned. Thus, in these two cases the randomness in the timing of treatment is as clearly established as in any field or laboratory experiment in which the researcher consciously establishes randomness through a lottery. Exploiting this situation, the "treated" group in the above-mentioned studies is the one that randomly receives the rebate in the period under consideration, whereas the "control" group encompasses all other households, which receive the rebate in a different period.[g]

[d] A valid reason for this approach for field or laboratory experiments is the fact that treatment is typically costly for the researcher.

[e] Johnson et al. (2006) and Parker et al. (2013) analyze consumption responses, whereas Agarwal et al. (2007) analyze the response of credit card spending and debt repayment to the 2001 federal income tax rebates. The 2008 Economic Stimulus Payments have been exploited by a number of studies, including Broda and Parker (2014) and Parker (2014) analyzing consumption responses, Gross et al. (2014) and Bertrand and Morse (2009) analyzing bankruptcy filing and repayment of payday loans, respectively, and Evans and Moore (2011) and Gross and Tobacman (2014) analyzing mortality and morbidity outcomes. Shapiro and Slemrod (2003) and a series of papers by Sahm et al. (2010, 2012) analyze self-reported propensities to consume and to save out of both rebate episodes. Misra and Surico (2014) analyze heterogeneity in consumption responses to both the 2001 and 2008 stimulus programs.

[f] For the 2008 Economic Stimulus Payment, the letter came only 1 week in advance.

[g] In the case of the 2008 Economic Stimulus Payments, part of the households received not a check but a direct deposit, for which the timing was somewhat different. Thus, the studies using the 2008 Economic Stimulus Payments suffer from larger measurement error than the studies using the 2001 federal income tax rebates, if they cannot determine whether a household received a check by mail or a direct deposit, which most cannot.

The week of rebate receipt is clearly exogenously determined. All of these studies find that household consumption adjusts at receipt of the rebates, in violation of the Permanent Income Hypothesis.

2.2.1.2 The "Narrative Approach"

In the absence of such a clear random treatment assignment, different strategies can be used. For example, Browning and Collado (2001) analyze quarterly consumption growth of Spanish workers who are part of one of two different payment schemes: the standard scheme used in the control group encompasses monthly wage payments of twelve equal amounts over the year, whereas the second payment scheme in the treatment group involves higher payments in the months of June (or July) and December. The payment scheme varies on the plant level, and because workers know which payment scheme their plant follows, workers in the treatment group should perfectly anticipate the unusually high monthly income growth between the months of May and June (or June and July, if the extra payment is in July rather than June) as well as November and December, each followed by a month of unusually low income growth. To test the prediction of the Permanent Income Hypothesis that consumption growth should not react to preannounced income changes, the authors then simply compare seasonal consumption patterns of the treatment group (called "bonus group") and the control group (called "nonbonus group").[h] Fig. 1 shows quarterly income growth per calendar week of both groups on the left, and quarterly expenditure growth on the right.[i] Despite strong differences in the income growth patterns between both groups, the expenditure growth patterns are very similar. Thus, the evidence in Browning and Collado (2001) is in line with the predictions from the Permanent Income Hypothesis.

The major challenge here is to argue about the random assignment of the payment scheme. For example, plants might use the second payment scheme because they know their workers have unusually strong preferences for seasonally high expenditures in June/July and December, for example, due to certain holiday traditions in their region. The authors explicitly discuss this assumption and give some historical account of how the two payment schemes arose. They also show that being part of either payment scheme is not significantly correlated with any observable household characteristics. We call this the "narrative approach," because it relies purely on carefully arguing about exogeneity of the treatment, and ruling out potential alternative stories of endogeneity. Placebo exercises, described below, are useful in this regard. Since Browning and Collado (2001) find that expenditure growth patterns of both groups over the year resemble each other, the argument about exogenous treatment becomes somewhat less important; any

[h] Similarly, Hsieh (2003) compares the seasonal consumption patterns of Alaskans to the seasonal consumption patterns in other US states, and Paxson (1993) compares seasonal consumption patterns of farmers and nonfarmers in Thailand.

[i] Income is measured as average income in the three quarters preceding the interview. While the extra payments are called "bonus," there is no performance component involved.

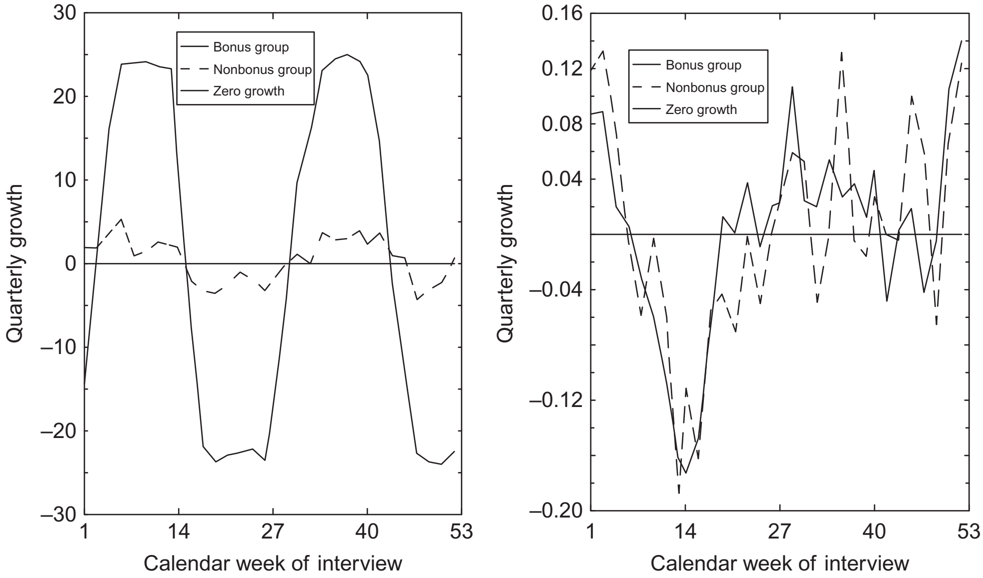


Fig. 1 Quarterly earnings growth (*left*) and quarterly total expenditure growth (*right*) in Browning and Collado (2001).

endogeneity should have led to the observation of stronger seasonal expenditure patterns correlated with the preannounced income changes for the treatment group.

2.2.1.3 Using Different Control Groups and the Matching Approach

Sometimes doubts about exogeneity of the treatment remain after a detailed description and careful analysis of the circumstances leading to treatment vs nontreatment in the "narrative approach." In this case, one can follow different strategies to still establish some confidence into a causal effect. The most basic strategy, followed by many papers, is to establish robustness of the results to the use of different control groups. Consider, for example, the study by Agarwal and Qian (2014b), who analyze the response of consumption and debt repayment to a unique cash pay-out by the government to each adult Singaporean. The pay-out happened at the same time for all eligible individuals, such that no randomness in the timing was present. Although amounts varied across individuals, this variation was not random, because the amount was a function of income and home values. Agarwal and Qian (2014b) use foreigners living in Singapore as a control group: foreigners make up almost 40% of the population living in Singapore and were not eligible to receive the pay-out. They show results of their analysis using this control group, as well as restricting the analysis to Singaporeans and exploiting only the (nonrandom) variation in amounts.

Both approaches clearly have their disadvantages. Specifically, foreigners only constitute a valid control group if their spending patterns are similar to those of Singaporeans in the absence of treatment. In a first step, the authors compare Singaporeans and foreigners along observable characteristics and find some significant differences. To control for these observable differences, they use propensity score matching methods (going back to Rosenbaum and Rubin, 1983) to construct two subsamples of matched treatment and control groups that are comparable across most observable characteristics. Researchers frequently use propensity score matching methods in microeconometric setups in which random treatment cannot be assumed. The basic idea behind a variety of submethods is that one constructs a subsample of the treatment group and a subsample of the control group which are as comparable as possible along a long list of observable characteristics. Importantly, Agarwal and Qian (2014b) also show that both subsamples have comparable seasonal spending patterns prior to the treatment, though this information is not used to construct the subsamples. Agarwal and Qian (2014b) find that Singaporeans increase their consumption already at announcement of the pay-out, and spread the consumption increase over the following 10-month period.

Abdallah and Lastrapes (2012) use a similar approach, analyzing the effect of a preannounced relaxation in the borrowing constraint among Texan home owners in 1997 on Texan retail spending. They start out using two control groups, the first consisting of all other US states except Texas, and the second consisting of all other US states that did not change sales tax rates during the estimation period. They allow for state-specific linear time trends, in order to ensure that a different general time trend in Texan retail spending is not mistakenly attributed to the policy change. In a next step, they also employ a form of matching methodology, specifically, the synthetic control method of Abadie and Gardeazabal (2003) and Abadie et al. (2010), which assigns optimally selected weights to each control group observation in order to minimize the distance between predicted sales in Texas and the control group during the pretreatment period. This study falls into the group of studies analyzing liquidity constraint relaxations (see the discussion in Section 2.2.3), and like all these studies, finds evidence for binding liquidity constraints.

2.2.1.4 The Use of Placebo Exercises

A formal way to gauge the validity of the control group is the use of placebo exercises. The idea of this approach is to define virtual "placebo treatments" and to compare the average effects of these "placebo treatments" to the one of the actual treatment. Consider again the study by Abdallah and Lastrapes (2012). In this study, the "placebo treatment" can be defined as dropping Texas from the analysis and assuming a state other than Texas introduced a similar relaxation of the borrowing constraint at the same point in time when Texas actually did. If one includes all US states in the analysis, one ends up with 49 different "placebo treatments."[j] For each of them, the baseline regression is run, and

[j] In addition, one could specify "placebo treatments" taking place in Texas, but at a different point in time, or even taking place in other states at a different point in time.

the baseline estimate on the true treatment is compared to the distribution of the estimated coefficients β from the placebo treatments. The treatment effect from the baseline regression should be well above the median placebo treatment effect in order to confirm a true effect. This approach can also be applied to control groups built based on a matching method. Gross et al. (2014), Mastrobuoni and Weinberg (2009), and Scholnick (2013) perform similar placebo exercises.

2.2.2 *The Presence of Liquidity Constraints*

As stated above, Eq. (6) holds under different concrete setups, for example, in an infinite or a finite life-time setting, under different assumptions of the functional form of the utility function, and so on. However, one important assumption has to be maintained: the consumer problem laid out in Eqs. (1) and (2) does not contain a liquidity constraint. If liquidity constraints are present and binding, the household will not be able to adjust consumption optimally at the announcement of a future income increase, but only at the implementation of the income increase.

We can deal with this complication in two ways. First, and most convincingly, one can analyze the consumption reaction to a preannounced income *decrease* rather than an increase. The presence of a liquidity constraint does not affect the optimal consumption change triggered by the announcement of a future income decrease: decreasing consumption is always a possibility. Unfortunately, the vast majority of the "natural" situations that researchers can analyze involve income increases rather than decreases. Here, the limitation of natural experiments, which cannot be designed to prevent certain limitations a priori, becomes clear in contrast to self-designed field or laboratory experiments. One paper that does analyze a decrease in income is the study by Souleles (2000). The paper analyzes the change in consumption upon an increase in college expenditure due to a child in the household starting college. Because the college entrance can be foreseen for some time, and college costs are also usually determined in the spring before the start of college, one can think of the increase in college expenditure as a perfectly anticipated increase in committed consumption and therefore as a decrease in net disposable income.[k] Souleles (2000) finds that expenditures on strictly nondurable goods and food do not fall significantly upon the anticipated decrease in net disposable income, or if anything, they fall by a very small amount, depending on the specification. This study thus finds behavior in line with the Permanent Income Hypothesis if households face an anticipated income decrease.

Apart from focusing on income decreases, one can explicitly analyze the importance of liquidity constraints by splitting the sample into potentially constrained and most likely

[k] This decrease in net disposable income is of course endogenous, because parents can choose whether and how much to spend on a child's college education. The paper includes robustness checks instrumenting for college expenditure.

unconstrained households. Because binding liquidity constraints are typically unobservable, this analysis is approximate rather than exact. Consider, for example, the paper by Johnson et al. (2006) cited above. They rely on three different measures that can proxy for liquidity constraints: age, income, and liquid assets, where the assumption is that young households, households with low income and/or a low levels of liquid assets are more likely to be liquidity constrained. Splitting the sample into "high," "medium," and "low" values of the respective variable, they then analyze whether the consumption change of the "low" and thus potentially liquidity constrained group is larger than that of the other groups.[l] Moreover, the coefficient β should be equal to zero for the unconstrained "high" group in the absence of measurement error in measuring liquidity constraints. The evidence points towards liquidity constraints: the "low" group increases consumption more upon receipt of the rebate check than the "high" group, though the difference is not always statistically significant. Moreover, even the "high" group shows a positive consumption response under some measures. These are two common findings in the literature: potentially liquidity constrained groups react stronger upon payment receipt, but groups that are likely not liquidity constrained still react significantly.[m] Misra and Surico (2014) point out that traditional measures of liquidity constraints might miss wealthy hand-to-mouth consumers (Kaplan and Violante, 2014a), who hold most of their assets in illiquid form, and provide evidence that indeed these consumers also react strongly to preannounced income changes. They argue that imposing homogeneous consumption responses not only misses interesting and systematic variation in the data that is potentially linked to liquidity constraints, but can also lead to biased estimates of the average consumption response. Indeed, analyzing measures of liquidity constraints that also incorporate wealthy hand-to-mouth consumers might allow reconciling some of the conflicting evidence on liquidity constraints found so far.

Another set of studies analyzes consumption during the payment cycle. Because income increases from zero to a positive value on the day of wage payment receipt, and then falls to zero again the day after payment receipt, these studies also encompass regular income decreases. These studies typically involve frequencies higher than the monthly one and analyze whether consumption is stable or decreases over the payment cycle (see Gelman et al., 2014; Mastrobuoni and Weinberg, 2009; Shapiro, 2005; Stephens, 2006, and Stephens, 2003). In principle, if one assumes the first payment comes at the end of a "consumption cycle," liquidity constraints could matter. However, if these regular payments have already been received for some time, arguing that households

[l] Part of the literature relying on credit card data can apply more direct measures of liquidity constraints, for example, whether individuals regularly pay interest on their credit card, or how close they are to the credit limits (see Agarwal et al., 2007; Agarwal and Qian, 2014b, and Scholnick, 2013).

[m] Unfortunately, not all studies analyzing liquidity constraints show results testing the latter hypothesis that nonliquidity constrained groups should not react significantly.

could not build up a buffer to smooth variations over the pay cycle is hard. Still, some of these studies employ explicit tests for liquidity constraints as described above and find evidence in favor of liquidity constraints (Gelman et al., 2014; Mastrobuoni and Weinberg, 2009, and Stephens, 2006). Because, in principle, liquidity constraints should not matter in this setup, the evidence in favor of liquidity constraints might indicate that the proxies for liquidity constraints are correlated with other behavioral traits that could drive the excess sensitivity of consumption to preannounced income changes.

2.2.3 Findings

More than two dozen papers test Eq. (8) in various ways. Describing these papers in detail is beyond the scope of this chapter. Nevertheless, Table A.1 at the end of this chapter provides a brief overview listing papers in alphabetical order. The table lists the nature of the specific episode that is analyzed and whether it involves an income increase or decrease, the data source (including country and specific data set used) and sample selection, the main dependent variable and its frequency, whether the paper finds significant evidence against the Permanent Income Hypothesis and what the main result is quantitatively, and finally whether any tests of liquidity constraints are carried out and what their results are. Unfortunately, because the concrete estimations run in each paper are different, one cannot indicate a comparable estimated coefficient β for each study, but we provide the main quantitative result as stated in the respective paper. All but two of the papers involve the use of household or individual data.

Table A.1 distinguishes between three different sets of studies. The first set includes 25 studies that analyze an experiment involving an anticipated change in disposable income, most often because of a direct gross or net income change, sometimes because of a change in payment commitments from mortgages or college expenditures. The dependent variable in these studies is some measure of consumption, which varies from standard measures of nondurable or durable expenditures over caloric intake to credit card spending or retail sales. Most of these studies find evidence against the Permanent Income Hypothesis: consumption growth reacts to the implementation of the preannounced income change. Notable exceptions are the studies by Agarwal and Qian (2014b), Browning and Collado (2001), Coulibaly and Li (2006), Hsieh (2003), Paxson (1993), and Souleles (2000).

The second set includes four studies that analyze the reaction to preannounced relaxations of borrowing constraints. These studies can be seen as direct tests of the presence of binding liquidity constraints: if liquidity constraints are not binding, then any preannounced relaxation of a constraint might lead to consumption reactions at announcement, but not at implementation.[n] If liquidity constraints are, however, currently

[n] Consumption might still react at announcement, because the possibility of liquidity constraints becoming potentially binding in the future affects consumption today.

binding, then any relaxation of the constraint should lead to increased consumption at implementation. The studies either involve an experiment directly relaxing the borrowing constraint and then use a measure of consumption as the dependent variable, or they involve an experiment relying on an expected income change as in the first set of studies but use a measure of loan take-up or bankruptcy filing as the dependent variable. Some of these studies still analyze whether potentially liquidity constrained households react stronger upon implementation than households who are less likely to be liquidity constrained. All of these studies find that liquidity constraints do matter.[o]

The third set includes two studies that deal with experiments that involve temporary price cuts. Under the assumption of forward-looking behavior, expenditures on goods subject to a temporary price cut increase during the period of the price cut, but at the same time, expenditures on these goods decrease before or after the period of the price cut if goods exhibit some degree of durability and the period of the price cut is relatively short.[p] These two studies find different results: Sales tax holidays seem to have long-lasting effects on purchases of some affected goods (Agarwal et al., 2013), whereas the 2009 "Cash for Clunkers" program merely shifted the purchases of new cars in time (Mian and Sufi, 2012).

2.2.4 Violation of Rational Expectations or Need for Model Extension?

A vast majority of the natural experiments investigating the Permanent Income Hypothesis find evidence against it by rejecting the null that $\beta = 0$. This seems to indicate that households are in fact not forward looking when making their consumption decisions, even if they have to take into account only income changes that will occur a few months ahead. How can one then assume that households look many years ahead, as required for retirement planning, saving for childrens' college expenditures, and so on? Thus, one of the major assumptions of the Permanent Income Hypothesis seems to be undermined. The evidence summarized in Section 2.2.2 suggests liquidity constraints can help reconcile theory and evidence, but often a significant consumption response to preannounced income changes can be found even among likely unconstrained households. Apart from analyzing liquidity constraints, the data are rarely rich enough to provide further insights into the sources of the failure of the Permanent Income Hypothesis.[q]

[o] DeFusco (2014) and Agarwal and Qian (2014a) are two recent papers analyzing similar experiments that involve a relaxation of borrowing constraints for home owners in the first case, and an unexpected tightening in the second case.

[p] Expenditures on other goods might also change, depending on the degree of substitutability or complementarity between goods.

[q] The recent study by Parker (2014) is a step in the right direction. It analyzes the spending response to the 2008 Economic Stimulus Payments using data from the Nielsen Consumer Panel, and augments these data with questionnaires that allow the author to draw conclusions about certain personal characteristics such as lack of planning, impatience, and inattention.

Table 1 Studies of the permanent income hypothesis sorted by size and regularity of the income change

	Small	Large
Regular	Aaronson et al. (2012) 0.03% Parker (1999)[a] 0.00038 % Parker (1999)[b] 0.82% Shea (1995) 0.0009%	**Browning and Collado (2001) 2.61%** **Hsieh (2003) 4.79%** **Paxson (1993) –** Souleles (1999) 1.24%
Irregular	Agarwal et al. (2007) 0.22% **Agarwal and Qian (2014b) 0.04%** Broda and Parker (2014) 0.31% **Coulibaly and Li (2006) 0.56%** Johnson et al. (2006) 0.10% Parker et al. (2013) 0.46% Scholnick (2013) 0.45% Souleles (2002) 0.01% Stephens (2008) 0.35%	**Souleles (2000)[c] 5.24%**

Note: Papers written in bold fail to reject the Permanent Income Hypothesis. The number after each study indicates the equivalent variation associated with the respective experiment. The equivalent variation is calculated as described in the text. Table A.1 provides details on the calculation of the equivalent variation for each paper.

[a] Change in social security tax rate

[b] Cap in social security withholding

[c] Because of the absence of suitable expenditure and income data, the equivalent variation is calculated with price-adjusted average quarterly spending from Johnson et al. (2006).

Table 1 distinguishes the existing natural experiment studies along two lines: how *large* the analyzed income change is, and whether it happens *regularly* over the life cycle. We consider an income change as *regular* if it is a repeated phenomenon that occurs to an individual several times over the life cycle, for example, tax refunds or payment schemes that double the income every year in July and December. On the other hand, unique government interventions such as tax rebates due to a fiscal stimulus program, or the running out of mortgage or college payments are considered irregular events. To classify an episode as large or small, we resort to the equivalent variation as a measure of the welfare loss associated with a certain behavior.[r] Specifically, we compare two hypothetical consumers over the course of 1 year only, considering monthly consumption,[s] and assuming additive separability of monthly utility and no discounting: the first "rational" consumer smoothes a preannounced income change x over the course of 1 year. This behavior is obviously not optimal, because optimality would call for smoothing the income change

[r] The distinction in small and large shocks has already been suggested by, among others, Browning and Collado (2001), Hsieh (2003), and Jappelli and Pistaferri (2010).

[s] The time unit is a month rather than a year because most experiments use monthly data and involve an episode of a predicted income change in a specific month, that is, most papers in the literature follow this timing logic.

over the entire life cycle, but it is a good approximation for those regular income changes that occur once a year, and otherwise provides a lower bound of the welfare losses. We calculate the utility of this consumer over a year as $U^{rational}(c) = 12 * u\left(y + \frac{x}{12}\right)$, where y is regular monthly consumption and x is the extra amount received in the experiment. The second "hand-to-mouth" consumer has the same baseline income as the "rational" consumer, but consumes the extra amount x analyzed in the specific episode entirely in the month of receipt rather than spreading it evenly over 12 months; that is, her utility is $U^{hand-to-mouth}(c) = 11 * u(y) + 1 * u(y + x)$. We then calculate the equivalent variation as the monthly consumption amount z we would have to add to the consumption of the "hand-to-mouth" consumer to make her as well off as the "rational" consumer, expressed as a percentage of regular monthly consumption.[t] In these calculations, we assume a constant relative risk aversion utility function with a risk aversion parameter of 2. We consider an experiment as *large* if the equivalent variation amounts to more than 1%. The spirit of this exercise and the specific threshold of 1% are in line with the study done by Chetty (2012), who analyzes bounds on labor supply elasticities, allowing for adjustment costs or inattention resulting in households not reacting to tax changes, as long as the associated utility loss amounts to less than 1% in a life-cycle setup. Table A.1 explains in detail which values are used to calculate the equivalent variation for each study.

In table 1, papers that fail to reject the Permanent Income Hypothesis are written in bold. As the table shows, four of the six studies that do not reject the Permanent Income Hypothesis analyze large income changes, three of these analyzing income changes that occur repeatedly over the life cycle, and one analyzing an irregular income change.[u] Among the studies analyzing large income changes, only one (Souleles, 1999, who analyzes tax refunds) rejects the Permanent Income Hypothesis. This study involves an experiment associated with an equivalent variation barely exceeding 1%, being the smallest among the "large" studies.

The remaining two studies that find support for the Permanent Income Hypothesis analyze small, irregular changes. Coulibaly and Li (2006) find that home owners smooth consumption over their last mortgage payment, after which disposable income increases. The episode is characterized as small, because the last mortgage payment is typically not high and these households are relatively well off; from a life-cycle perspective, mortgage payments, however, constitute substantial consumption commitments and thus reductions in disposable income. The last study, by Agarwal and Qian (2014b), analyzes the

[t] In other words, z solves $11 * u[y + z] + 1 * u[y + x + z] = 12 * u\left[y + \frac{x}{12}\right]$, and the equivalent variation is calculated as $EV = z/y$.

[u] Paxson (1993) does not provide enough information to calculate the equivalent variation as in the other studies. Still, it is clear that the utility loss for farmers not smoothing income fluctuations over the year would be large in the sense of an equivalent variation exceeding 1%.

2011 Singaporean Growth Dividends, which amounted to around 500 USD per individual. This study is different from the others in that it explicitly analyzes the consumption reaction at implementation *and* announcement, and finds that consumption increases already at announcement, but remains higher for almost 1 year.

Taken together, the evidence appears to imply that households tend to behave consistently with the Permanent Income Hypothesis when the stakes are high, that is, when dealing with large or repeated changes in their income. A simple way to rationalize the different results of the papers may be to consider models that allow for monetary or psychological adjustment costs of reoptimization. Moreover, learning on the part of the consumer might play a role. Monetary or psychological adjustment costs would point to near-rationality, as defined, for example, by Cochrane (1989). The evidence in favor of the Permanent Income Hypothesis coming from the studies analyzing large income changes is in line with the natural experiment studies analyzing income shocks in Section 2.1, which all look at large shocks to income and do not find evidence against the Permanent Income Hypothesis. Moreover, in the two studies that analyze temporary price cuts, the study analyzing large price cuts (Mian and Sufi, 2012) finds evidence in favor of rational behavior, whereas the paper analyzing relatively small price cuts (Agarwal et al., 2013) finds evidence against rationality. Near-rationality is in the spirit of the concept of inattentive consumers as in Reis (2006) or of inaction inertia as discussed in, for example, Anderson (2003).[v] In Reis (2006), households with high planning costs become inattentive savers, which live according to a saving plan and let consumption absorb all income changes that are not large enough to trigger a reoptimization. Small income changes might fall into this category, and thus an inattentive saver would not adjust his or her consumption at arrival of new information on a future small income change, but rather consume the extra income at arrival. The announcement of a large future income change would instead trigger reoptimization.[w]

Perhaps the most convincing evidence in favor of adjustment costs or near rationality comes from Hsieh (2003). Hsieh (2003) uses data on Alaskan households from the consumer expenditure survey in order to analyze two natural experiments on the same set of households. The first experiment involves tax refunds also analyzed by Souleles (1999). These refunds are anticipated, because the taxpayer has to calculate them when filing the tax return. Around three quarters of all taxpayers receive refunds, and the average refund

[v] Early studies of near-rationality include Akerlof and Yellen (1985) and Mankiw (1985). For a recent study, see Hassan and Mertens (2014).

[w] The survey responses analyzed by Shapiro and Slemrod (2003) do not, however, support the implications of models of inattentive savers or inattentive consumers: in contrast to the model's prediction, individuals who report to target spending are more likely to spend the 2001 tax rebates than individuals who don't target spending, whereas individuals who report to target saving show the same propensity to save the rebate as those who don't. Also, the survey evidence by Parker (2014) does not point towards inattentiveness.

on the household level amounts to 700 USD to 850 USD (in 1982–84 USD, see Souleles, 1999). Hsieh (2003) finds that Alaskan households receiving a tax refund consume 28% of it in the quarter of receipt. He then runs a regression on the same set of households, in which he analyzes a different preannounced income change, namely, payments from Alaska's Permanent Fund. The Alaska Permanent Fund redistributes receipts from oil royalties as dividend payments to residents of Alaska. The amount of the payment has been increasing over time and varies between around 300 USD per person in 1984 and almost 2000 USD per person in 2000. Because every resident of Alaska, regardless of income and age, is entitled to this payment, the average household payment is quite high, substantially higher than the average tax refund. Hsieh (2003) finds that the same households that show significant excess sensitivity of consumption to the preannounced income changes caused by tax refunds do not show such excess sensitivity to the preannounced income changes caused by dividends from the Alaska Permanent Fund. This is strong evidence that the size of the welfare cost associated with failing to smooth the income change in question matters, because it comes from exactly the same set of households. Scholnick (2013) also reports direct evidence that the magnitude of the analyzed income change matters. He analyzes the reaction of credit card spending to the predictable changes in disposable income resulting from a household's final mortgage payment. Because the mortgage payment amounts vary by households, also relative to their income, he can analyze whether the size of the income change matters.[x] Indeed, he finds a positive reaction of credit card spending to the income increase after the final mortgage payment, in violation to the prediction of the Permanent Income Hypothesis, but the larger the preannounced income change is, the smaller the reaction.

Summarizing, the literature on the Permanent Income Hypothesis finds that liquidity constraints clearly matter for some households. For households that are not constrained, near-rationality is a likely candidate to explain their excess sensitivity to small anticipated income changes. Faced with large income changes, households seem to react in line with the Permanent Income Hypothesis and are thus forward looking when making their consumption decisions.

3. QUANTIFICATION: THE FISCAL MULTIPLIER

The size of the fiscal multiplier is a highly controversial topic in macroeconomics. The fiscal multiplier measures the size of the output change associated with a change in a fiscal instrument; that is,

[x] One might, however, be worried whether the variation in size is endogenous, and how this endogeneity would affect the estimates. In a recent working paper, Kueng (2015) similarly finds evidence for near-rationality by exploiting variation in the relative size of the income change across households relying on the same experiment as Hsieh (2003).

$$\Delta Y_{t+1} = \alpha + \beta \Delta F_{t+1} + \gamma' \Delta X_{t+1} + \epsilon_{t+1}, \tag{9}$$

where β is the fiscal multiplier, Y is a measure of output, F is a measure of the fiscal instrument, and X is a vector of potential control variables, typically including lagged growth measures.

Although, in principle, running this regression with a time series of macroeconomic data is easy, the macroeconomic literature on fiscal multipliers faces one serious challenge: the change in government spending must be exogenous to economic growth. A standard measure of total government spending is certainly subject to reverse causality, such that the assumption $Cov[\Delta F_{t+1}, \epsilon_{t+1}] = 0$ does not hold. For example, automatic stabilizers such as medicaid and unemployment insurance lead to an increase in fiscal spending precisely when output growth is low, such that $Cov[\Delta F_{t+1}, \epsilon_{t+1}] < 0$, biasing the estimate of β towards zero. On the other hand, procyclical government spending components might exist if governments have limited ability to accumulate debt, in which case $Cov[\Delta F_{t+1}, \epsilon_{t+1}] > 0$, and the estimate of β would be biased upwards.

The macroeconomic literature typically addresses the issue of endogeneity using vector auto regression (VAR) methods imposing identifying restrictions (see, eg, Blanchard and Perotti, 2002 and Mountford and Uhlig, 2009), for example, that government spending does not react to current economic conditions at the quarterly frequency, or relying on the so-called narrative approach, which establishes exogeneity of fiscal policies to current economic conditions based on government records (Romer and Romer, 2010). Relying on natural experiments provides an alternative way of establishing exogeneity: for example, a war initiated by another country may create a natural experiment that causes increased government spending not motivated by current economic conditions in the home country. Although the absence of reverse causality might be easier to argue, the approach still faces some important challenges. First, one needs a critical number of these events over time or geographical variation in order to carry out an empirical analysis. A further hurdle lies in controlling for effects of the "natural experiment" on economic growth that do not play out via government spending. For example, a war might affect patriotism in the home country, potentially increasing the demand for home-produced goods, or a war taking place in the home country likely affects the capital stock. We review two lines of this literature: the first one relies on exogenous variation in military spending, and the second one estimates local fiscal multipliers, relying on different natural experiments.

3.1 Permanent Income Hypothesis Studies and the Fiscal Multiplier

Before analyzing the use of natural experiments in establishing exogeneity of fiscal spending, we want to point out the intimate link between the literature on the Permanent Income Hypothesis and the question of the size of the fiscal multiplier. Some of the Permanent Income Hypothesis papers involving natural experiments mentioned above lend themselves naturally to answering questions about the effectiveness of fiscal policy.

Are tax rebates an effective means to stimulate the economy? The answer depends on whether households save or spend the tax rebates that they receive. However, answering this question with the studies above has four important caveats: the Permanent Income Hypothesis would predict that households adjust their consumption at *announcement* of the stimulus, whereas the papers cited above analyze consumption reaction at the *implementation* of the stimulus. In that sense, using them to test the Permanent Income Hypothesis and to at the same time analyze the response of household consumption to an economic stimulus is inconsistent. On the other hand, one can consider the estimates found in these papers as a lower bound of spending, because households could have adjusted their consumption already partly at the announcement or during the time between announcement and implementation. Agarwal and Qian (2014b), which is the only paper that explicitly analyzes consumption reactions to a temporary income increase both at announcement and at implementation, find a significant consumption response already at announcement, which carries over to the time period after receipt of the payment. Second, although the receipt of a rebate check can be considered exogenous for an individual household, for the economy as a whole, it is certainly not exogenous. For example, the 2008 Economic Stimulus Payments were explicitly designed to stimulate the economy. Third, the papers analyze only a partial equilibrium response of households, not taking into account any general equilibrium effects. Fourth, most of the cited studies analyze expenditure on nondurables, whereas for the fiscal multiplier, total spending matters.

That said, the studies that analyze fiscal policy measures find evidence for a large spending response by households. Parker et al. (2013) find that households spent between 50% and 90% of the 2008 Economic Stimulus Payment on durable and nondurable goods in the quarter following receipt, indicating that the majority of the payments were consumed, not saved. Johnson et al. (2006) find that nondurable consumption increased by 20 to 40% of the payments in the quarter following receipt of the 2001 tax rebates, and that around two thirds of the rebates were spent in total on nondurable consumption, considering the cumulative effect over the 6-months period following receipt.[y] Misra and Surico (2014) stress the heterogeneity of the consumption response to both the 2001 and 2008 episodes. They show that the aggregate consumption response is smaller once this heterogeneity is taken into account than the estimates imposing homogeneity would suggest. The total disbursements of $38 billion in 2001 and $96 billion in 2008 lead to an estimated aggregate consumption response of $16 ($26) billion in nondurable consumption in 2001 based on the heterogeneous (homogeneous) estimates, and to an increase of $15 ($56) billion in total consumption in 2008.[z] Sahm et al. (2010) use survey

[y] They do not analyze expenditure on durable goods.

[z] Kaplan and Violante (2014b) point out potential reasons for the generally lower impact of the 2008 reimbursements, among others the larger size of the individual household transfers, and the phasing out at the lowest income levels.

data on spending intentions. Their estimated spending responses are smaller than the ones estimated from actual expenditure data, indicating that around one third of the 2008 Economic Stimulus Payments were spent.[aa] Analyzing an older episode, Hausman (2015) finds that within one year, World War I veterans spent between 60% and 75% of bonuses that they received in 1936, mostly on cars and housing.

3.2 Military News Shocks as Natural Experiments

One way to address the potential endogeneity of government spending is to use military events as natural experiments. The work of Barro (1981) and Hall (1986) recognized the usefulness of military events in this regard early on. Geopolitical events leading to a large buildup of military expenditure are often plausibly exogenous, because they arise due to actions of some other nations. Thus, they can potentially be used as natural experiments to isolate exogenous changes in government spending. The first paper systematically following this approach is the study by Ramey and Shapiro (1998). Based on newspaper articles, they identify three major military news shocks in the post-World-War II area: the Korean War news shock in the third quarter of 1950, the Vietnam War news shock in the first quarter of 1965, and the Carter-Reagan buildup after the Soviet invasion of Afghanistan in the first quarter of 1980. Ramey (2011) adds to this list the shock of September 11, 2001. Because the identification is based on newspaper articles, these studies also fall under the "narrative approach," but they try to identify the dates of military shocks that can be used as natural experiments. Note that timing the news shock is not trivial, especially during the Vietnam War. Ramey and Shapiro (1998) document that newspaper articles only started arguing about a military buildup after the February 1965 attacks on the US Army barracks, long after the military coup of November 1, 1963, in Vietnam. Military actions of foreign entities caused all four events, such that the argument that they were exogenous to current economic conditions in the United States is very plausible.

However, using these events as natural experiments to analyze the size of the fiscal multiplier still poses some challenges. Specifically, these news might affect other relevant variables that influence GDP, rather than only government spending. This argument especially holds for World War II, in which rigid price controls were introduced and patriotism was strong, both of which might have had a direct effect on labor supply. For this reason, Ramey and Shapiro (1998) exclude World War II from the analysis. A thorough discussion of potential confounding factors in the other episodes is, however, somewhat missing from the literature. For example, the terrorist attacks of 9/11/2001

[aa] Using the same methodology and data, Sahm et al. (2012) find that the form of payment matters, and reducing tax withholdings leads to a smaller consumption response than explicitly distributing a rebate. Misra and Surico (2014) provide some reconciliations of the different estimates of consumption responses based on expenditure data and survey responses.

likely affected uncertainty about the future path of the economy, and uncertainty shocks can matter for economic performance (see, eg, Bloom, 2009).

The early study by Ramey and Shapiro (1998) runs a regression of GDP growth on these quarterly military event dummies with up to eight lags, whereas Ramey (2011) uses these dummies in a VAR approach. She finds that the military event dummies significantly precede increases in military spending, and concludes that traditional VAR approaches might fail to identify anticipation effects. This failure can potentially reconcile the different outcomes that the traditional VAR approach and the "narrative approach" relying on military buildups have found in the effect of government spending on private consumption and real wages: traditional VAR approaches typically find a positive effect of government spending shocks on consumption and the real wage, whereas the papers relying on military news shocks find a negative effect on these two variables. Edelberg et al. (1999) provide two robustness exercises for Ramey and Shapiro (1998): first, they show that results are robust to small disturbances in the event dates. Second, and more importantly, they run placebo exercises in the spirit of the placebo exercises discussed in Section 2.2.1 and find that using arbitrary event dates leads on average to response functions outside of the confidence bands of the true responses.

Ramey (2011) goes one step further than relying simply on dummy variables, and constructs a defense news variable that measures the change in expected net present value of future military spending at the quarterly frequency based on newspaper accounts. This variable is a strong predictor of government spending as long as World War II is included. The implied fiscal multiplier relying on a VAR estimation lies between 1.1 and 1.2, but falls to between 0.6 and 0.8 if World War II is excluded.[ab] In line with this evidence, Hall (2009) stresses that the identifying variation in these studies comes from large wars, especially World War II and the Korean War. Focusing on differential effects by spending categories, Auerbach and Gorodnichenko (2012) find that military spending is the spending variable associated with the largest multiplier.

3.3 Local Fiscal Multipliers

The literature on fiscal multipliers has recently started using natural experiments to establish the exogeneity of the fiscal instrument by relying on regional data and estimating local fiscal multipliers. These multipliers are local in the sense that they analyze the effects of changes in local spending financed by the federal administrative level, which therefore constitute windfall payments from the point of view of the localities and are not associated

[ab] In Owyang et al. (2013), this defense news variable is constructed for the United States from 1890 to 2010, and for Canada from 1921 to 2011. The authors then also analyze whether the government spending multiplier is larger in recessions than in booms, and find evidence in favor for Canada, but not for the United States. Barro and Redlick (2011) also use this variable on an annual frequency to estimate multipliers, and find somewhat smaller multipliers than in Ramey (2011).

with an increase in taxation or local debt.[ac] Importantly, this approach ignores general equilibrium effects at the national level. Nakamura and Steinsson (2014) provide a theoretical discussion of how the local fiscal multiplier estimate can be tied to a standard aggregate fiscal multiplier; we come back to this shortly at the end of this section.

A typical regression run in the local fiscal multiplier literature is the following variant of (9):

$$\Delta Y_{i,t+1} = \alpha + \beta \Delta \hat{F}_{i,t+1} + \gamma' \Delta X_{i,t+1} + \delta_i + \eta_t + \epsilon_{i,t+1}, \tag{10}$$

where subscript i stands for the local entity, δ_i are regional fixed effects, and η_t are year fixed effects.[ad]

An advantage of the local fiscal multiplier regression (10) over specification (9) is that it allows the inclusion of regional and year fixed effects. The regional fixed effects capture any time-invariant regional characteristics that could lead to systematically lower or higher growth in a respective region (eg, in urban vs rural regions).[ae] More important in the context of estimating the fiscal multiplier might, however, be the inclusion of year fixed effects. These effects allow one to control for any national fiscal and more importantly monetary policies that happen concurrently with the local fiscal policy. Monetary policy is often correlated with fiscal policy, and disentangling the effects of both is therefore a major challenge for any macro estimation of fiscal multipliers. However, because monetary policy is conducted exclusively on the national level, it can easily be controlled for by including year fixed effects in an estimation of local fiscal multipliers. Given that local fiscal multiplier studies all analyze multiple local subentities of a country, they lend themselves naturally to analyzing potential heterogeneous effects of fiscal multipliers depending on local characteristics such as business cycle conditions, openness, financial development, and so on.

3.3.1 Instrumental Variables

To address potential endogeneity of the fiscal instrument, regression (10) is estimated via instrumental variables, where $\Delta \hat{F}_{i,t+1}$ is the predicted change in fiscal spending based on a first-stage regression involving the instrument I:

$$\Delta F_{i,t+1} = \kappa + \theta I_{i,t} + \zeta' \Delta X_{i,t+1} + \varepsilon_{i,t+1}. \tag{11}$$

The challenge of the estimation is to find a valid instrument for the fiscal measure. The exclusion restriction for the instrument is that it affects output growth only through its

[ac] An exception to this rule is the paper by Clemens and Miran (2012).

[ad] Note that Corbi et al. (2014) and Shoag (2013) instead regress output growth on the level of the fiscal instrument, rather than its change.

[ae] The papers by Serrato and Wingender (2014) and Chodorow-Reich et al. (2012) are exceptions in this regard by not including regional fixed effects. The former instead includes state-decade fixed effects, with the local level being the county level. The latter one does not include a time dimension, thereby preventing the use of local fixed effects.

effect on the fiscal measure. Here is where natural experiments step in. Consider, for example, the paper by Serrato and Wingender (2014). Their instrument relies on the fact that federal spending at the local level is tied to the size of the local population. The estimates of the local population size come from different sources in different years: the census carried out every 10 years provides direct counts of the local population, whereas in the years in between two censuses, population estimates are updated based on vital statistics and estimated migration flows. As a result, substantial fluctuations occur in measured population in the year before a decennial census and the census year, which are called census "error of closure." The authors use this census error of closure to instrument the change in federal spending on the local level in the affected years.

For the census error of closure to be a valid instrument, it has to predict fiscal spending, and the exclusion restriction has to hold. The first of these two conditions can relatively easily be established by showing the first-stage regression results and running an F-test for the joint significance of the instruments. Concerning the second condition, the authors show theoretically that under classical measurement error in both census counts and administrative estimates, or under the weaker condition that both estimates are biased in the same order of magnitude, the exclusion restriction holds. Moreover, they explain in detail how the two estimates arise and what the literature concludes on their accuracy and potential biases. For example, one could imagine that the population estimates between censuses systematically underestimate population growth in fast-growing counties, such that the error of closure is always more positive in counties that experience an economic expansion. Any persistence in growth would then result in a direct effect of the error of closure on growth, violating the exclusion restriction. Controlling for past growth helps address this concern but does not rule it out completely. To provide further evidence that the exclusion restriction holds, the authors show that the census error of closure shows only minimal geographical correlation at the county level and no time-series correlation. Most importantly, it is not positively correlated with growth in the years before the error of closure should actually affect federal fiscal spending on the county level.[af] Although in the end these exercises remain suggestive, addressing potential concerns about the exclusion restriction in further evidence of this kind is good practice.

This literature uses two other interesting natural experiments. The first one, used in the paper by Acconcia et al. (2014), is an Italian law specifying the dismissal of elected local officials and their replacement with three external commissioners for 18 months upon evidence of Mafia infiltration in city councils. This replacement leads on average to sharp reductions in spending on public work at the provincial level, the reason being

[af] The error of closure should only affect fiscal spending 2 years after the census is run, given that publishing the results takes 2 years. In fact, a significant negative correlation exists between the error of closure and employment and income growth in previous years. This finding might raise the worry that spending rises in past recession areas, and that mean reversion might lead to future growth in these areas. The authors argue that controls for past growth in the second-stage regression take care of these concerns.

that this sector is typically a lucrative source of business for the Mafia. The authors show that growth rates prior to dismissal are not significantly different in treated and control provinces. The second natural experiment in the paper by Cohen et al. (2011) consists of changes in congressional committee chairmanships, which influence government spending in the state of the new chairman.[ag] Because chairmanship in a committee is largely determined by seniority, and because chair turnover results from election defeat or resignation of the incumbent, it is driven by political circumstances in other than the home state of the incoming chairman and can thus be seen as exogenous.

There exist other papers in this literature that rely on the same approach of finding a valid instrument for the fiscal measure in Eq. (10), but in which this instrument is less clearly a natural experiment. There is an obvious "grey zone" of what can be considered a natural experiment. In the spirit of our definition that a natural experiment is an historical event that provides exogenous variation to give a plausible identifying assumption, the census error of closure, the law specifying replacement of local officials upon evidence of Mafia infiltration, and the chairmanship in congressional committees, are such historic episodes. A variety of papers in this literature use instruments to identify exogenous variation in government spending, where the instruments do not rely on historical episodes and are themselves more directly linked to fiscal policies. Examples of these papers are Nakamura and Steinsson (2014), who exploit different state-level sensitivities to national military spending,[ah] Chodorow-Reich et al. (2012), who use precrisis state-level Medicaid spending to extract the exogenous component of increases in federal match components of state Medicaid expenditure during the 2009 American Recovery and Reinvestment Act (ARRA), and Wilson (2012), who uses exogenous formulary allocation factors such as federal highway miles in a state or a state's youth share to instrument government spending under the 2009 ARRA. Similarly, Clemens and Miran (2012) use fiscal institutions on the state level, specifically how strict balanced budget rules are, to identify exogenous variation in government spending. Shoag (2013) and Shoag (2015) exploit variations in returns to state pension plans.[ai] Kraay (2012) and Kraay (2014) focus

[ag] Feyrer and Sacerdote (2012) take a similar approach, relying on average seniority of House members, when analyzing the effectiveness of the 2009 American Recovery and Reinvestment Act.

[ah] Fishback and Cullen (2013) analyze the effect of state-level military spending during the Second World War, but do not use instruments for state-level military spending, but rather rely on narratives to establish exogeneity.

[ai] Both papers use the same instrument. However, the analysis in Shoag (2013) focuses on the Great Recession years 2008 and 2009, whereas Shoag (2015) exploits information from 1987 to 2008. Brückner and Tuladhar (2014) also analyze a local fiscal multiplier using data from Japanese Prefectures. Whereas their study shares the use of year and region fixed effects with the studies cited above, it addresses the issue of endogeneity of local government expenditures by using a system GMM approach. The paper by Fishback and Kachanovskaya (2015) analyzes federal spending on the state level during the Great Depression, relying on instruments similar in spirit to Nakamura and Steinsson (2014), Chodorow-Reich et al. (2012), and Wilson (2012).

on World Bank lending, in which project financing is spread out over several years after the lending decisions. In all of these cases, the instrument is quite closely related to the research question, which makes it harder to argue that the exclusion restriction holds. Therefore, all of these papers spend considerable effort on providing additional evidence for the validity of the exclusion restriction. A common strategy to establish exogeneity of the instrument is to include further controls and exploit specifics in the timing to argue that the identifying assumption holds; see, for example, Clemens and Miran (2012) and Kraay (2012).[aj] Wilson (2012) controls for a variety of variables that are potentially correlated with post-2009 growth and also 2009 ARRA spending, for example, pre-2009 employment growth.

3.3.2 Regression Discontinuity

There are only few papers in the local fiscal multiplier literature using a regression discontinuity approach. Since the exploited policy rules that generate the discontinuities are close to the research question at hand, these papers are closer in spirit to the second group of papers cited above than to the first group, which relies more obviously on natural experiments. Corbi et al. (2014) exploit the fact that, as in the United States, Brazilian federal transfers to municipal governments rely on the population at the local level. In contrast to the United States, a specific step function exists that specifies the total transfer amount for certain population classes. Thus, sharp discontinuities arise in the transfers per capita around the cut-off values in this step function, whereas all other variables should change smoothly around the cut-off. This is the identifying assumption for their regression discontinuity approach. Another advantage of their experiment is that several cut-off values exist (rather than, eg, only one cut-off value as in the paper by Trezzi and Porcelli, 2014), which gives the test high statistical power. Sixty percent of the municipalities in the sample switch the population class at least once in the sample period. Because the data show that adherence to the cut-off is not implemented 100%, and, in fact, some judiciary disputes surround them, the authors confront a fuzzy regression discontinuity design and use the theoretically predicted transfers based on the actual population count as an instrument for the actual transfers. They employ different bandwidths around the cut-off values, and also report results from regressions with a rectangular kernel. In a similar spirit, Becker et al. (2010) and Becker et al. (2013) use a regression discontinuity design to analyze the effect of EU grants on local development, relying on a discontinuity in regional GDP per capita eligibility for the grants.

3.3.3 Estimates of the Local Fiscal Multiplier

The estimates of the local fiscal multipliers in the studies described above show surprising consistency and range between 1.5 and 2, despite the different identifying restrictions

[aj] Kraay (2012) in addition uses predicted project-level disbursements, given the economic sector and the geographic region, rather than actual ones to address concerns about endogeneity.

based on natural experiments or other institutional details, different econometric approaches, different nations, and different fiscal spending measures.[ak] Nakamura and Steinsson (2014) show that an open economy New Keynesian model with sticky prices and complementarity between consumption and labor is best able to replicate these large local fiscal multipliers, which they call "open economy relative multiplier." They also show that the model implied "closed economy aggregate multiplier," which is commonly considered in the literature on fiscal multipliers, depends in its size very much on the accompanying monetary policy.

Another common feature of these studies is that the IV estimates are 5 to 15 times larger than the simple OLS estimates of Eq. (10).[al] This indicates that OLS estimates might be systematically downward biased, for example, because of automatic stabilizers (which increase spending in downturns of the economy), general endogeneity of government spending (increased discretionary spending to stimulate the economy in downturns), and interaction with monetary policy (fiscal spending might step in specifically when monetary policy does not work due to a binding zero lower bound).

4. IDENTIFICATION: CAUSAL FACTORS IN ECONOMIC GROWTH

4.1 The Fundamental Causes of Growth

In the previous two sections, we have explored a number of natural experiments that inform us about the validity of established macroeconomic models and the direction and magnitude of causal relationships that operate within them. In this section, we ask what natural experiments can teach us about the kind of models that we *should* be writing. In other words, what are the important determinants of economic outcomes that are outside of our standard models?

By far the most influential application of natural experiments in this respect is the search for the "fundamental" reasons of why some countries are rich while others are poor. Corresponding to the breadth of this question, the empirical work in this area is very diverse and much less focused than the work we covered in the previous sections. For this reason, we find it useful to first lay out a rough framework to fix ideas and organize the different strands of this literature into a coherent narrative.

To distinguish the fundamental causes from proximate causes of economic growth, consider a typical production function that relates output Y_t to the input of physical

[ak] These multipliers to windfall income are larger than the multiplier estimated by Clemens and Miran (2012), who take local tax or debt adjustments into account.

[al] For example, in Serrato and Wingender (2014), the 2SLS estimate is 15 times larger than the OLS estimate, 7 times larger in Acconcia et al. (2014), and 5 (based on regional data) to 15 (based on state data) times larger in Nakamura and Steinsson (2014).

capital K_t, human capital H_t, the number of workers L_t, and the level of labor-augmenting technology A_t:

$$Y_t = K_t^{\alpha} H_t^{\beta} (A_t L_t)^{1-\alpha-\beta}, \tag{12}$$

where $0 < \alpha < 1, 0 < \beta < 1$, and $\alpha + \beta < 1$. By dividing both sides of this expression by L, we can see that output per capita is a function of technology and the intensity of physical and human capital per worker. Eq. (12) thus describes a causal relationship: having better technology and more capital and education per worker makes a country richer.[am]

Although this statement is undoubtedly true, it is also not very helpful for understanding the large and persistent differences in income per capita across countries, without understanding the process by which countries accumulate capital and technology. This question is the object of a large theoretical literature on economic growth.

Consider, for example, the canonical growth model based on Solow (1956) and Swan (1956). In this model, A_t grows at the exogenous rate g and at every point in time households invest a fixed proportion of their income into physical and human capital according to

$$\dot{K}_t = s_k Y_t - \delta_k K_t$$

and

$$\dot{H}_t = s_h Y_t - \delta_h H_t,$$

where the savings rates s and depreciation rates δ are between zero and one, and $s_k + s_h < 1$.

Using lowercase variables to denote the level of output, physical, and human capital in terms of effective labor,

$$x_t \equiv X_t/(A_t L_t), X = Y, K, H,$$

we can then show that at the balanced growth path, consumption, output, physical, and human capital per effective unit of labor are all constant. In particular, output per effective labor is

$$y^* = \left(\frac{s_k}{g + \delta_k}\right)^{\frac{\alpha}{1-\alpha-\beta}} \left(\frac{s_h}{g + \delta_h}\right)^{\frac{\beta}{1-\alpha-\beta}}.$$

It follows directly that output per capita (as well as physical and human capital per capita) grows at the fixed rate g.

The Solow–Swan model again describes a causal relationship: having better technology and saving more for investment in human and physical capital makes a country richer. Although this relationship gives us a better idea of *how* growth happens (by saving

[am] Here we are assuming that all countries share the same production function (12) but do not otherwise interact with each other.

and accumulating capital), it does arguably little for our understanding of *why* people in a poor country like Zimbabwe use worse technology and/or invest less in physical capital and education than people in a rich country like the United States.

In the half century since the publication of the Solow (1956) and Swan (1956) papers, growth theory has made tremendous advances in understanding the mechanics of growth. Aside from endogenizing the savings rates (s), one of the major advances of the literature has been to explain the dynamics of the technology process, such that A_t becomes effectively an endogenous accumulated factor itself. With one major exception, which we will discuss in detail below, these advances, however, do not change the fundamental nature of the problem: physical capital, education, and technological progress are proximate rather than fundamental causes of growth. Understanding their dynamics helps us understand the mechanics of growth (how it happens) but conceptionally does not tell us why a poor country like Zimbabwe is fundamentally different from a rich country like the United States.

For the purposes of our following discussion, we will define the fundamental causes of economic growth as *the political, institutional, and social reasons that are preventing many countries from investing enough in technology and productive factors*. Formally, we may think of a typical growth model as producing a mapping from a vector of parameters ϕ_i governing the production and accumulation of technology and productive factors in a country i to its level of output per capita $\frac{Y_{i,t}^*}{L_{i,t}}$ on the balanced growth path:

$$F(\phi_i) \rightarrow \frac{Y_{i,t}^*}{L_{i,t}}. \tag{13}$$

Whereas we think of the proximate causes of growth as operating within the function F, the fundamental causes are the political, institutional, and social factors that generate cross-country variation in the parameters ϕ_i.

In the remainder of this chapter, we consider the evidence that natural experiments have uncovered for three such fundamental causes:

1. institutions,
2. social structure, and
3. culture.

For the purposes of our following discussion, we define as "institutions" the broad set of rules, regulations, laws, and policies that affect economic incentives and thus the incentives to invest in technology, physical capital, and human capital (Acemoglu, 2009). By social structure we mean "patterned relations," that is, the network of friendships and family ties between large groups of individuals that affects the diffusion of information and the ability of individuals to enforce contracts, as well as the system of socioeconomic stratification (eg, the class structure) of a society. As "culture" or "civic capital," we define those persistent and shared beliefs and values that help a group overcome the free-rider problem in the pursuit of socially valuable activities (Guiso et al., 2012). Note that the literature often refers to the latter two categories jointly as "social capital" (Putnam, 2000). We prefer to separate them here because both of these elements have been empirically shown to have independent effects on growth, but otherwise have no preference for one definition over another.

Taken together, these three causes determine the social environment in which economic activity takes place. Because conducting controlled experiments on this social environment at the scale of countries or even regions is generally impossible, natural experiments are arguably our best bet for making causal inference at this scale. In each case, we will first discuss a selection of natural experiments that provide evidence of the causal relationship between GDP per capita and institutions, social structure, and culture, respectively. Then, whenever available, we will discuss some of the evidence on the dynamics of each of these forces and avenues for future research. Because each of these literatures is vast, we cannot give a comprehensive overview of each of them (Guiso et al., 2012; Algan and Cahuc, 2013; Alesina and Giuliano, 2015; Nunn, 2013, and Chaney, 2014 provide excellent surveys). Instead, we focus on giving a broad overview of the types of measurement and identification issues that arise in these applications, and provide specific examples.

To our list of three fundamental causes, we add a set of natural experiments that speak to the question of whether multiplicity might exist in the mapping (13). In other words, the mechanics of economic growth may be such that two countries with identical ϕ_i may end up on different balanced growth paths by pure coincidence or luck. This scenario is the major exception mentioned above: models that generate multiple equilibria may be able to explain why Zimbabwe is poor and the United States is rich, without taking recourse to any fundamental causes of growth:

4. luck and multiple equilibria.

Here again we select papers to give a broad overview of the measurement and identification issues.

One striking feature that the studies in all four categories have in common is that they tend to make relatively little use of the structure provided by growth models. That is, most approaches relate variation in ϕ_i directly to $\frac{Y^*_{i,t}}{L_{i,t}}$, without making use of the structure of F. In this sense, the literatures on the proximate causes and fundamental causes of growth have developed relatively independently. We will comment on this issue further below.

4.2 Institutions and Political Economy

Of the three fundamental causes of economic growth, "institutions"—the rules, regulations, laws, and policies that affect the incentives to invest in accumulated factors of production—have received by far the most attention from empiricists. The reason for this attention is that institutions, and in particular the protection of property rights, fit neatly into standard ways of thinking about incentives: investors will be reluctant to invest in capital and technology if the returns from these investments are likely to disappear into the pockets of corrupt bureaucrats. Clearly, bad institutions must thus be bad for growth. A correspondingly long history of thought in economics links institutions to economic development (North and Thomas, 1973; Jones, 2003; North, 1981; Mauro, 1995; Hall and Jones, 1999; La Porta et al., 1998).

However, despite this long intellectual history, which way the causality runs is far from clear: for example, we may suspect that richer countries can also afford a better judicial system, and thus better protect property rights. Similarly, researchers disagree about what makes governments *want to* protect property rights. In practice, the set of institutions a society adopts depends in complex ways on the society's level of economic development, culture, social structure, the education of the population, and its historical experience. Identifying a causal effect of institutions on economic growth thus requires an exogenous source of variation in the type of institutions that different societies adopt.

Consider, for example, the following structural equation of interest:

$$log(Y_i) = \alpha + \beta R_i + X_i' \gamma + \epsilon_i, \tag{14}$$

where Y_i is income per capita in country i, R_i is a measure for the protection of property rights, and X_i is a vector of controls. The fundamental problem in identifying β is that R_i and Y_i are likely to be jointly determined such that $cov(R_i,\epsilon_i) \neq 0$, resulting in biased OLS estimates of (14). Unbiased estimates of β thus require exogenous variation in R_i.

In this section, we first discuss a set of natural experiments that attempt to identify such exogenous variation, and establish a causal link between institutions and growth. We then turn to the mechanism that links institutions to growth and to the dynamics of institutions.

4.2.1 The Effect of Institutions on Growth

In an influential study, Acemoglu et al. (2001), consider the natural experiment surrounding the colonization of virtually all of Africa, North and South America, Australia, and large parts of Asia by Europeans. Today, the countries that emerged from these former colonies differ widely in their level of economic development and in the functioning of their institutions—and these differences appear highly persistent over time. For example, the United States has been richer than the Congo for more than a century, and there is little evidence that this will change in the near future. Acemoglu, Johnson, and Robinson (AJR) argue that the mortality rates Europeans settling in these different colonies faced explain part of this heterogeneity: Europeans who made the trip to the Congo could expect to be plagued by malaria and yellow fever and consequently had much lower life expectancies than settlers heading for the present-day United States. This health cost that

Europeans faced when considering whether to settle in different parts of the world determined the form of colonization that each country experienced.

AJR's idea is that in countries with a disease environment that was hazardous to Europeans (eg, the Congo), the early colonizers set up institutions that would maximize the extraction of natural resources, requiring as small a European presence as possible. These institutions did not introduce much protection of property rights and did not involve checks and balances against the power of the ruling elites that were dominated by Europeans. By contrast, colonies that were attractive destinations for European settlers (eg, Australia and the United States) imported European institutions that fostered safe property rights and checks and balances in government. The resulting historical differences in institutions across countries then tended to persist to the present day, resulting in persistent differences in economic performance that last to the present day.

Fig. 2 shows the relationship between settler mortality and a contemporary measure for average expropriation risk for the 64 countries included in AJR's sample. The authors argue that this negative and highly significant relationship delivers quasi-random variation in R_i because the mortality rates of European settlers in the 17th to 19th centuries (or any omitted variables correlated with them) are unlikely to have a direct effect on GDP per capita today, other than their effect through institutions.

When instrumenting R_i with the mortality rates of soldiers, bishops, and sailors stationed in country i between 1600 and 1800 CE, AJR's estimates of the parameter β are positive and statistically highly significant. They imply that an improvement in Nigeria's level of protection of property rights to match that of Chile would, in the long-run, increase Nigeria's GDP per capita by a factor of 7. Heterogeneity in institutions may thus account for a large part of the differences that exist today in wealth between countries.

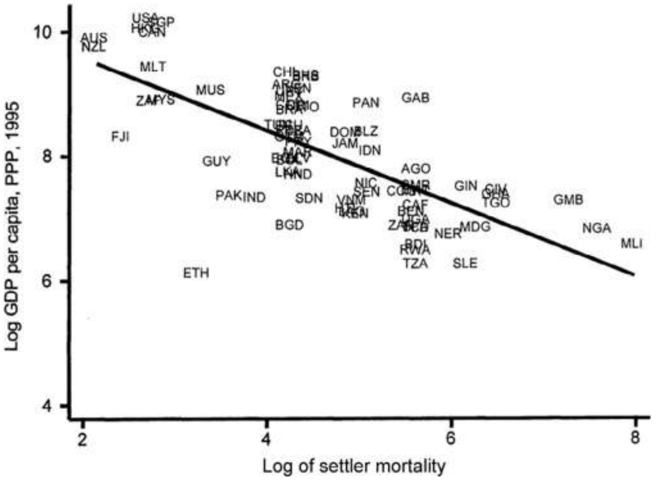


Fig. 2 Relationship between settler mortality and expropriation risk in Acemoglu et al. (2001).

In an additional application of the same natural experiment, Acemoglu et al. (2002) argue that heterogeneity in institutions can also explain the large changes in the relative prosperity that have occurred since 1500 CE. Before the onset of European colonization, civilizations in Meso-america, the Andes, India, and Southeast Asia appear to have been richer than those in North America, Australia, or New Zealand. Today, the relationship is reversed: a simple regression of GDP per capita in 1995 on urbanization in 1500 (as a proxy for historical GDP per capita) yields a negative and highly statistically significant coefficient. AJR show that differences in the institutions imposed by European colonizers can again explain this fact. When adding measures of institutions to the regression and instrumenting them with settler mortality, the negative correlation between present-day GDP per capita and historical urbanization disappears. Because settler mortality and population density in 1500 are positively correlated, European colonizers tended to introduce or perpetuate extractive institutions in places where they found a high density of local populations, while creating good institutions in places where they settled in large numbers.[an]

The main identifying assumption for a causal interpretation of the results in both papers is that, conditional on the controls included in the vector X, the mortality rates of European settlers in the 17th to 19th centuries (or any omitted variables correlated with them) have no effect on GDP per capita today, other than their effect through institutions.

This assumption raises two main concerns. First, omitted variables that are correlated with historical mortality rates could be correlated with present-day GDP per capita. The most obvious of these variables is the current disease environment. For example, places that were hard to settle historically may be difficult to live in today, resulting in lower economic growth. AJR argue that this is not the case. The historical disease environment had large negative effects on Europeans, but much smaller effects on the indigenous populations. For example, local troops serving the British army in Bengal had mortality rates that were comparable to those of British troops stationed in Britain, whereas British troops in Bengal had mortality rates that were 7 to 10 times higher. In addition, AJR's estimates for β change little when controlling for elements of the current disease environment, suggesting that the effect of the historical disease environment indeed transmits itself through differential European settlement at the time. They also show that their results are robust to controlling for other correlates of GDP per capita, such as the identity of the colonizer, the origin of the legal system, and various measures of climate conditions.

[an] For a more in-depth discussion of the merits and weaknesses of AJR (2001)'s analysis, see Easterly and Levine (2003); Glaeser et al. (2004); Olsson (2004); Rodrik et al. (2004); Acemoglu et al. (2012); Albouy (2012); Easterly and Levine (2012), and Acemoglu et al. (2014a).

The second main concern is much harder to address: the effect of European settlement may transmit itself through a range of different mechanisms that may be correlated with the protection of property rights. In particular, Europeans may also have imported their own culture and high levels of civic capital or they may have brought valuable social ties to their countries of origin. Thus, AJR are unable to distinguish the effect of institutions from the effect of other persistent variables that may be correlated with both European settlements and institutions, and in particular from the two other fundamental causes of economic growth discussed in the next two subsections. More generally, any instrumental variables approach that relies on cross-sectional variation will run into this problem.

Michalopoulos and Papaioannou (2014) attempt to make progress on distinguishing the effects of institutions and culture using data on light intensity in combination with a regression discontinuity approach. They argue that the borders between many African countries were drawn in the mid to late 19th century by Europeans who were largely uninformed about local conditions. As a result, these borders partition more than 200 historic homelands of ethnicities between two different modern-day countries in a quasi-random way, subjecting identical cultures residing in geographically homogeneous territories to different country-level institutions. Their main specification takes the following form:

$$y_{p,e,i} = \alpha + \beta IQL_i^{HIGH} + f(BD_{p,e,i}) + \gamma PD_{p,e,i} + X'_{p,e,i}\Phi + \eta_e + \epsilon_{p,e,i}, \tag{15}$$

where $y_{p,e,i}$ is a dummy variable that is 1 if pixel p (an area corresponding to about 12 × 12 km) in the historic homeland of ethnic group e in country i is lit—a simple measure of whether residents of that piece of land can afford nighttime lighting. IQL^{HIGH} is an indicator that is 1 if the pixel is in the part of the partitioned homeland that is located in the country with the relatively higher institutional quality, $f(BD_{p,e,i})$ is a polynomial of shortest distance of the pixel centroid to the country border, PD is population density, X contains additional controls, and η_e is an ethnic homeland fixed effect. Under the identifying assumption that at the country border, institutions change discretely, while all other relevant influences on $y_{p,e,i}$ (including culture and geography) change continuously, β measures the effect of relatively better institutions at the country border.

In contrast to the results of AJR (2001)'s cross-country analysis, the authors find no effect of institutions at the country border: across different variations, they cannot distinguish β from zero, suggesting that, at the border, exogenous variation in national institutions is not associated systematically with higher wealth (more light). These results may suggest that, at least for Africa, omitted factors such as culture (civic capital) or social structure might explain the strong association between institutions and GDP per capita. However, more consistent with AJR (2001)'s results, Michalopoulos and Papaioannou find a positive and significant effect of better institutions for split groups that are located close to national capitals. The absence of an effect at the border of the average African country could thus simply be explained by the fact that the influence of many African countries' governments is weak in remote regions.[ao]

In a closely related paper, Pinkovskiy (2013) also uses a regression discontinuity approach to test for discontinuities in the amount of light per capita at country borders, but applies his analysis to the entire world, rather than only to Africa. Because most borders in the world were not drawn randomly but might be determined by the location of ethnic homelands, culture, or other variables, he restricts his estimates to within 50km of the border. Within this narrow bandwidth, one may plausibly argue that the *exact* location of the border is quasi-random, even though borders outside of Africa were not drawn by largely uninformed imperial powers. In contrast to Michalopoulos and Papaioannou (2014), Pinkovskiy finds a large and statistically significant discontinuity at national borders, although it is not significant when restricting the sample to Africa. In addition, this effect of the country border becomes statistically insignificant once Pinkovskiy controls for differences in the rule of law.

4.2.2 The Effect of Institutions on Business Cycles and Conflict

Given this evidence of a causal link between institutions and growth, an obvious question is *how* this effect transmits itself in practice. One interesting, and often overlooked, piece of evidence comes from yet another application of AJR (2001)'s natural experiment. Acemoglu et al. (2003) show that countries with worse institutions (again instrumented with settler mortality) also have more volatile business cycles and are more prone to episodes of economic crises. Once the authors control for this effect of institutions, the standard macroeconomic policies that are often blamed for macroeconomic instability (eg, high government spending and high inflation rates) are no longer systematically associated with macroeconomic volatility.

The authors draw two main conclusions from this finding. First, bad macroeconomic policies are often a symptom of underlying institutional problems. Second, they are often not the primary mediating channel through which institutions affect macroeconomic stability. These results suggest that a more useful line of thought might be that policies such as excessive government spending and high inflation rates are just two of a large set of tools that politically powerful groups use to extract rents from the economy. Any policy (perhaps imposed by the IMF or another international organization) that shuts down one of these distortions but does not resolve the underlying institutional problems, may then simply result in the use of a different tool (another macroeconomic or microeconomic distortion) that achieves the same aim. Macroeconomic instability and low growth may then be thought of as collateral damage from this rent-extraction process.[ap]

[ao] There is also a technical debate on the bleeding of nighttime light across borders that may attenuate any estimated border-effect, see the discussion in Pinkovskiy (2013).

[ap] Most macroeconomic models imply that macroeconomic volatility reduces growth; see Baker and Bloom (2013) for estimates of this relationship.

In a second application of their natural experiment surrounding the quasi-random drawing of African boundaries by relatively uninformed Europeans, Michalopoulos and Papaioannou (2011) show that ill-designed institutions may also slow economic development by prompting long-term conflict. Using data on the precolonial locations of 834 ethnic groups, they show that the drawing of African boundaries indeed appears to have quasi-randomly partitioned 229 of these groups between two adjacent states: partitioned and nonpartitioned ethnicities do not appear to differ systematically in their precolonial characteristics, except that partitioned groups tended to cover larger land areas and to have historical homelands with larger areas under water. The authors then show that between 1997 and 2010, partitioned ethnic homelands had a 30% higher likelihood of experiencing political violence (eg, battles between government forces, rebel groups, and militias) and a 40% higher likelihood of experiencing violence against civilians (eg, murders, abductions, and child-soldiering raids) than nonpartitioned ethnic homelands. This higher incidence of violence is also associated with worse economic outcomes and lower provision of public goods.

4.2.3 Persistent Effects of Historical Institutions

The main conclusion from this set of studies is that dysfunctional institutions are a major obstacle to economic development because they deter investment, create macroeconomic instability, and potentially lead to violent conflict. A crucial question for policy is then whether replacing such institutions reverses these adverse effects. Several papers study this question using natural experiments surrounding the imposition and subsequent abolition of historical institutions.

Banerjee and Iyer (2005) examine the long-term impact of colonial land revenue systems in British India. In some Indian districts, British officials levied taxes on agricultural income, whereas in other districts, the collection of taxes was left to a class of native landlords who were free to set the revenue terms for the peasants under their rule and to dispossess them if they did not pay their dues. After Indian independence in 1947, all direct taxes on agricultural income (and thus both institutions) were abolished. Nevertheless, Banerjee and Iyer find that to the present day, districts that used the landlord-based system have lower agricultural yields and investment, lower public investment in health and education, as well as worse health and educational outcomes. To show that these relationships are causal, the authors instrument the choice of land revenue system with a dummy variable equal to 1 if the district was annexed between 1820 and 1856, a period during which the authors argue the British preferred the nonlandlord-based systems for political and intellectual reasons that are orthogonal to the characteristics of the annexed districts. Banerjee and Iyer argue that these adverse effects of the landlord-based system persist because of its effect on the social structure of the affected areas: it created a class-based antagonism that limits the capacity for collective action in the affected districts to the present day. As a result, these districts are relatively less able to muster the political influence to claim their fair share of expenditure in education, health care, and public goods.

Rather than focusing on heterogeneous institutions within districts that were directly controlled by the British, Iyer (2010) studies the long-term effects of direct British colonial rule vs indirect rule across 415 districts in present-day India. Although all of India was under British political control by the middle of the 19th century, the British administered directly only part of this area (British India), while Indian Kings (Princely States) who had considerable autonomy ruled the remaining parts. After the end of colonial rule in 1947, all of India then came under a uniform administrative and political structure. The major problem when attempting to estimate the causal effect of direct British control is that the British did not randomly annex areas, but presumably were more eager to annex richer than poorer areas. Iyer deals with this selection problem by using the "Doctrine of Lapse," a policy that was in place from 1848 to 1856, under which the British annexed princely states whose rulers died without a natural heir. Using lapse as an instrument for direct British rule, she finds no effect on agricultural productivity and investment, but a negative effect of direct British rule on the provision of public goods, education, and health care that lasts through the 1990s, although these persistent effects appear to be decreasing over time.

Whereas both Banerjee and Iyer (2005) and Iyer (2010) rely on instruments in their identification strategy, Dell (2010) uses a regression discontinuity design to study the long-term effects of the *mita* forced labor system that the Spanish instituted in Peru and Bolivia from 1573 to 1812. Under this system, villages within a geographically precisely determined area were required to send one seventh of their male adult population to work in silver and mercury mines. Similar to Pinkovskiy (2013), she argues that the exact location of the *mita* boundary is quasi-random within a 50km band, such that all other relevant influences on household consumption vary smoothly at the border of the *mita* area. Subject to this assumption, Dell estimates that the mita had a persistent negative effect that lowers household consumption by 25% almost 200 years after its abolition. She argues that this effect most likely transmits itself again through the lower provision of public goods and education services to these areas.

Alesina and Fuchs-Schündeln (2007) analyze the long-term effect of institutions on individual economic preferences, as opposed to economic outcomes. They explore German separation and reunification as a natural experiment to analyze how 45 years of communist rule affected individuals' attitudes toward market capitalism, and the role of the state in providing insurance and redistribution from the rich to the poor. Seven years after reunification, East Germans are still much more likely to favor a strong role of the government. This difference in preferences is larger for older cohorts, who lived under different systems for a longer time period. The micro data allow the authors to include rich controls for individual economic motives to favor a strong government, such that the difference can confidently be attributed to the effect of living under a market system vs a communist system. Similar to the convergence found in Iyer (2010), the longer East Germans live under the new market system, the more their preferences resemble those of West Germans. Under the assumption of linear convergence, full convergence will take one to two generations. This convergence is the product of two forces: around one third of the convergence is due to a shift in the cohort composition towards younger cohorts, and around two thirds can be attributed to actual convergence of preferences within individuals.[aq]

The main conclusion from this set of natural experiments is that historical institutions have long-lasting effects—even after their abolition—that transmit themselves through their effect on the distribution of political power, social structure, preferences, or some other mechanism. Although all three studies focusing on economic outcomes commit considerable effort to narrow down the potential channels of persistence, their design does not allow them to identify the exact channel. A remaining challenge for future work is thus to go beyond causal identification of the treatment effect of historical institutions and to identify natural experiments that speak simultaneously to the treatment effect and its channel of persistence.

4.2.4 Determinants and Dynamics of Institutions

If institutions have a large causal effect on economic growth, a crucial question is how countries acquire well-functioning institutions and what might determine their evolution over time. The natural experiments covered so far suggest colonization has generated large and persistent differences in the "level" of the quality of governance and institutions across countries. This view is also consistent with a large literature on the economic consequences of legal origins, which has shown that the identity of the colonizing country determines the type of legal system used in former colonies as well as its performance along a broad range of dimensions (La Porta et al., 1998, 1997, 2008). Another example of institutional change resulting from foreign domination and conquest is the abolition of feudalism and the imposition of French civil law in German states conquered during the Napoleonic wars (Acemoglu et al., 2011a).

Aside from these large-scale cross-sectional experiments, a more practical question, at least from a policy perspective, may be what determines the dynamics of institutions over time. To address this question, a large set of studies attempts to identify exogenous shocks to the political balance of power between different groups within a polity, and how these shocks may foster or retard the development of good political and economic institutions. A closely related set of studies, which we discuss in Section 4.3.3, focuses instead on identifying exogenous shocks to a society's social structure.

4.2.4.1 Shocks to the Political Balance of Power

One interesting approach to identifying exogenous shocks to the political balance of power uses shocks to the natural environment as an instrument. For example,

[aq] Fuchs-Schündeln and Schündeln (2015) similarly find evidence for the endogeneity of democratic preferences.

Brueckner and Ciccone (2011) combine information about GDP per capita and average annual rainfall with measures of democratization and constraints on the executive in sub-Saharan African countries 1981 to 2006.[ar] Because many sub-Saharan countries rely heavily on agriculture, negative rainfall shocks (droughts) may serve as a good instrument for recessions in these countries. Brueckner and Ciccone's main specification relates changes in constraints on the executive to lagged GDP per capita and a full set of time and country fixed effects, while using lagged average rainfall as an instrument for lagged GDP per capita. They find a negative and highly significant coefficient on GDP per capita. Because the specification uses only variation within countries (due to the inclusion of country fixed effects), this finding implies that transitory recessions are associated with democratization, a result that is in line with a literature that has argued theoretically that autocracies tend to become vulnerable in times of economic crisis (eg, Lipset, 1959; Huntington, 1991; Acemoglu and Robinson, 2005).

The main advantage of using changes in the natural environment rather than "man-made" historical events as an instrument is that ruling out reverse causation between GDP per capita and droughts is easy. The main disadvantage is that droughts, and shocks to the natural environment in general, may affect democratization (and many other things) through a variety of channels. Brueckner and Ciccone have two main responses to this concern. First, their standard specification relates changes in constraints on the executive to last year's GDP growth, such that the timing of the effects mitigates the possibility that droughts may affect GDP through institutions rather than the other way around. Second, they show that the reduced-form effect of rainfall on democratization is much smaller in countries that rely less on agriculture.[as]

Chaney (2013) applies a similar empirical strategy to a completely different context. He argues that in medieval Egypt, the political power of religious leaders tended to increase during years with deviant Nile floods, because economic crises increased their capacity to coordinate a revolt. His main specification shows that in years with deviant floods, Egypt's main religious figure was less likely to be replaced and that those years also showed more evidence of popular unrest. He then shows evidence from a variety of sources to distinguish his interpretation from a variety of other channels through which deviant floods might have affected Egyptian society.

Rather than using historical variation in floods over time, Hornbeck and Naidu (2014) study the effect of a single event, the Great Mississippi Flood of 1927, on the balance of power between black laborers and white landowners in the affected areas.

[ar] Constraints on the executive typically refer to the extent of institutional constraints on the decision-making powers of the chief executive, such as the president.

[as] Also see Miguel et al. (2004) and Franck (2015) for similar natural experiments. Other empirical papers studying the relationship between income and political institutions include Barro (1999); Glaeser et al. (2007); Acemoglu et al. (2005b); Persson and Tabellini (2009); Burke and Leigh (2010), and Acemoglu et al. (2014b).

In 1927, the Mississippi River broke its banks in an unprecedented flood that inundated 26,000 square miles and displaced the affected population. Hornbeck and Naidu argue that this event represented a significant shock to oppressive racial institutions that were geared towards keeping black laborers on the land and in jobs with depressed wages. Subject to the identifying assumption that flooded and nonflooded areas in the same state and with similar preflood characteristics would have developed similarly absent the flood, they show that flooded areas experienced an immediate and persistent out-migration of black laborers. In the following decades, these areas, deprived of cheap labor, modernized agricultural production and increased capital intensity relative to nonflooded areas.

Instead of shocks to a country's natural environment, a large literature on the "resource curse" studies the effect of exogenous changes in the value of countries' endowment of commodities on the quality of their institutions. For example, Caselli and Tesei (2011) calculate the flow of resource rents accruing to commodity-exporting countries. They argue that most commodity exporters are small relative to the world market, such that, for example, a change in the world price of oil is exogenous to Venezuela's political system. Their main specification relates the 1-year change in an index of the quality of a country's political institutions to the lagged, 3-year average change in the price of its principle export commodity. Their findings are consistent with the idea that an increase in the value of a nondemocratic country's natural resources tips the balance of political power in favor of the ruling elite: when autocratic countries receive a positive shock to their flow of resource rents, they tend to respond by becoming more autocratic subsequently.[at]

4.2.4.2 Popular Mobilization

The main conclusion from the set of papers studying the effect of shocks to a country's political balance of power is that some of these shocks may make it easier to place constraints on the ruling elite, and thus obtain "good" institutions. However, *how* such constraints may be imposed in practice remains unclear, particularly in nondemocratic societies that lack a functioning mechanism for replacing the ruling elite. Part of the reason for this lack of evidence is that typical measures of institutional quality vary only at the annual or lower frequency and are thus hard to tie to particular events. In a recent study, Acemoglu et al. (2015) suggest using daily financial data to measure the real-time effects of popular mobilization in street protests on investors' expectations of economic rents from future favoritism and corruption accruing to politically connected firms. Their approach generalizes the event study methodology typically used in the finance literature that estimates the value of political connections from changes in the relative stock market valuations of politically connected firms (Roberts, 1990; Fisman, 2001).

[at] Also see Brückner et al. (2012) for a similar exercise.

During Egypt's Arab Spring, street protests first brought down Hosni Mubarak's government and then ushered in an era of competition between three groups that repeatedly rotated in and out of power: elites associated with Mubarak's National Democratic Party (NDP), the military, and the Islamist Muslim Brotherhood. In their main specification, Acemoglu, Hassan, and Tahoun estimate the direct effect of street protests on the valuation of firms connected to the group currently in power relative to their effect on the valuation of nonconnected firms:

$$R_{i,t} = I'_{i,t}\beta + \left(P_t \times I'_{i,t}\right)\gamma + X'_i \times \delta_t + \delta_t + \eta_s + \epsilon_{i,t}, \tag{16}$$

where $R_{i,t}$ is the log return of firm i on day t and P_t denotes the daily number of protesters in Tahrir Square estimated from Egyptian and international print and online media. $I_{i,t}$ denotes a vector of two dummies, the first reflecting political connections to the group currently in government and the second recording connections to the two other rival (nonincumbent) power groups. X_i is a vector of firm-level controls, and δ_t and η_s denote, respectively, time and sector dummies. The coefficients of interest are the entries of the vector γ, and measure the effect of the number of protesters in Tahrir Square on the relative stock market valuation of firms connected to the incumbent group and the relative valuation of firms connected to the two rival (currently nonincumbent) groups.

The estimates of γ show a robust and quantitatively large negative effect of larger protests on the returns of firms connected to the incumbent group, but no effect on the valuation of their rivals. For example, a turnout of 500,000 protesters in Tahrir Square lowers the market valuation of firms connected to the incumbent group by 0.8% relative to nonconnected firms, but triggers no offsetting gain in the value of "rival" (nonincumbent) connected groups. These results hold even when excluding periods from the analysis during which protests may have resulted in a change in regime or any kind of formal institutions.

The main identifying assumption for a causal interpretation of the effect of protests on the relative valuation of firms connected to the incumbent group and their rivals is that (i) no omitted variables should exist that fluctuate at the daily frequency and are correlated with both stock returns and the number of protesters in Tahrir Square, and (ii) no reverse causality should result from daily differential returns on firms connected to different power groups to the intensity of protests. Key to corroborating this assumption is the daily frequency of the data and the timing of the effect: if both stock market valuations and protests respond to some other slow-moving variable, then one might expect stock returns to be correlated with future protests (the lead of protests should be statistically significant). The authors show that this is not the case in the data. Instead, stock returns respond only during and immediately after the protest. According to their preferred interpretation, these results provide evidence that popular mobilization and protests have a role in restricting the ability of connected firms to capture excess rents and thus may limit favoritism and corruption even if they do not result in changes in formal institutions or the identity of the government.

This interpretation of a causal link between popular mobilization and corruption is also in line with evidence from a natural experiment that links popular mobilization to a range of other political outcomes in the United States: Madestam et al. (2013) study the effect of the first large rallies by the Tea Party on April 15, 2009 (the day income tax filings became due). In many regions, these rallies were the first large event organized by the movement. Using rainfall on that day as an instrument, the authors show that higher attendance at these initial rallies (due to good weather) is associated with higher subsequent political support for the Tea Party's political positions, more votes for the Republican party in the midterm elections of 2010, and more conservative voting records of congressmen subsequently elected in the region.

4.3 Social Structure

A powerful idea in economic sociology is that the economic success of an entity, be it an individual, a household, or a geographic region, depends on its position in the social structure of the marketplace. Well-connected individuals who bridge "holes" in this social structure are more likely to be economically successful and may generate a competitive advantage for the firms at which they work and the regions in which they live (Loury, 1977; Burt, 1992; Granovetter, 1985, 2005). For example, one might imagine that well-connected individuals provide "social" collateral for economic transactions that would not otherwise be feasible, or reduce informational frictions by providing a credible channel for communication (Coleman, 1988; Greif, 1993; Stiglitz, 1990). According to this view, social ties that are formed and maintained for historical or personal reasons are a fundamental cause of economic growth and thus affect the economic development of entire geographic regions.

Saxenian (1999) gives an example of this view. She analyzes the biographies of Indian engineers who migrated to California in the 1970s. Following the liberalization of the Indian economy in 1991, these immigrants were in a position to leverage their social ties to relatives and friends in Hyderabad and Bangalore. Many excelled in their personal careers, managing outsourcing operations for US firms. Saxenian argues that by connecting Silicon Valley firms to low-cost and high-quality labor in their regions of origin, these Indian immigrants became instrumental in the emergence of their home regions as major hubs of the global IT services industry.

Although empirical evidence of the effect of social ties on a range of microeconomic outcomes is compelling, estimating the effect of social ties on economic growth poses additional difficulties.[au] The reason is that social ties are likely to be nonrandom across as well as within regions. Individuals who have social ties may have common unobserved

[au] These microeconomic outcomes range from education (Sacerdote, 2001) and employment (Munshi, 2003) to performance in the financial industry (Cohen et al., 2008) and agricultural yields (Conley and Udry, 2010). Also see Bertrand et al. (2000), Hochberg et al. (2007), Beaman (2012), Kuhnen (2009), Shue (2013), and Banerjee et al. (2013).

characteristics, sort endogenously across regions, or form social ties in anticipation of future economic benefits (Manski, 1993; Glaeser et al., 2002). Identifying a causal link between social ties and macroeconomic outcomes thus requires exogenous variation both in (a) the economic value of social ties and (b) the formation of these ties across geographic regions. In general, and in the Indian example above in particular, such exogenous variation either does not exist or cannot be measured.

In this section, we first discuss a natural experiment that attempts to address this identification problem, and establish a causal link between social ties and growth. We then turn to the effect of social ties on other aggregates such as trade and foreign direct investment. Compared to the large body of literature on institutions, the study of this fundamental cause of economic growth is in its infancy. In particular, we are unaware of natural experiments that study the dynamics of the formation and value of social ties over time.

4.3.1 The Effect of Social Ties on Growth

Burchardi and Hassan (2013) use the natural experiment surrounding the fall of the Berlin Wall to identify the causal effect of social ties on economic growth. They argue this setting's key advantage is that the partition of Germany was generally believed to be permanent. After the physical separation of the two German states in 1961, private economic exchange between the two Germanies was impossible. As a result, West Germans who maintained social ties with East Germans during this period did so for purely noneconomic reasons. After the fall of the Berlin Wall, trade between the two Germanies suddenly became feasible. To the extent that social ties facilitate economic exchange, the fall of the Berlin Wall thus generates exogenous variation in the value of these ties (condition (a) stated above).

In their main specification, Burchardi and Hassan study the growth in income per capita across West German regions in the 6 years following the fall of the Berlin Wall as a function of the share of the region's population that has ties to relatives in East Germany. Their structural equation of interest is

$$Y_r^{95} - Y_r^{89} = \beta T_r^{89} + Z_r' \gamma + \varepsilon_r, \tag{17}$$

where Y_r^t is log income per capita in region r in year t, T_r^{89} is a proxy for the share of the region's population that has ties to the East, and Z_r is a vector of controls that contains a complete set of federal state fixed effects, log income per capita in 1989 $\left(Y_r^{89}\right)$, the growth rate of income per capita between 1985 and 1989, and the distance from region r to the inner-German border. Because the specification controls for the pretrend in growth, the coefficient β estimates the *differential change* in the growth rate of income per capita after 1989 for regions with different intensities of social ties to the East.

Eq. (17) consistently estimates the parameter of interest if $Cov\left(T_r^{89}, \varepsilon_r\right) = 0$. However, this covariance restriction may not hold in the data, because the strength of social ties to East Germany may be correlated with differences in growth prospects across regions. An unbiased estimate of β thus requires exogenous variation in the regional distribution of social ties (condition (b) stated above). Burchardi and Hassan argue that such variation arises as the result of a large-scale migration post World War II when millions fled from East Germany to the West. An overwhelming concern for the migrants arriving in the West at the time was an acute lack of housing. During World War II, almost a third of the West German housing stock was destroyed. Variation in wartime destruction thus made settling in some parts of West Germany more difficult than in others at a time when millions of migrants were arriving from the East. Burchardi and Hassan argue that the extent of regional destruction in 1946 thus provides the exogenous source of variation in the regional distribution of social ties needed for identifying β.

Their key identifying assumption combines a difference in differences approach with instrumental variables: it states that, conditional on the covariates in Z, (i) no omitted variable drives both wartime destruction and differential changes in income growth post-1989, and (ii) wartime destruction in 1946 has no effect on changes in the growth rate of income per capita after 1989 other than through its effect on the settlement of migrants who have social ties to the East.

Using the degree of wartime destruction as an instrument for the share of the population with social ties to the East, Burchardi and Hassan then explicitly test the hypothesis that a concentration of households with social ties to East Germany in 1989 in a given West German region is causally related to a rise in the growth rate of income per capita after the fall of the Berlin Wall. They find that regions that received a one-standard-deviation larger share of migrants from the East prior to 1961 experienced a 4.6-percentage-point higher growth rate of income per capita in the 6 years following the fall of the Berlin Wall. Although both entrepreneurs and nonentrepreneurs who live in regions with strong social ties to the East experience a rise in their incomes post-1989, the incomes of entrepreneurs increase at a significantly higher rate. Moreover, the share of the population engaged in entrepreneurial activity rises in regions with strong social ties to the East. Consistent with this increase in entrepreneurial activity, West German firms that are headquartered in regions with strong social ties to the East are more likely to invest in East Germany between 1989 and 2007.

A crucial question for the interpretation of these results is whether this link between social ties and growth is a purely "microeconomic" phenomenon in the sense that a few people who have ties to the East internalize all of the benefits from this tie, or whether it is a "macroeconomic" phenomenon that involves positive spill-overs from one person's tie to the East to the income growth of unconnected individuals living in the same region.

Using household-level data, Burchardi and Hassan show that the income growth of households with ties to at least one relative in the East is on average 6 percentage points higher in the 6 years following the fall of the Berlin Wall than that of comparable households with no such ties. The authors then relate the effects of social ties on household and region-level income growth using a model in which household income is a function of direct and indirect (higher-order) social ties to the East, where individual households may benefit from having friends with ties to the East, even if they themselves have no such ties. Their preferred estimates imply that, other things equal, a direct social tie to the East has the same effect on individual household income as a 50-percentage-point (or 3.5-standard-deviation) increase in the regional share of households with such ties. From the perspective of an individual household, the incremental benefit from a direct social tie to the East is thus large compared to the incremental benefit from higher-order social interaction. Nevertheless, indirect social ties to the East account for two thirds of the aggregate effect, because all of a region's households benefit from indirect social ties, whereas only a subset of the population benefits from direct social ties.

4.3.2 The Effect of Social Ties on Trade and FDI

Aside from an interest in identifying the fundamental causes of economic growth, trade economists have long documented a strong association between social ties and the pattern of international trade and foreign direct investment (FDI) as part of a broader effort to understand the role of informal barriers in shaping economic interactions across borders (Gould, 1994).[av] The central puzzle in this literature is that geographic distance and country borders have a strong negative impact on bilateral trade flows, even after controlling for any measurable barrier to trade. Part of this puzzle may be resolved if social ties are effective means of overcoming informal barriers to trade (eg, informational frictions or problems with contract enforceability) and are negatively correlated with geographic distance and country borders.

Perhaps the most influential of these studies is Rauch and Trindade (2002), who study the effect of ethnic Chinese networks on international trade. They study Chinese networks in particular because data on the population share of ethnic Chinese are readily available and because Chinese migrants are present in most countries of the world. In their main specification, the authors estimate a conventional gravity equation that explains bilateral trade as a function of the product of the GDP of the two countries, geographic distance, and the product of the share of the population of the two countries that is ethnic Chinese. This product can be interpreted as a proxy for the strength of social ties between Chinese residents in the two countries—the probability that if one selects at random an individual in each country, both will be ethnic Chinese. Focusing on countries in which ethnic Chinese constitute at least 1% of the population (eg, as in all of Southeast Asia), this variable has remarkable predictive power for bilateral trade, suggesting trade would be on average around 60% lower in the absence of the Chinese ethnic

[av] A number of other papers show that measures of affinity between regions, such as trust, telephone volume, and patterns of historical migration, correlate strongly with other aggregate outcomes, such as foreign direct investment (Guiso et al., 2008b) and international asset flows (Portes and Rey, 2005).

network. This conditional correlation is stronger for trade in differentiated goods, which is consistent with the view that the Chinese ethnic network facilitates trade by reducing asymmetric information. Combes et al. (2005) find a similarly strong conditional correlation between the volume of trade between a given pair of French regions and the bilateral stock of migrants between the two regions, where they measure the stock of migrants as the number of individuals born in region *i* that work in region *j*. Once they add this variable to their gravity equation, the estimated effects of distance and region borders on trade are reduced to a much more reasonable level (both effects drop by about 50%). Garmendia et al. (2012) find similar results for Spain.

The main conclusion from this set of papers is that failing to account for the effect of social ties across borders and within countries may explain why traditional gravity equations have found unreasonably large effects of distance and country borders on trade. However, none of the three studies attempt to deal with potential endogeneity or reverse causality, such that the conditional correlations between trade flows and social ties that they document should not be interpreted as causal effects. In other words, the effect of social ties on trade flows is not identified in these papers for the same reason that (17) is not identified without an appropriate instrument.

Three recent studies address this issue using different natural experiments. Parsons and Vezina (2014), Cohen et al. (2014), and Burchardi et al. (2015) evaluate the causal impact of migrant networks on the variation in international trade and FDI across locations in the United States. To the extent that these papers look at trade and FDI originating from the same country, the United States, some of the concerns of reverse causality are mitigated: the regulatory environment, most direct barriers to trade and investment, as well as the ease with which migrants from different countries can emigrate are all relatively uniform.

Parsons and Vezina's strategy is similar to that of Burchardi and Hassan (2013) in the sense that they study variation in social ties that results from a historical migration that occurred while economic interaction with the migrants' region of origin was impossible. At the end of the Vietnam war, the US government imposed a trade embargo on Vietnam and evacuated 130,000 Vietnamese citizens to the United States. Upon arrival at one of four processing centers located in Arkansas, California, Pennsylvania, and Florida, charitable organizations were charged with finding sponsors who were willing to provide food and shelter for the refugees. Parsons and Vezina stress that a main objective of this process was to disperse the Vietnamese refugees as much as possible across the United States to avoid an agglomeration of refugees, similar to that of Cubans in Florida. Consistent with this view, they show that the resulting allocation of refugees is uncorrelated with a range of variables that may proxy for the potential for trade with Vietnam, which is quite plausible because of the trade embargo that was in force at the time the refugees were allocated. Instead, they argue that the variation in the allocation of refugees across states was driven by quasi-random variation in the capacity of the charitable organizations operating in different states. Importantly, Parsons and Vezina also show that the initial allocation of this first wave of refugees in 1975 is highly predictive of the location of ethnic Vietnamese in 1995, the year in which the trade embargo was finally lifted. Their structural equation of interest takes the form

$$X_i = \beta V_i + C_i'\gamma + \epsilon_i, \tag{18}$$

where X_i is the average share of exports of state i to Vietnam between 1995–2010, V_i is the stock of Vietnamese migrants in 1995, and C_i is a of controls. The standard specification uses the allocation of refugees across the United States in 1975 as an instrument for V_i. The identifying assumption for a causal interpretation of β is thus that the initial allocation of refugees is uncorrelated with ϵ_i. Subject to this assumption, Parsons and Vezina's main specification implies that a doubling of the population share of Vietnamese migrants relative to the mean increases the ratio of exports to Vietnam over GDP by 19.8%.

Cohen et al. (2014) instead use the forced relocation of ethnic Japanese into Japanese internment camps during World War II as an exogenous shock to the location of ethnic Japanese across US Metropolitan Statistical Areas (MSAs). These camps were established in remote areas (away from any industrial activity that may have been considered sensitive to the war effort) to house Japanese-Americans who lived predominantly on the West Coast prior to the bombing of Pearl Harbor. After the war, the residents were released from the camps. However, having lost their jobs and sold off their possessions in their regions of origin, they often resettled in MSAs that were geographically close to the location of their internment. Cohen et al. (2014) show that this relocation had a persistent effect on the regional distribution of the Japanese-American population that lasts to the present day: MSAs that are within a 250-mile radius of the location of a former internment camp have a 62% larger Japanese population today than other comparable MSAs. To corroborate this finding, they show, importantly, that internment camps predict higher populations of Japanese-Americans but not of other Asian-Americans.

This persistent effect on the location of Japanese-Americans has a sizable effect on the probability that firms located in a given MSA trade with Japan. Using the location of Japanese internment camps as an instrument, they show that a one-standard-deviation increase in the share of an MSA's population that are Japanese-Americans doubles the likelihood that a given firm will export to Japan. They find similarly large effects on the probability of importing, the volume of trade, and even on the likelihood that a given MSA has a sister city in Japan today.

Although both Parsons and Vezina (2014) and Cohen et al. (2014) make convincing arguments for a causal link between social ties and trade, both use natural experiments that resulted in a shock to the allocation of migrants from one particular country, such that they cannot control for unobserved destination effects as one would in a gravity equation. Assessing the external validity of these results and how they generalize to other ethnicities is therefore difficult.

To quantify the more general causal effect of social ties on US trade and FDI, Burchardi et al. (2015) instead study the natural experiment that arises from the entire history of settlement of the United States. Their strategy isolates quasi-random variation in the allocation of migrants across destinations within the United States that results from the interaction of two facts: First, migrants from different origins tended to arrive in the United States at different times. Second, the set of destinations that are most economically attractive to the typical migrant arriving in the United States changed over time.

They motivate their approach using a reduced-form dynamic model of migrations. Migrations from a given foreign origin country *o* to a given destination county *d* in period *t* depend on the total number of migrants arriving in the United States from *o* (a push factor), the relative economic attractiveness of destination county *d* to migrants arriving at the time (a pull factor), and the size of the preexisting local population of ancestry *o*, allowing for the fact that migrants tend to prefer settling near others of their own ethnicity (a recursive factor). Solving the model shows the number of residents today who are descendants of migrants from *o* is a function of simple and higher-order interactions of the sequence of pull and push factors. Burchardi et al. (2015) then use these interactions to construct instruments for each US county's present-day ancestry composition.

To prevent omitted variables that affect both migrations and FDI from driving their results, they measure the pull factor of each US destination for migrants from *o*, using the number of migrants arriving in *d* at the same time from a continent other than *o*'s continent of origin. That is, they predict a migrant's choice of destination within the United States using the revealed behavior of the average migrant arriving at the same time but from a different continent. Similarly, they measure the push factor using the total number of migrants arriving in the United States from *o* at time *t*, excluding those who settle in the vicinity of *d*. Interacting these measures of pull and push factors for each vintage of census data since 1880 isolates variation in the present-day ancestry composition of US counties that derives solely from the interaction of the staggered arrival of migrants from different origins with time-series variations in the relative attractiveness of different destinations within the United States.

In their main specification, Burchardi, Chaney, and Hassan find that doubling the number of individuals with ancestry from a given foreign country increases by 4.2 percentage points (or 237% relative to the mean) the probability that at least one firm from that US county engages in FDI with that country. The main identifying assumption for a causal interpretation of this result is that omitted variables making a given location within the United States differentially more attractive for migrants from a specific origin for both settlement and FDI do not affect the location choices of the average migrant originating from other continents and simultaneously have large effects outside of the surrounding states of the destination in question.

Flexibly applying this instrumentation strategy to the entire set of origins and destinations, the authors estimate heterogeneous effects of ancestry on FDI across origins and destinations. Across origins, they find that the effect increases with the geographic distance and the judicial quality of the origin country. Across destinations, they find consistent evidence of a positive impact of regional diversity on FDI: the more diverse and less ethnically homogeneous the local population, the larger the total effect of ancestry on FDI. Similarly, the effect of ancestry on FDI falls with the population of migrants from the same origin in neighboring counties, and from neighboring origins, so that a small minority from a distant part of the world that is not otherwise represented in the local ethnic mix has the largest marginal impact on FDI.

By instrumenting separately for each wave of immigration, Burchardi et al. (2015) are also able to distinguish the effect of first-generation immigrants from the effect of their descendants. They find a significantly smaller effect for the first generation, implying the full effect of ancestry on FDI is long lasting and takes multiple generations to fully unfold.

Simultaneously instrumenting for migration from multiple origins also enables the authors to base their estimation on a gravity equation, including destination fixed effects that control for differences in size, market access, and productivity across US destinations. Although the results with respect to FDI remain stable regardless of the inclusion of destination and other (more complicated) fixed effects, the authors find no systematic causal impact of ancestry on the patterns of international trade of US states whenever they control for destination fixed effects. This latter finding is in stark contrast to the results in Parsons and Vezina (2014) and Cohen et al. (2014), and suggest that the causal relationship between social ties and trade found in these papers may not generalize to a larger set of origin countries.

4.3.3 The Effect of Internal Social Structure on Institutions and Growth

Although the literature on social ties focuses on the external relationships of a social entity, a number of studies have also used natural experiments to assess the effect of the *internal* relationships between different groups within a given society on aggregate economic outcomes, and in particular on the ability of a society to develop a functional political system and "good" institutions.

Perhaps the most convincing of these studies is Dippel (2014), who considers the natural experiment surrounding the formation of Native American reservations. In the 19th century, the US government formed several reservations consisting of members of ethnically and linguistically homogeneous tribal bands. While respecting ethnic differences, this process largely ignored differences in historical institutions, such that some (mixed) reservations received constituents of several previously politically independent subtribal bands while others did not. Dippel shows that these mixed reservations have significantly worse contemporary economic outcomes, even when conditioning only on variation within reservations belonging to the same tribe. To account for potential confounding factors in the formation of reservations, he instruments the likelihood of a mixing of several previously independent bands with historical mining activity in the historic homeland of the tribe, where mining activity generated incentives for the US government to form fewer, and thus more likely mixed, reservations. He then shows that the majority of the divergence in economic outcomes appears only after the 1980s, when the Bureau of Indian Affairs ceded reservation governance to the local reservations. Using information on contemporary political conflict and corruption within reservations, he argues convincingly that the adverse economic effects are explained by the fact that mixed reservations tend to have more dysfunctional political institutions.

Another set of studies has linked the emergence of good institutions to the relative influence of the middle class. Acemoglu et al. (2005a) use a difference-in-differences approach to argue that Western European countries that developed a large and politically influential merchant class were able to develop constraints on the executive and safe property rights that were crucial for subsequent economic growth. Their evidence comes from a panel data set of GDP per capita, institutional quality, and the number of Atlantic voyages undertaken by each European country 1300–1850. They show that the onset of the Atlantic trade after 1500 was a major positive shock to GDP per capita in countries on the Atlantic coast. However, this shock led to an increase in constraints on the executive only in Atlantic countries that already had relatively higher constraints on the executive in medieval times: the interaction between Atlantic trade and medieval institutions explains most of the variation in institutional quality across European countries. The authors argue that these results are consistent with their view that a merchant class could only emerge in countries in which rulers were relatively constrained and did not monopolize the Atlantic trade.

Although AJR (2005) provide historical and anecdotal evidence to corroborate their interpretation, a problem with their analysis is that they do not observe the size of the merchant class directly. More generally, a major challenge for the literature attempting to establish a causal link between social structure and institutions is that consistent measures of social structure are rarely available for a sufficiently long period of time, and in particular for historical episodes in which one might observe quasi-exogenous variation in social structure.

While stopping short of claiming success at having identified a causal effect, Acemoglu et al. (2011b) make some progress in this dimension by studying the mass-murder of Jews following the Nazi invasion of Russia during World War II. Uniquely, Soviet authorities kept extensive records of the size and ethnic composition of the middle class (white-collar workers) in Russian regions (oblasts), dating back to 1926. Before the outbreak of World War II, Jews were heavily overrepresented in white-collar occupations, such that their persecution and murder by the occupying forces represented a significant shock to the size of the middle class. Using variation both within oblasts occupied by the Nazis and across occupied and nonoccupied oblasts, the authors show that oblasts in which the Holocaust most severely reduced the size of the middle class have worse political and economic outcomes today. These oblasts grew less since 1945 both in terms of population and GDP per capita, exhibited greater vote shares for communist candidates during the 1990s, and more support for preserving the Soviet Union in a referendum held in 1991. Moreover, the shock to the relative size of the middle class in oblasts most adversely affected by the Holocaust appears to persist over time, until the last Soviet census held in 1989. Acemoglu, Hassan, and Robinson argue that their evidence is consistent with the view that a shock to the size of the middle class may have permanent effects because it reduces the core constituency for policies that advance constraints on the executive and safe property rights (most notably in the form of more political support for the preservation of communism).

4.4 Trust and Civic Capital

The literature on culture and economics offers various competing definitions of social capital, culture, trust, and related concepts. For the purposes of structuring our discussion, we focus on the concept of "civic capital," which Guiso et al. (2012) define as "those persistent and shared beliefs and values that help a group overcome the free rider problem in the pursuit of socially valuable activities." We prefer to use this relatively narrow definition mainly because it allows a convenient grouping of the existing empirical literature. First, it clearly distinguishes civic capital from social structure, the focus of the literature discussed in the previous section. We may loosely think of "social capital" as the union of the two concepts. Second, it closely describes the variables used in the empirical literature, which often focuses on various measures of how much individuals are willing to trust a stranger, and other measures of beliefs about the intentions and actions of others.

The idea that the set of norms and beliefs that make cooperation among individuals easier should be a driver of economic growth has a long tradition in the social sciences, going back at least to Banfield (1967), Coleman (1988), and Greif (1993). Putnam et al. (1993) famously argued that Southern Italy is less developed economically than the North because of a lower level of civic capital, and conjectured that this difference is a result of the fact that Northern cities had a long tradition of self-rule, which fostered a tradition of civic engagement, at a time when Southern cities were tightly controlled by Norman kings.

The basic idea in this literature is that culture in general and civic capital in particular changes only slowly over time. Parents pass on beliefs and values to their children, such that civic capital is akin to a slow-moving, accumulated factor. Societies with a higher level of civic capital have a higher capacity for economic growth because they develop better tools for overcoming market failures and collective action problems.

The prime example for such a belief is trust. Even in a society with an efficient police force and functioning courts, most commercial transactions involve some element of trust (Arrow, 1972). For example, when you hire an accountant to do your taxes, you trust that she will not abuse your private information to commit credit card fraud. If you cannot trust your accountant, you might prefer to do your taxes yourself. Similarly, when you take a taxi, you trust that the driver knows how to drive, is not intoxicated, has not manipulated the meter, and will not simply lock the door, drive you off to the desert, and hold you for ransom. Although you can take accountants and taxi drivers to court, doing so will cost time and money, and even the most efficient court systems cannot enforce all the rules all the time. Other transactions that rely on trust include employment contracts in which managers cannot perfectly monitor employees, sales in which goods are delivered before or after payment is made, and many financial transactions and investment decisions. (Thinking about it this way, the amount of trust people in developed societies put in perfect strangers is remarkable!) The more complex an economy becomes, and the more labor is divided into specialized tasks, the more important may be the shared belief that strangers can generally be trusted.

4.4.1 The Effect of Trust on Growth

A number of papers have used cross-country data to document a conditional correlation between measures of civic capital and economic development. Knack and Keefer (1997) show that measures of trust and civic cooperation are strongly associated with higher GDP, higher economic growth, and higher investment-to-GDP ratios, even after controlling for education, institutions, and other factors. However, these results do not speak to causation. The obvious problem is that people who live in wealthy countries with good institutions may rationally put more trust in strangers because they know they will be partially protected by a well-functioning police force and efficient courts. Clearly, institutions, civic capital, and economic development are mutually interdependent variables. A key challenge in demonstrating a causal effect of civic capital on growth is thus not only to identify exogenous variation in civic capital, but also to separate its effect from the effect of institutions.

Three papers attempt to tackle this challenge using natural experiments that rely on the idea that civic capital depends on the experiences of each generation and is at least partially transmitted from one generation to the next.[aw] For example, one might expect individuals whose parents grew up under an authoritarian dictatorship to be less trusting than otherwise similar individuals whose parents grew up in a democracy.

Tabellini (2010) uses the fact that some Western European countries emerged as the union of several very heterogeneous historical political entities. He studies the variation in gross value added per capita across 69 regions within these countries. To the extent that present-day institutions vary only at the country level, rather than the regional level, country fixed effects absorb any differences in current institutions. He then instruments

[aw] Several studies show trust indeed has an inherited component; see Rice and Feldman (1997), Putnam (2000), and Guiso et al. (2006).

for current measures of civic capital using literacy in 1880 and a measure of constraints on the executive between 1600 and 1850, while controlling for urbanization in 1850. The key identifying assumption is that these historical instruments affect present-day economic development only through their effect on the persistent component of civic capital. Conditional on this (rather demanding) assumption, Tabellini shows that the exogenous component in civic capital has a large positive effect on regional economic development. A more conservative interpretation of the same fact, that is nevertheless interesting in its own right, is that distant political history is an important determinant of current economic performance not just across but also within countries.

Guiso et al. (2008a) consider a similar experiment within Italy. In the Northern part of Italy, some cities achieved the status of a free city during medieval times, whereas others remained under the control of a feudal lord or the Holy Roman Emperor. Consistent with the results in Tabellini (2010), they find that those cities that achieved self-rule earlier (by 1136 or 1300 CE) exhibit higher measures of social capital today. However, Guiso et al. (2008a) then go one step further by instrumenting the dummy variable for early self-rule with two variables that they argue historically affected the cost of achieving self-rule but are unlikely to affect directly the level of civic capital or the level of output today: whether the city was a seat of a bishop who may have been able to coordinate the struggle for independence, and whether the city was founded in pre-Roman (Etruscan) times and is therefore located in a geographic position that is easy to defend militarily. Using these two instruments, they confirm that a longer history of self-rule is significantly associated with higher social capital and higher GDP per capita today.

Although more robust than Tabellini (2010), the identifying assumption for a causal interpretation of the results in Guiso et al. (2008a) is still a tall order: self-rule, the two variables driving the cost of self-rule, or any omitted variables correlated with these measures cannot have a direct effect on GDP per capita today, except through their effect on civic capital. Even if formal institutions did not vary within Italy (which they do to some extent), administrative capacity, the functioning, and the quality of these institutions surely vary across regions even though they might follow the same letter of the law. As a result, both of these natural experiments may still confound the effects of civic capital and institutions or any other omitted variable that is correlated with these variables.

Two recent papers attempt to tackle this problem using clever strategies that are not natural experiments according to our definition. Algan and Cahuc (2010) use the inherited trust of descendants of US immigrants covered in the General Social Survey to recover a long time series of trust for their origin countries, reaching back to the beginning of the 20th century. This long time series then allows them to relate changes in GDP per capita to changes in inherited trust over several generations, while controlling for all of the time-invariant effects that complicate the interpretation of the results in Tabellini (2010) and Guiso et al. (2008a). In addition, they also control for time variation in the quality of institutions in order to convincingly distinguish the effect of institutions from the effect of trust. Their estimated effect of trust is positive, highly statistically significant, and quantitatively large. For example, according to these estimates, GDP per capita in Russia and Mexico would have been 60% higher had these two countries inherited the same level of trust as Swedes. Gorodnichenko and Roland (2010) also rely on the idea that cultural traits are inherited and instrument culture (in their case, a measure of individualism rater than trust) with the genetic distance of the population to the most individualist countries in the world (the United States and United Kingdom). Consistent with the other studies, they also find a large effect of culture on growth.

4.4.2 Effect of Trust on Financial Development and Other Aggregates

Although the papers surveyed in the previous section may have convincingly identified a causal effect of civic capital on growth, they have little to say about the mechanism through which this effect transmits itself and about *how* civic capital affects growth. One obvious candidate is financial development: financial contracts are arguably "trust intensive," in the sense that handing over cash to a stranger today in the hopes of receiving returns in the future requires a large amount of trust in that stranger. Higher levels of civic capital may thus enable a society to sustain a more sophisticated financial system that may then in turn facilitate economic growth. Guiso et al. (2004) study this channel.

Similar to Guiso et al. (2008a), this paper measures variation in civic capital across regions within Italy and relates these measures to the use of financial instruments by households responding to a survey by the Italian central bank. The identification strategy again relies on the idea that part of social capital is inherited from previous generations. Thus, when a household moves from one Italian region to another, the level of civic capital (but not the quality of institutions) in its region of origin still influences its behavior. In their main specification, Guiso et al. (2004) relate a household's use of financial instruments to the level of civic capital (measured as voter turnout or the volume of blood donations) in its region of origin, a set of region fixed effects, and a number of household-level controls. They find that households that originate in regions with higher levels of social capital are more likely to use checks, invest more in the stock market, and rely less on informal loans from friends and family. These effects tend to be stronger in regions with weak law enforcement.

The authors' preferred interpretation of these results is that the civic capital plays an important role in the degree of financial development across Italy. If these results generalize to the variation in financial development across countries, they may thus account for part of the observed effect of civic capital on economic growth.

Apart from this evidence of a financial channel, additional evidence on the mechanism by which civic capital affects growth is scarce, and if it exists it is largely not causally identified. A promising avenue for future applications of natural experiments may be the relationship between trust and regulation. For example, Aghion et al. (2010)

document that levels of trust and the level of government regulation are strongly negatively correlated across countries.

4.4.3 Determinants and Dynamics of Trust

The main conclusion from the series of studies that document an effect of civic capital on growth and financial development is that, from a macroeconomic perspective, we may think of civic capital as a slow-moving state variable that codetermines a society's capacity for economic growth. A crucial question then is what governs the dynamics of this state variable. That is, how do some societies end up with a high level of civic capital while others suffer from low levels of civic capital? The existing literature has examined natural experiments that identify three factors determining the dynamics of civic capital: historical institutions, experiences of violence and conflict, and the climate.

4.4.3.1 Historical Institutions

As part of their identification strategies, Tabellini (2010) and Guiso et al. (2008a) show that historical institutions appear to affect the level of trust, decades or even centuries later. Both papers show that the descendants of residents that lived in areas that had more constraints on the executive historically exhibit higher levels of trust today.

Becker et al. (2015) examine a natural experiment in which one might expect similar results. Parts of five Eastern European countries (Montenegro, Poland, Romania, Serbia, and Ukraine) were under the rule of the multiethnic Habsburg empire until the end of World War I. In some parts, this rule lasted for hundreds of years. Compared to both its contemporaries and some of its successor states, the Habsburg empire had a reputation for having a restrained and effective bureaucracy, courts, and police. The descendants of residents of areas formerly under Habsburg control thus had a longer history of living under "good" institutions in this sense than their present-day countrymen. Becker et al. (2015) study the effect of this treatment using a regression-discontinuity design.

Their main specification relates responses from a survey covering individuals in all five countries to their location relative to the former border of the Habsburg empire. They find that individuals living within 200 km of the former Habsburg side of the border are not significantly more trusting of strangers or more likely to be members of a civic organization than their countrymen on the other side of the former border. Although this finding may be due to a lack of power in their specification, they do find that individuals on the former Habsburg side are significantly more trusting of the police and less likely to pay bribes to officials.

4.4.3.2 History of Violence or Conflict

One interesting detail in the results of Algan and Cahuc (2010) is that inherited trust in Sweden increased after 1935 while it decreased in continental Europe and the United Kingdom. They conjecture that this differential change may be the effect of World Wars I and II: individuals that experience violence and conflict may pass down a lower level of trust to their descendants. Two papers examine this link between violence and trust in different historical settings.

Nunn and Wantchekon (2011) link the variation in levels of trust across individuals in Africa to the history of slave trade. They argue that the demand for slaves by Europeans (and later Americans) based predominantly in Atlantic ports created conditions that would result in distrust among the indigenous population. Particularly in the later phases of the slave trades, individuals found themselves enslaved not as the result of inland raids by foreigners but as the result of kidnappings or trickery on the part of other members of the indigenous population. By selling others into slavery, one could obtain the means to purchase iron weapons, thus protecting oneself from enslavement. A number of historical sources show that the majority of individuals were sold into slavery by kidnappers or even by their own relatives. Exposure to the slave trade may therefore have severely reduced the level of trust in the affected societies.

Nunn and Wantchekon's structural equation of interest takes the following form:

$$trust_{j,e,d,i} = \alpha_i + \beta\ slave\ exports_e + X'_{j,e,d,i}\Gamma + X'_{d,i}\Omega + X'_e\Phi + \epsilon_{j,e,d,i}, \tag{19}$$

where α_i denotes country fixed effects; *slave exports*$_e$ measures the number of slaves taken from ethnic group e during the slave trade per square kilometer of area settled by the ethnic group; $X'_{j,e,d,i}$ denotes a rich set of individual-level controls including ethnicity, education, and age; $X'_{d,i}$ controls for the ethnic composition of district d in country i; and X'_e is a vector of ethnicity-level controls that capture subnational variation in colonial rule, in particular, the disease environment and measures of precolonial prosperity.

In their main specification, Nunn and Watchenkon instrument slave exports with the distance of an ethnic group from the coast, which is where European slave traders kept their bases. Because geographic features in general are correlated with all kinds of things, they then make a careful argument that, conditional on the controls they include in their standard specification, the distance to the coast is plausibly uncorrelated with other factors that affected trust. First, they argue that Africans did not engage in overseas trade before the slave trade, such that distance to overseas trade is not a confounding factor in their case. Second, they stress the importance of the ethnicity-level controls for other forms of European contact. Third, they include additional controls for each ethnicity's historical reliance on fishing. Conditional on these controls, they argue that the exclusion restriction is plausibly satisfied. Their estimates show that a one-standard-deviation increase in exposure to the slave trade is associated with approximately a 0.2-standard-deviation decrease in various measures of trust of neighbors and trust of other ethnicities.

A fairly common problem with papers using large natural experiments such as the enslavement of Africans is that unobservables could potentially bias the result. For example, in Nunn and Wantchekon (2011), we may worry that despite the large number of controls the authors propose, differences in preexisting trust and prosperity or some other unobservable correlated with slave trade and trust may not be adequately accounted for. To assuage these concerns, Nunn and Wantchekon use a technique developed by Altonji et al. (2005) and Bellows and Miguel (2009) that calculates how much stronger selection on unobservables, relative to selection on observables, would have to be to overturn the estimated effect:

$$\hat{\beta}^R / (\hat{\beta}^R - \hat{\beta}^F),$$

where $\hat{\beta}^R$ and $\hat{\beta}^F$ are the coefficients of interest estimated with a restricted and the full set of controls, respectively. If including observable covariates does not have a large effect on the coefficient of interest, this number is large and selection on unobservables would have to be multiple times more severe than selection on observables to overturn the result. Nunn and Wantchekon find that including their full set of controls changes their coefficient of interest so little that this statistic is 3 in all of their specifications. Whatever their specifications are missing would thus have to have a very large selection effect to overturn their qualitative result.

A second paper probes the relationship between inter-state conflict and trust. Jancec (2012) uses a difference-in-differences approach to show that individuals living in regions within the present-day countries of Slovenia, Croatia, Serbia, Montenegro, Romania, and Ukraine that experienced more frequent changes in the ruling nation state between 1450 and 1945 exhibit lower trust in political institutions today. However, similar to Becker et al. (2015), he finds no effect on measures of civic behavior and trust toward strangers.

4.4.3.3 Geography and Climate

Rather than identifying "man-made" shocks, such as wars and historical migrations, Durante (2010) takes a more radical approach and links civic capital directly to environmental factors that determine the need for cooperation. The idea (similar to Ostrom, 1990) is that the earliest societies, centered around subsistence farming, would develop a culture of cooperation and trust where it is needed for survival. He argues that in regions where precipitation and temperature are highly variable from year to year, societies needed to develop civic capital to sustain investment in irrigation and other large works that facilitated survival. Similarly, in regions with very diverse climatic conditions, developing civic capital increases the probability of survival, by facilitating trade and risk-sharing. Using long-term climate data reaching back to 1500 CE, he indeed finds a significant association between trust and these climatic variables in the cross section of European regions. Interestingly, these results are robust to including country fixed effects and controlling for early institutions as in Tabellini (2010), suggesting that part of the effect of climate on trust may indeed transmit itself through civic capital.

4.5 Multiple Equilibria and Path Dependence

Some growth models that feature nonconvexities produce multiple equilibria, such that the mapping between the steady state GDP per capita and ϕ_i in (13) is not unique. For example, Murphy et al. (1989) show that multiple equilibria can arise in a simple model with monopolistic competition and a fixed cost of production. In their model, a given firm finds that incurring the fixed cost of production is profitable only if other firms do the same, due to a demand spillover. In the "bad" equilibrium, firms do not invest, because they expect that other firms will also not invest. This scenario is an example of a coordination failure—economic growth does not happen because economic actors have the "wrong" expectations and cannot coordinate to invest simultaneously. Although it seems implausible that countries might be persistently poor just because its residents cannot coordinate to simultaneously change their expectations, many more complicated models also feature multiple steady states. In these models, long-run GDP per capita is path-dependent, and once an economy finds itself on the path to a bad steady state, it may be hard to reverse its trajectory.

To our knowledge, no single natural experiment has been used to test the hypothesis that multiple steady states may explain cross-country income differentials. Part of the reason for the absence of such an experiment might be that viewing the income differential between the Congo and the United States purely as the result of a historical accident seems unsatisfactory. Instead, the literature has focused on the more modest goal of showing that the interaction of nonconvexities and historical accidents can have an effect on the sectoral composition of production or on its spatial distribution within a given country.

Juhasz (2014) uses data on 19th-century France to show a causal effect of temporary trade protection during the Napoleonic wars on the long-run location of the cotton-spinning industry in France. She argues that the continental blockade that prohibited direct shipping between Britain and France during the war differentially affected the costs of shipping goods from Britain to different regions within France. In particular, regions in Northern France that historically had relatively low costs of trading with Britain were temporarily more protected from British imports, because these imports now had to be shipped through Spain. Under the identifying assumption that more and less protected regions would have developed similarly absent the effect of the continental blockade on trade costs, she shows that relatively more protected regions significantly increased their capacity in mechanized cotton spinning (a new technology at the time). She then shows that this change in the regional distribution of the cotton-spinning industry within France persisted 30 years after the end of the blockade. Her results are thus consistent with nonconvexities in the adoption of new technologies, and the view that infant industry protection can have a lasting effect on the location of production within and possibly also across countries.[ax]

[ax] See Kline and Moretti (2014) for similar evidence from the Tennessee Valley Authority, where subsidies appear to have permanently increased the level of manufacturing employment.

A large body of work in urban economics has used natural experiments to examine this question at the scale of cities. Although the focus of this literature is not explicitly on macroeconomic variables, we summarize it briefly, keeping in mind that factors of production, in particular labor, are much less mobile at the country or region level, such that it is unclear how results would change at larger levels of aggregation. In an influential study, Davis and Weinstein (2002, 2008) use a difference-in-differences approach to show that the bombing of Japanese cities during World War II had no long-run effect on the relative size of cities and even their industrial composition, suggesting that fundamentals rather than chance govern the long-run spatial distributions of these variables. Miguel and Roland (2011) show similar results using an instrumentation strategy for the US bombing of Vietnam.

Bleakely and Lin (2012) show contrasting evidence that a temporary locational advantage can have long-lasting effects on population density. They argue that many cities in North America formed in places where natural obstacles, such as waterfalls, blocked continued water transport and thus required overland hauling. These places (portage sites) attracted transportation services and commerce such that large settlements and cities would often form. However, this natural advantage is no longer relevant today, because trains, trucks, and airplanes have supplanted ships as the primary transportation technology, and locks and canals now make hauling cargo from one ship to another unnecessary. In this sense, technological progress generated a temporary positive shock to locational fundamentals of portage sites that plausibly no longer exists today. Nevertheless, Bleakly and Lin show not only that portage sites are significantly associated with higher population density today (including large cities such as Washington DC and Philadelphia), but also that no evidence exists of a relative decline in these areas, because their natural advantage as portage sites dissipated. Their findings are thus consistent with the existence of multiple equilibria in the location of cities and towns.

A series of papers in this literature also uses the division and reunification of Germany as a natural experiment. Most directly focused on testing for multiple equilibria is Redding et al. (2011). Using a simple difference-in-differences approach on a panel of passenger-traffic data for German airports, they show that the division of Germany led to a significant increase in the growth rate of passenger traffic in Frankfurt (previously a minor destination) and to a simultaneous shrinking of passenger traffic at Berlin airport (the previous main hub). However, after reunification, this trend did not reverse, leaving Germany's main airport hub in Frankfurt. Using various methods the authors argue that this apparently permanent shift cannot be explained by fundamentals, and instead is evidence of multiple equilibria in the location of a country's main airport hub.

Two closely related papers document evidence that is consistent with a specific source of multiple equilibria—the agglomeration externalities that operate in new economic geography models. The basic idea in this literature is that consumers like locating close to firms, and that firms like being close to other firms for various reasons, most commonly to economize on commuting and trade costs (Krugman, 1991). If these agglomeration forces are strong enough relative to dispersion forces (eg, the costs of congestion), multiple equilibria may arise. One simple prediction of these models is that market access is an important driver of economic development. Redding and Sturm (2008) test this hypothesis using German division and reunification. Using a difference-in-differences approach, they find that over the 40-year period of German division, the population of West German cities close to the inner-German border declined significantly relative to other West German cities, a finding that is consistent with the predictions of the model. Their main identifying assumption is that, absent the effect of the inner-German border on market access, cities closer to and farther away from the border would have developed similarly. In Ahlfeldt et al. (2015), the authors take the analysis one step further and use the division and reunification of Berlin as a natural experiment, allowing them to quantitatively estimate agglomeration and dispersion forces at the city level. This paper is methodologically distinct from all the other papers covered in this chapter in that it focuses explicitly on structural, rather than reduced-form estimation. The authors develop a quantitative model of city structure that features multiple agglomeration and dispersion forces. They are able to separately identify these forces, because of the variation in the surrounding economic activity of city blocks that results from the division and reunification of the city. Interestingly, the identifying assumption for this structural model is just a sharper version of that used in the previous paper: that Berlin's division and subsequent reunification affects the systematic change in the pattern of economic activity across city blocks only through its effect on commuting costs and changes in access to production and residential externalities.

Taken together, these studies seem to suggest that historical accidents can have an effect on the spatial distribution of production within a given country. However, it is important to note that, conceptually, none of these studies can distinguish multiple equilibria from long-term persistence. Moreover, none of them speak to the existence of multiple equilibria at the country level.

5. CRITICAL ASSESSMENT AND OUTLOOK

In this chapter, we describe the use of natural experiments in macroeconomics for three distinct purposes: to verify underlying model premises, to quantify policy parameters, and to identify causal mechanisms that are absent from conventional models. We do this by covering the use of natural experiments in the literatures on the Permanent Income Hypothesis, on fiscal multipliers, and on the fundamental causes of growth.

An easy test of the assumption of forward-looking behavior of the Permanent Income Hypothesis can be carried out if one can identify instances when households receive payments resulting from a preannounced change in income. A series of natural experiments that identify such preannounced changes finds that agents adjust consumption at receipt, which is in contrast to the assumption of forward-looking rational behavior. Some of these experiments suggest binding liquidity constraints can partly, but not entirely, explain this finding. They tend to find that more liquidity constrained households react more strongly at receipt, but even unconstrained households appear to adjust their consumption upon receiving a preannounced payment. The pattern of results across different types of natural experiments suggests a degree of near-rational behavior: in response to large income changes, households do appear to adhere to the Permanent Income Hypothesis, and only when faced with small income changes does the evidence not line up with the model predictions. Modeling near-rationality explicitly and embedding a natural experiment directly into such a model would be useful to generate further insights. Moreover, analyzing new measures of liquidity constraints that incorporate wealthy hand-to-mouth consumers could possibly help reconcile some of the conflicting evidence on the importance of liquidity constraints. A clear disadvantage of natural experiments also becomes apparent in this application: analyzing preannounced income *decreases* would directly rule out the possibility that binding liquidity constraints drive the results, and analyzing *large* preannounced income changes would rule out near-rational behavior. Whereas a researcher running field or laboratory experiments would thus make sure any given experiment covers both features, natural episodes encompassing these two features are rare.[ay]

The literature on the fiscal multiplier faces the challenge of identifying fiscal policies that are orthogonal to current economic conditions. Two approaches involving natural experiments can help: the first uses increases in military spending caused by geopolitical events as exogenous shocks to national fiscal policies, and the second exploits cross-regional variation at the subnational level driven by historical episodes to establish exogeneity. Whereas the first approach typically estimates aggregate fiscal multipliers smaller than 1, the second approach consistently returns estimates of the local fiscal multiplier between 1.5 and 2. However, these aggregate and regional multipliers are different concepts and cannot be directly compared. A model is needed to gain further insights into their relationship.[az]

Identifying the fundamental reasons some countries are rich while others are poor is a major challenge in economics. The main empirical difficulty is that most of the likely drivers of growth, such as the accumulation of physical and human capital, the quality of institutions, the level of trust, and social structure, depend in complicated ways on the level of income and on each other. In the last two decades, natural experiments have

[ay] Replicating large income changes in a designed experiment would, however, be costly and thus also difficult.

[az] See the discussion of Nakamura and Steinsson (2014) in Section 3.

become the most widely used tool to try to resolve this question by identifying the causal determinants of economic growth. Using such large-scale natural experiments as the age of European colonization or German reunification, the literature has produced convincing evidence that the major fundamental causes of growth are outside traditional models of economic growth. Instead, cross-country variation in political and economic institutions, cultural traits, and social structure appear to explain the majority of cross-country and often even the majority of within-country variation. The literature has been less successful at distinguishing these causal channels from each other. In particular, much of the work causally linking institutions and trust to growth does not convincingly differentiate between these two channels. In addition, much work remains to be done in trying to understand the dynamics and the interaction of these fundamental causes over time. The study of social structure, meaning social ties and networks as well as the social stratification of society, as a fundamental determinant of economic growth is a new, promising field that has arisen in this context, allowing a deeper understanding of the direct effect of social structure on growth and its effect on the evolution of institutions. A common thread across the different studies identifying the fundamental causes of economic growth is that they focus almost exclusively on the reduced-form relationship between the fundamental causes and economic outcomes, but make no use of the structure of existing models of economic growth. Bridging the gap between these two literatures may prove a promising avenue for future research.

How researchers use natural experiments in practice differs between the three literatures we have covered. The literature on the Permanent Income Hypothesis employs natural experiments directly in identifying preannounced income changes, whereas the other two strands of the literature mainly rely on microeconometric techniques such as instrumental variables and regression discontinuity designs. Some studies that do not rely on specific historical events for identification also use these techniques. As a result, deciding what can and cannot be considered a natural experiment in these latter two strands of literature involves some judgment.

The fundamental challenge, however, is the same in all three literatures, namely, to argue the historical episode in question provides the quasi-random variation that is necessary to identify causal effects. Although creating a comprehensive checklist that all analyses involving natural experiments can use is impossible, we list below some common features that distinguish successful papers that rely on the use of natural experiments:

Identifying assumption: The identifying assumption underlying the study should be clearly stated. What does the reader need to believe to causally interpret the effect? What aspect of the natural experiment does the study exploit and how does it translate into exogenous variation? Why can reverse causality be ruled out? Can we think of any omitted factors that jointly affect the natural experiment and the outcome variable of interest?

Supporting evidence: Researchers should provide as much evidence as possible to corroborate the identifying assumption. Such evidence may include showing treatment and control groups do not differ on observables prior to treatment, an analysis of pre- and posttreatment trends, the timing of the effect, and the use of multiple instruments. At a minimum, a careful description of the specific historical episode is needed that makes clear the origin of the quasi-random variation and allows the reader to gauge the potential scope for endogeneity and other confounding factors. Thus, detailed institutional knowledge on the episode surrounding the experiment is required. Researchers should carefully consider and explicitly discuss which omitted factors one might reasonably worry about.

Additional methods to support a causal interpretation: If possible, researchers should use additional analyses to establish that the treatment and not other factors correlated with treatment drives the outcome variable. Specifically, placebo exercises that analyze placebo treatments can provide credibility to the claim of causality. Moreover, showing robustness to the use of different control groups if more than one is possible, and using matching methods to guarantee the control group lines up well with the treatment group in terms of observable characteristics can help assuage concerns about randomness.

Quantitative implications: Arguing convincingly in favor of a causal interpretation is usually impossible without explicitly stating the quantitative implications of the estimates. Is the size of the effect reasonable? How does it compare to the influence of other factors that are known to affect the outcome variable?

A common promising avenue for future research identified in all three strands of literature is to move beyond reduced form approaches and incorporate natural experiments into structural models. Models will be useful to identify specific mechanisms at work and channels through which effects transmit, and to argue more forcefully about the plausibility of the quantitative sizes of effects. Natural experiments could then potentially prove useful in identifying a wider variety of structural model parameters.

Our discussion of the use of natural experiments in macroeconomics has also shown the "natural" limits of this approach. Because the researcher does not create the historical episode constituting the experiment, it is often not ideally suited to analyze the question at hand. As a result, different natural experiments often produce different answers to the same question. In addition, many of the more tightly identified natural experiments leave open the question of external validity. Finally, because natural experiments rely on history for identification, they are naturally more useful for understanding the past rather than the future. Although this is not problematic in many contexts, it does preclude natural experiments from addressing the effects of unprecedented events, such as climate change.

Nevertheless, we believe that two decades of research using natural experiments have produced a range of substantial insights into the functioning of the economy, and have made major contributions to the field of macroeconomics.

APPENDIX

Table A.1 Overview of Permanent Income Hypothesis papers relying on natural experiments

Study	Experiment	Data source	Sample	Main dependent variable	Frequency of data	PIH	Quantitative reaction at implementation	Liquidity constraint
Standard PIH studies								
Aaronson et al. (2012)	Minimum Wage Hikes; income increase	CEX interview survey (1982–2008); CPS (1980–2007) and SIPP (1983–2007) for data on income; proprietary dataset from national financial institution (1995–2008): credit card account and each credit card holder's auto, home equity, mortgage, and credit card balance	200,500 Household-survey observations on spending, of which 11% derive some income from minimum wage work	Spending on nondurables and durables, including separate subcategories, and change in debt (total and subcategories)	Quarterly	Reject	Households for which minimum wage labor is the source of at least 20% of household income spend 3.4 times the short-term increase in income, but mostly on cars	based on liquid assets; some evidence in favor

Continued

Table A.1 Overview of Permanent Income Hypothesis papers relying on natural experiments—cont'd

Study	Experiment	Data source	Sample	Main dependent variable	Frequency of data	PIH	Quantitative reaction at implementation	Liquidity constraint
Agarwal et al. (2007)	2001 Federal Income Tax Rebates; income increase	Proprietary data set from a large financial institution that issues credit cards nationally (2000–02): data on credit bureau reports, credit card spending, balance, debt	75,000 Credit card accounts open as of June 2000, followed monthly for 24 months	Credit card spending, balances, and debt, as well as credit limit	Monthly	Reject	Spending increases by 40% of the average household rebate cumulatively in the 9 months after receipt for consumers whose most intensively used credit card account is in the sample	Based on credit limit, utilization rate, and age; strong evidence in favor
Agarwal and Qian (2014b)	2011 Singapore Growth Dividends; income increase	Credit card, debit card, and bank checking account data from leading bank in Singapore (2010–12)	Random sample of all bank customers of leading Singaporean bank (180,000 observations)	Spending, debt, credit card usage	Monthly	Cannot reject	Increase in card spending by 8 cents per month for every dollar received, corresponding to a total increase of 80 cents in the 10 month period after announcement; monthly increase similar after announcement and implementation	based on liquid assets and credit card limit; liquidity constrained consumers react stronger, though also already at announcement, ie, likely incompletely binding constraints

Broda and Parker (2014)	2008 Economic Stimulus Payments; income increase	Nielsen's Consumer Panel (2008): scanned purchases in grocery stores, drugstores and mass-merchandise sectors	Data on weekly purchases of 28,937 households (1,131,208 observations)	weekly spending on household goods based on barcode scanners used by households	Weekly	Reject	3 Months cumulative increase in spending on Nielson Consumer Panel goods amounts to 7% of ESP receipt	based on income and survey question regarding the availability of easily accessible funds; evidence in favor
Browning and Collado (2001)	Spanish Bonus Payment Scheme; income increase and decrease	Spanish Encuesta Continua de Presupuestos Familiares (1985–95): data on earnings and spending	2341 Households (16,143 observations) of which about 80% are classified as bonus	Spending on nondurables and durables, including separate subcategories	Weekly	Cannot reject	Spending patterns of bonus and nonbonus groups are indistinguishable	No test
Coulibaly and Li (2006)	Last mortgage payment; disposable income increase	CEX interview survey (1984–2000)	70,593 Observations, including 286 with last mortgage payments	Spending on nondurables and durables, plus subcategories	Quarterly	Cannot reject	Homeowners do not increase consumption after the last mortgage payment	No test (none of respondents likely to be constrained)
Gelman et al. (2014)	Regular paycheck or social security check arrival; income increase and decrease	Check, a financial aggregation and service application combining information from different financial accounts (2012–13)	75,000 randomly sampled US Check users	Total spending, nonrecurring spending, and fast food and coffee shop spending	Daily	Cannot reject	Nonrecurring spending and fast food and coffee shop spending show only mild comovement with regular payments	Based on ratio of average daily balance on saving and checking account to average daily spending; strong evidence in favor

Continued

Table A.1 Overview of Permanent Income Hypothesis papers relying on natural experiments—cont'd

Study	Experiment	Data source	Sample	Main dependent variable	Frequency of data	PIH	Quantitative reaction at implementation	Liquidity constraint
Hsieh (2003)	Alaska Permanent Fund Payments and income tax refunds; income increase and decrease	CEX interview survey (1980–81, 1984–2001)	806 Alaskan households	Spending on nondurables and durables, and different subcategories	Quarterly	Cannot reject	A 10% increase in household income increases nondurable consumption insignificantly by only 0.002%	Based on current income; no evidence in favor
Johnson et al. (2006)	2001 Federal Income Tax Rebates; income increase	CEX interview survey (2000–02) with added questions on tax rebates	13,066 Observations on households who received the tax rebate	Spending on nondurables, strictly nondurables, and food	Quarterly	Reject	Households spend 20–40% of their rebates on nondurable consumption goods during the 3-months period in which the rebates were received, and roughly two thirds during 6-months period	Based on age, income, and liquid assets; some evidence based on income and liquid assets, not age
Mastrobuoni and Weinberg (2009)	Social Security benefits payments; income increase and decrease	Continuing Survey of Food Intake by Individuals (1994–96)	745 Observations from households in which Social Security income makes up at least 80% of total income	Caloric intake	Daily	Mixed	Retirees with savings above $5000 smooth caloric intake over the pay cycle, while those with less than $5000 in savings have 24% lower caloric intake during the final few days of pay cycle than during first week	Based on liquid assets; strong evidence in favor (see main result)

Misra and Surico (2014)	2001 Federal Income Tax Rebates and 2008 Economic Stimulus Payments; income increase	CEX interview survey (2000–02 and 2007–08) with added questions on tax rebate and stimulus payments	17,718 Household who received rebates in 2001 or 2008	Spending on nondurables and durables, plus subcategories	Quarterly	Reject	During 6-months period in which payment was received, households spent 43% of payments on nondurable consumption in 2001, and 16% on total consumption in 2008	based on high income and high mortgage debt (wealthy hand-to-mouth consumers); evidence in favor
Parker (1999)	Caps on Social Security tax and changes in Social Security tax withholding; income increase and decrease	CEX interview survey (1980–93)	133,820 Observations on 57,051 households	Spending on nondurables and durables, and different subcategories	Quarterly	Reject	When a household's Social Security contributions fall so that income rises by $1, nondurable consumption rises by 20 cents	Based on age and liquid assets; weak evidence in favor
Parker et al. (2013)	2008 Economic Stimulus Payments; income increase	CEX interview survey (2007–08) with added questions on stimulus payments	17,478 Household observations, of which 11,239 received economic stimulus payments	Spending on nondurables and durables, plus subcategories	Quarterly	Reject	During 3-months period in which payment was received, households increase their expenditures on nondurable goods by 12–39% of the payment, and on overall consumption by 50–90%	Based on age, income, and liquid assets; some evidence based on income and age, not on liquid assets

Continued

Table A.1 Overview of Permanent Income Hypothesis papers relying on natural experiments—cont'd

Study	Experiment	Data source	Sample	Main dependent variable	Frequency of data	PIH	Quantitative reaction at implementation	Liquidity constraint
Paxson (1993)	Seasonal income patterns in Thai agriculture; income increase and decrease	Thai Socioeconomic Surveys (1975–76, 1981, and 1986)	27,963 Economically active (ie, not retired) households that did not engage in forestry or fishing	Spending on nondurables	Monthly	Cannot reject	Spending patterns of farm and nonfarm households are indistinguishable	No test
Scholnick (2013)	Last mortgage payment; disposable income increase	proprietary data set from a Canadian bank (2004–06): credit card and mortgage accounts data	4147 Individuals who have paid off their mortgage or who have less than 1 year of mortgage payments left	Credit card expenditures	Monthly	Reject	No quantitative interpretation reported; reaction of consumption to preannounced income increase is weaker the larger the income increase	Based on households paying positive interest on credit card debt; also non liquidity constrained consumers react significantly
Shapiro (2005)	Food stamps; income increase and decrease	Continuing Survey of Food Intake by Individuals (1989–91); Nationwide Food Consumption Survey (1987–88): data on market value and nutritional characteristics of food eaten; survey conducted to document	6652 Observations from surveyed individuals who receive food stamps	Caloric intake	Daily	Reject	caloric intake declines by a statistically significant 0.40% per day after receipt of food stamps	No test

		effects of Maryland's adoption of the Electronic Benefit Transfer system (1992–93)						
Shea (1995)	Unionized wage; income increase and decrease	PSID survey (1981–86): data on food consumption matched with data on wage growth from union contracts	647 Observations drawn from 285 households whose head is a union member and can be reasonably assigned to specific union	Spending on food consumed at home and in restaurants, plus the bonus value of food stamps	Annual	Reject	A 1 percentage point increase in wage growth is associated with a 0.89 percentage point increase in food consumption	Based on liquid assets and heterogeneous reactions to increases vs decreases; mildly supportive of liquidity constraints in traditional measures; yet, reaction stronger to income decrease
Souleles (2002)	1981 Economic Recovery Tax Act (Reagan Tax Cuts); income increase	CEX interview survey (1982–83)	2399 Household-quarter observations, head aged 24 to 64	Spending on nondurables and total consumption	Quarterly	Reject	For each dollar increase in take-home pay, nondurable consumption rises by about two-thirds of a dollar	Based on age, income, and liquid assets; no evidence in favor
Souleles (2000)	Paying for college; income decrease	CEX interview survey (1980–93)	7200 Household observations with child aged 16–24, of which 1249 have positive college expenditure	Spending on nondurables	Quarterly	Cannot reject	A one dollar decrease in income due to paying college tuition leads to an 8 cent increase (not decrease) in nondurable consumption	NA

Continued

Table A.1 Overview of Permanent Income Hypothesis papers relying on natural experiments—cont'd

Study	Experiment	Data source	Sample	Main dependent variable	Frequency of data	PIH	Quantitative reaction at implementation	Liquidity constraint
Souleles (1999)	Income tax refunds; income increase	CEX interview survey (1980–91)	4121 Observations on households receiving income tax refunds, head aged 24 to 64	Spending on nondurables and total consumption	Quarterly	Reject	One dollar of refund receipt raises strictly nondurable consumption by 2.6 cents, and total consumption by 18 cents	Based on liquid assets; some evidence in favor
Stephens (2008)	Final payment of a vehicle loan; disposable income increase	CEX interview survey (1984–2000)	4583 Observations on households who have an expiring vehicle loan	Spending on nondurables, except for public transportation and gas and motor oil	Quarterly	Reject	A 10% increase in after-tax income increases nondurable consumption by 2.8%	Based on age, liquid wealth, and maturity of expiring vehicle loan; evidence in favor based on age and liquid assets, but not based on maturity of prior loan
Stephens (2006)	Regular paycheck receipt; income increase and decrease	UK Family Expenditure Survey: 2 week diary of all expenditures (1986–98)	12,827 Households with a dependently employed monthly paid primary earner aged 25–59	Total spending, spending on strict nondurables, food at home, and instant consumption goods	Weekly	Reject	Instant consumption increases by 5% in week when households receive monthly paychecks	Based on asset income; strong evidence in favor

Stephens (2003)	Receipt of Social Security check on "3rd of tha Month"; income increase and decrease	CEX diary survey (1986–96)	9942 consumer units which contribute a total of 123,034 potential expenditure days	Spending on instant consumption, food, and nondurables	Daily	Reject	Households that receive at least 70 percent of income from Social Security increase instant consumption and food away from home by roughly 20% in week following arrival of Social Security check	No test
Stephens and Unayama (2011)	Change in frequency of Japanese public pension payments (February 1990); income increase and decrease	Japanese Family Income and Expenditure Survey (1986–94): diary data on expenditures and income	2503 Retirees and employees before reform (pension paid once every 3 months), and 3595 after reform pension paid once every 2 months)	spending on nondurables and durables	Monthly	Reject	Nondurable consumption increases by 4% in the month of check receipt, while strict nondurable and food consumption both increase by over 2% when checks are received	Based on age, total net financial assets and demand deposits; no evidence in favor
Wilcox (1989)	Changes in Social Security amounts; income increase	aggregate data on retail sales and personal consumption expenditures (1965–85)		Total retail sales, also divided in durable and nondurable good stores, and all commodities	Monthly	Reject	An increase in benefits by 10% increases total retail sales by 1.4%, with a 3% increase in durable goods sales (mostly cars)	No test

Continued

Table A.1 Overview of Permanent Income Hypothesis papers relying on natural experiments—cont'd

Study	Experiment	Data source	Sample	Main dependent variable	Frequency of data	PIH	Quantitative reaction at implementation	Liquidity constraint
Liquidity constraint relaxation								
Abdallah and Lastrapes (2012)	1998 Texas constitutional amendment on home equity loans allows mortgages for nonhousing expenditures	county (1992, 1997, 2002) and state (1992–2002) level data on retail sales; American Housing Survey (1994 and 2002): household-level data on second-lien equity homes	3006 Counties/ 45 states with complete data	Retail sales (measured directly on county level, or as ratio of sales tax revenues to sales tax rates at state level)	Annual (county level: 1992, 1997, 2002 only)	NA	Real per capita spending of average Texas county increases by 2–4% due to relaxation of borrowing constraints	NA
Bertrand and Morse (2009)	2008 Economic Stimulus Payments; income increase	Customers of a payday lending chain (March to September 2008)	881 Active payday loan customers	Payday loan take-up	Weekly	NA	Payday loan customers reduce borrowing on average by $46 after receipt of the rebate check of on average $600; frequency of borrowing also falls significantly	Based on frequency of use of payday loans; no evidence in favor
Gross and Tobacman (2014)	2001 Federal Income Tax Rebates and 2008 Economic Stimulus Payment; income increase	Data set compiled based on public access to Court Electronic Records system	All consumer bankruptcy filings in 81 out of 94 US courts (1998–2008)	Chapter 7 and chapter 13 bankruptcy filings	Weekly	NA	Bankruptcies increase by 2% after 2001 rebates and 6% after 2008 rebates; also shift from chapter 13 to chapter 7 bankruptcies	Based on income, share of subprime borrowers, and home ownership rate at ZIP code level; no evidence in favor

Leth-Petersen (2010)	1992 Danish credit market reform allows mortgages for nonhousing expenditures	Danish public administrative registers: data on wealth, income, household composition, characteristics of dwelling (1987–96)	63,613 Households aged between 25 and 65 in 1991	Imputed expenditures (income minus change in wealth)	Annual	NA	Credit constrained households with an equity to house value ratio of 0.5 or higher increase their expenditure by 1–3% in the years following change in law	Based on liquid assets; some evidence in favor if equity to home value is high
Temporary price cuts								
Agarwal et al. (2013)	Sales Tax Holidays	CEX diary survey (1997–2011); proprietary credit card transactions data from a large financial institution (2003, February 8–October 20)	Over 700,000 household-date observations from CEX diaries; over 10 million consumer-date observations from credit card data	Spending on specific categories (esp. children's clothing), credit card transactions	Daily	Reject	Sales tax holidays increase daily clothing spending by $1.17 (ie, 29% of daily household clothing spending); no significant reduction in spending before/after sales tax holiday on these categories, or during sales tax holidays on other categories	No test

Continued

Table A.1 Overview of Permanent Income Hypothesis papers relying on natural experiments—cont'd

Study	Experiment	Data source	Sample	Main dependent variable	Frequency of data	PIH	Quantitative reaction at implementation	Liquidity constraint
Mian and Sufi (2012)	2009 Cars Allowance Rebate System (Cash for Clunkers)	Data on car purchases from R.L. Polk for US metropolitan or micropolitan statistical areas (2004–10), augmented by data from Census, Equifax Predictive Services, Federal Housing Finance Agency, BLS and IRS	957 Metropolitan or micropolitan statistical areas	New car purchases	Monthly	Cannot reject	Cities with lots of qualifying clunkers have significantly higher increase in car sales in months of program, but then reduce them, such that cumulative response over 12 months is zero	No test

ACKNOWLEDGMENTS

The chapter has benefitted from helpful discussions and comments from Daron Acemoglu, Jonathan Dingel, Yuriy Gorodnichenko, Chang-Tai Hsieh, Nathan Nunn, Elias Papaioannou, Matthias Schündeln, Rob Vishny, Mirko Wiederholt, and seminar participants of the Handbook of Macroeconomics Conference at Stanford and at the Deutsche Bundesbank. Leonhard Czerny, Denis Gorea, and Philip Xu provided excellent research assistance. N.F.-S. gratefully acknowledges financial support from the Cluster of Excellence "Formation of Normative Orders" and the European Research Council under Starting Grant No. 262116.

REFERENCES

Aaronson, D., Agarwal, S., French, E., 2012. The spending and debt response to minimum wage hikes. Am. Econ. Rev. 102 (7), 3111–3139.

Abadie, A., Gardeazabal, J., 2003. The economic costs of conflict: a case study of the Basque country. Am. Econ. Rev. 93 (1), 113–132.

Abadie, A., Diamond, A., Hainmueller, J., 2010. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. J. Am. Stat. Assoc. 105 (490), 493–505.

Abdallah, C.S., Lastrapes, W.D., 2012. Home equity lending and retail spending: evidence from a natural experiment in Texas. Am. Econ. J. Macroecon. 4 (4), 94–125.

Acconcia, A., Corsetti, G., Simonelli, S., 2014. Mafia and public spending: evidence on the fiscal multiplier from a Quasi-experiment. Am. Econ. Rev. 104 (7), 2185–2209.

Acemoglu, D., 2009. Introduction to Modern Economic Growth. Princeton University Press, Princeton, NJ.

Acemoglu, D., Robinson, J.A., 2005. Economic Origins of Dictatorship and Democracy. Cambridge University Press, Cambridge, United Kingdom.

Acemoglu, D., Hassan, T.A., Tahoun, A., 2015. The Power of the Street: Evidence from Egypt's Arab Spring. Fama-Miller Working Paper.

Acemoglu, D., Johnson, S., Robinson, J.A., 2001. The colonial origins of comparative development: an empirical investigation. Am. Econ. Rev. 91, 1369–1401.

Acemoglu, D., Johnson, S., Robinson, J.A., 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. Q. J. Econ. 117, 1231–1294.

Acemoglu, D., Johnson, S., Robinson, J., Thaicharoen, Y., 2003. Institutional causes, macroeconomic symptoms: volatility, crises and growth. J. Monet. Econ. 50 (1), 49–123.

Acemoglu, D., Johnson, S., Robinson, J.A., 2005a. The rise of Europe: Atlantic trade, institutional change and economic growth. Am. Econ. Rev. 95, 546–579.

Acemoglu, D., Johnson, S., Robinson, J.A., Yared, P., 2005b. From education to democracy. Am. Econ. Rev. Pap. Proc. 95 (2), 44–49.

Acemoglu, D., Cantoni, D., Johnson, S., Robinson, J.A., 2011a. The consequences of radical reform: the French revolution. Am. Econ. Rev. 101 (7), 3286–3307.

Acemoglu, D., Hassan, T.A., Robinson, J.A., 2011b. Social structure and development: a legacy of the holocaust in Russia. Q. J. Econ. 126 (2), 895–946.

Acemoglu, D., Johnson, S., Robinson, J.A., 2012. The colonial origins of comparative development: an empirical investigation: reply. Am. Econ. Rev. 102 (6), 3077–3110.

Acemoglu, D., Gallego, F.A., Robinson, J.A., 2014a. Institutions, human capital, and development. Ann. Rev. Econ. 6, 875–912.

Acemoglu, D., Naidu, S., Restrepo, P., Robinson, J.A., 2014b. Democracy does cause growth. NBER Working Paper No. 20004.

Agarwal, S., Qian, W., 2014a. Access to home equity and consumption: evidence from a policy experiment. Working Paper.

Agarwal, S., Qian, W., 2014b. Consumption and debt response to unanticipated income shocks: evidence from a natural experiment in Singapore. Am. Econ. Rev. 104 (12), 4205–4230.

Agarwal, S., Liu, C., Souleles, N.S., 2007. The reaction of consumer spending and debt to tax rebates-evidence from consumer credit data. J. Polit. Econ. 115 (6), 986–1019.

Agarwal, S., Marwell, N., McGranahan, L., 2013. Consumption responses to temporary tax incentives: evidence from state sales holidays. Working Paper.

Aghion, P., Algan, Y., Cahuc, P., Shleifer, A., 2010. Regulation and distrust. Q. J. Econ. 125 (3), 1015–1049.

Ahlfeldt, G.M., Redding, S.J., Sturm, D.M., Wolf, N., 2015. The economics of density: evidence from the Berlin wall. Econometrica 83 (6), 2127–2189.

Akerlof, G., Yellen, J., 1985. A near-rational model of the business cycle with wage and price inertia. Q. J. Econ. 823–838.

Albouy, D.Y., 2012. The colonial origins of comparative development: an empirical investigation: comment. Am. Econ. Rev. 102 (6), 3059–3076.

Alesina, A., Fuchs-Schündeln, N., 2007. Good bye Lenin (or not?): the effect of communism on people. Am. Econ. Rev. 97 (4), 1507–1528.

Alesina, A., Giuliano, P., 2015. Culture and institutions. J. Econ. Lit. 53 (4), 898–944.

Algan, Y., Cahuc, P., 2010. Inherited trust and growth. Am. Econ. Rev. 100 (5), 2060–2092.

Algan, Y., Cahuc, P., 2013. Trust and growth. Ann. Rev. Econ. 5 (1), 521–549. http://dx.doi.org/10.1146/annurev-economics-081412-102108.

Altonji, J.G., Elder, T.E., Taber, C.R., 2005. Selection on observed and unobserved variables: assessing the effectiveness of catholic schools. J. Polit. Econ. 113 (1), 151–184.

Anderson, C.J., 2003. The psychology of doing nothing: forms of decision avoidance result from reason and emotion. Psychol. Bull. 129, 139–167.

Arrow, K.J., 1972. Gifts and exchanges. Phil. Publ. Aff. 1 (4), 343–362. http://www.jstor.org/stable/2265097.

Auerbach, A.J., Gorodnichenko, Y., 2012. Measuring the output responses to fiscal policy. Am. Econ. J. Econ. Pol. 4 (2), 1–27.

Baker, S.R., Bloom, N., 2013. Does uncertainty reduce growth? Using disasters as natural experiments. NBER Working Paper No. 19475.

Banerjee, A., Chandrasekhar, A.G., Duflo, E., Jackson, M.O., 2013. The diffusion of microfinance. Science 341 (6144), 363.

Banerjee, A., Iyer, L., 2005. History, institutions and economic performance: the legacy of colonial land tenure systems in India. Am. Econ. Rev. 95 (4), 1190–1213.

Banfield, E.C., 1967. The Moral Basis of a Backward Society. Free Press, Glencoe, IL.

Barro, R.J., 1981. Output effects of government purchases. J. Polit. Econ. 89 (6), 1086–1121.

Barro, R.J., 1999. Determinants of democracy. J. Polit. Econ. 107 (S6), S158–S183.

Barro, R.J., Redlick, C.J., 2011. Macroeconomic effects from government purchases and taxes. Q. J. Econ. 126 (1), 51–102.

Beaman, L.A., 2012. Social networks and the dynamics of labor market outcomes: evidence from refugees resettled in the U.S. Rev. Econ. Stud. 79 (1), 128–161.

Becker, S.O., Egger, P.H., von Ehrlich, M., 2010. Going NUTS: the effect of EU structural funds on regional performance. J. Publ. Econ. 94 (9-10), 578–590.

Becker, S.O., Egger, P.H., von Ehrlich, M., 2013. Absorptive capacity and the growth and investment effects of regional transfers: a regression discontinuity design with heterogeneous treatment effects. Am. Econ. J. Econ. Pol. 5 (4), 29–77.

Becker, S.O., Boeckh, K., Hainz, C., Woessmann, L., 2015. The empire is dead, long live the empire! Long-run persistence of trust and corruption in the bureaucracy. Econ. J. 126 (590), 40–74.

Bellows, J., Miguel, E., 2009. War and local collective action in Sierra Leone. J. Publ. Econ. 93 (11-12), 1144–1157. ISSN 0047-2727 http://dx.doi.org/10.1016/j.jpubeco.2009.07.012. http://www.sciencedirect.com/science/article/pii/S0047272709000942.

Bertrand, M., Morse, A., 2009. What do high-interest borrowers do with their tax rebate? Am. Econ. Rev. 99 (2), 418–423.

Bertrand, M., Luttmer, E.F.P., Mullainathan, S., 2000. Network effects and welfare cultures. Q. J. Econ. 115 (3), 1019–1055.

Blanchard, O., Perotti, R., 2002. An empirical characterization of the dynamic effects of changes in government spending and taxes on output. Q. J. Econ. 117 (4), 1329–1368.

Bleakely, H., Lin, J., 2012. Portage and path dependence. Q. J. Econ. 127, 587–644.

Bloom, N., 2009. The impact of uncertainty shocks. Econometrica 77 (3), 623–685.

Bodkin, R., 1959. Windfall income and consumption. Am. Econ. Rev. 49 (4), 602–614.

Broda, C., Parker, J.A., 2014. The economic stimulus payments of 2008 and the aggregate demand for consumption. J. Monet. Econ. 68 (S), 20–36.

Browning, M., Collado, M.D., 2001. The response of expenditures to anticipated income changes: panel data estimates. Am. Econ. Rev. 91 (3), 681–692.

Brückner, M., Tuladhar, A., 2014. Local government spending multipliers and financial distress: evidence from Japanese prefectures. Econ. J. 124 (581), 1279–1316.

Brückner, M., Ciccone, A., 2011. Rain and the democratic window of opportunity. Econometrica 79 (3), 923–947.

Brückner, M., Gradstein, M., 2013. Effects of transitory shocks to aggregate output on consumption in poor countries. J. Int. Econ. 91, 343–357.

Brückner, M., Ciccone, A., Tesei, A., 2012. Oil price shocks, income, and democracy. Rev. Econ. Stat. 94 (2), 389–393.

Burchardi, K.B., Hassan, T.A., 2013. The economic impact of social ties: evidence from German reunification. Q. J. Econ. 128 (3), 1219–1271.

Burchardi, K.B., Chaney, T., Hassan, T.A., 2015. Migrants, trade, and investments. Working Paper.

Burke, P.J., Leigh, A., 2010. Do output contractions trigger democratic change? Am. Econ. J. Macroecon. 2 (4), 124–157. http://dx.doi.org/10.1257/mac.2.4.124.

Bursztyn, L., Cantoni, D., 2016. A tear in the iron curtain: the impact of western television on consumption behavior. Rev. Econ. Stat. 98 (1), 25–41.

Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, MA.

Carroll, C.D., Summers, L.H., 1991. Consumption growth parallels income growth: some new evidence. In: Bernheim, B.D., Shoven, J.B. (Eds.), National Saving and Economic Performance. University of Chicago Press, Chicago, IL, pp. 305–348.

Caselli, F., Tesei, A., 2011. Resource windfalls, political regimes, and political stability. Rev. Econ. Stat. http://dx.doi.org/10.3386/w17601.

Chaney, E., 2013. Revolt on the Nile: economic shocks, religion and political power. Econometrica 81 (5), 2033–2053.

Chaney, T., 2014. Networks in international trade. In: Bramoulle, Y., Galleoti, A., Rogers, B. (Eds.), Oxford Handbook of the Economics of Networks. Oxford University Press, Oxford, United Kingdom, pp. 754–775.

Chetty, R., 2012. Bounds on elasticities with optimization frictions: a synthesis of micro and macro evidence on labor supply. Econometrica 80 (3), 969–1018.

Chodorow-Reich, G., Feiveson, L., Liscow, Z., Woolston, W.G., 2012. Does state fiscal relief during recessions increase employment? Evidence from the American recovery and reinvestment act. Am. Econ. J. Econ. Pol. 4 (3), 118–145.

Clemens, J., Miran, S., 2012. Fiscal policy multipliers on subnational government spending. Am. Econ. J. Econ. Pol. 4 (2), 46–68.

Cochrane, J., 1989. The sensitivity of tests of the intertemporal allocation of consumption to near-rational alternatives. Am. Econ. Rev. 79 (3), 319–337.

Cohen, L., Frazzini, A., Malloy, C., 2008. The small world of investing: board connections and mutual fund returns. J. Polit. Econ. 116 (5), 951–979.

Cohen, L., Coval, J., Malloy, C., 2011. Do powerful politicians cause corporate downsizing? J. Polit. Econ. 119 (6), 1015–1060.

Cohen, L., Gurun, U., Malloy, C., 2014. Resident networks and firm trade. Working Paper, October 2014.

Coleman, J.S., 1988. Social capital in the creation of human capital. Am. J. Sociol. 94, 95–120.

Combes, P.P., Lafourcade, M., Mayer, T., 2005. The trade-creating effects of business and social networks: evidence from France. J. Int. Econ. 66, 1–29.

Conley, T.G., Udry, C.R., 2010. Learning about a new technology: pineapple in Ghana. Am. Econ. Rev. 100 (1), 35–69.

Corbi, R., Papaioannou, E., Surico, P., 2014. Federal transfer multipliers. Quasi-experimental evidence from Brazil. NBER Working Paper No. 20751.

Coulibaly, B., Li, G., 2006. Do homeowners increase consumption after the last mortgage payment? An alternative test of the permanent income hypothesis. Rev. Econ. Stat. 88 (1), 10–19.

Davis, D.R., Weinstein, D.E., 2002. Bones, bombs, and breakpoints: the geography of economic activity. Am. Econ. Rev. 92 (5), 1269–1289.

Davis, D.R., Weinstein, D.E., 2008. A search for multiple equilibria in urban industrial structure. J. Reg. Sci. 48 (1), 29–65.

DeFusco, A.A., 2014. Homeowner Borrowing and Housing Collateral: New Evidence from Expiring Price Controls. Mimeo, University of Pennsylvania.

Dell, M., 2010. The persistent effects of Peru's mining mita. Econometrica 78 (6), 1863–1903.

Dippel, C., 2014. Forced coexistence and economic development: evidence from native American reservations. Econometrica 82 (6), 2131–2165.

Duffy, J., 2008. Macroeconomics: a survey of laboratory research. In: Kagel, J., Roth, A.E. (Eds.), Handbook of Experimental Economics, vol. 2. Princeton University Press, Princeton, NJ.

Durante, R., 2010. Risk, Cooperation and the Economic Origins of Social Trust: An Empirical Investigation. Mimeo, Science Po.

Easterly, W., Levine, R., 2003. Tropics, germs and crops: how endowments influence economic development. J. Monet. Econ. 50, 3–39.

Easterly, W., Levine, R., 2012. The European origins of economic development. NBER Working Paper No. 18162.

Edelberg, W., Eichenbaum, M., Fisher, J.D.M., 1999. Understanding the effects of a shock to government purchases. Rev. Econ. Dyn. 2 (1), 166–206.

Evans, W.N., Moore, T.J., 2011. The short-term mortality consequences of income receipt. J. Publ. Econ. 95 (11), 1410–1424.

Feyrer, J.D., Sacerdote, B., 2012. Did the Stimulus Stimulate? Effects of the American Recovery and Reinvestment Act. Mimeo, Dartmouth University.

Fishback, P., Cullen, J.A., 2013. Second world war spending and local economic activity in US counties, 1939-58. Econ. Hist. Rev. 66 (4), 975–992. http://EconPapers.repec.org/RePEc:bla:ehsrev:v:66:y:2013:i:4:p:975-992.

Fishback, P.V., Kachanovskaya, V., 2015. In Search of the multiplier for federal spending in the states during the great depression. J. Econ. Hist. 75 (1), 125–162.

Fisman, R., 2001. Estimating the value of political connections. Am. Econ. Rev. 91 (4), 1095–1102.

Franck, R., 2015. The political consequences of income shocks: eplaining the consolidation of democracy in France. Rev. Econ. Stat. 98 (1), 57–82.

Friedman, M., 1957. A Theory of the Consumption Function. Princeton University Press, Princeton, NJ.

Fuchs-Schündeln, N., 2008. The response of household saving to the large shock of German reunification. Am. Econ. Rev. 98 (5), 1798–1828.

Fuchs-Schündeln, N., Schündeln, M., 2005. Precautionary savings and self-selection: evidence from the German reunification "experiment" Q. J. Econ. 120, 1085–1120.

Fuchs-Schündeln, N., Schündeln, M., 2015. On the endogeneity of political preferences: evidence from individual experience with democracy. Science 347 (6226), 1145–1148.

Garmendia, A., Llano, C., Minondo, A., Requena, F., 2012. Networks and the disappearance of the intranational home bias. Econ. Lett. 116, 178–182.

Gebhardt, G., 2013. Does relationship specific investment depend on asset ownership? Evidence from a natural experiment in the housing market. J. Eur. Econ. Assoc. 11 (1), 201–227.

Gelman, M., Kariv, S., Shapiro, M.D., Silverman, D., Tadelis, S., 2014. Harnessing naturally occurring data to measure the response of spending to income. Science 345 (6193), 212–215.

Glaeser, E.L., Laibson, D., Sacerdote, B., 2002. An economic approach to social capital. Econ. J. 112 (483), F437–F458.

Glaeser, E.L., La Porta, R., Lopez-De-Silanes, F., Shleifer, A., 2004. Do institutions cause growth? J. Econ. Growth 9, 271–303.

Glaeser, E.L., Ponzetto, G., Shleifer, A., 2007. Why does democracy need education? J. Econ. Growth 12 (2), 77–99.

Gorodnichenko, Y., Roland, G., 2010. Culture, institutions and the wealth of nation. NBER Working Paper No. 16368.

Gould, D.M., 1994. Immigrant links to the home country: empirical implications for U.S. bilateral trade flows. Rev. Econ. Stat. 76 (2), 302–316.

Granovetter, M., 1985. Economic action and social structure: the problem of embeddedness. Am. J. Sociol. 91 (3), 481–510.

Granovetter, M., 2005. The impact of social structure on economic outcomes. J. Econ. Perspect. 19 (1), 33–50.

Greif, A., 1993. Contract enforceability and economic institutions in early trade: the Maghribi traders' coalition. Am. Econ. Rev. 83 (3), 525–548. ISSN 00028282. http://www.jstor.org/stable/2117532.

Gross, T., Tobacman, J., 2014. Dangerous liquidity and the demand for health care: evidence from the 2008 stimulus payments. J. Hum. Resour. 49 (2), 424–445.

Gross, T., Notowidigdo, M.J., Wang, J., 2014. Liquidity constraints and consumer bankruptcy: evidence from tax rebates. Rev. Econ. Stat. 96 (3), 431–443.

Guiso, L., Sapienza, P., Zingales, L., 2004. The role of social capital in financial development. Am. Econ. Rev. 94 (3), 526–556.

Guiso, L., Sapienza, P., Zingales, L., 2006. Does culture affect economic outcomes? J. Econ. Perspect. 20 (2), 23–48. http://dx.doi.org/10.1257/jep.20.2.23.

Guiso, L., Sapienza, P., Zingales, L., 2008a. Long-term persistence. NBER Working Paper No. 14278.

Guiso, L., Sapienza, P., Zingales, L., 2008b. Social capital as good culture. J. Eur. Econ. Assoc. 6 (2-3), 295–320.

Guiso, L., Sapienza, P., Zingales, L., 2012. Civic capital as the missing link. In: Benhabib, J., Bisin, A., Jackson, M.O. (Eds.), The Handbook of Social Economics. Elsevier, Amsterdam, Netherlands, pp. 417–480.

Hall, R.E., 1986. The role of consumption in economic fluctuations. In: Gordon, R.J. (Ed.), The American Business Cycle: Continuity and Change. University of Chicago Press, Chicago, IL, pp. 237–266.

Hall, R.E., 2009. By how much does GDP rise if the government buys more output? Brook. Pap. Econ. Act. 40 (2), 183–249.

Hall, R.E., Jones, C.I., 1999. Why do some countries produce so much more output per worker than others? Q. J. Econ. 114 (1), 83–116.

Hassan, T., Mertens, T., 2014. The social cost of near-rational investment. NBER Working Paper No. 17027.

Hausman, J.K., 2015. Fiscal policy and economic recovery: the case of the 1936 veterans' bonus. Am. Econ. Rev. 106 (4), 1100–1143.

Hochberg, Y., Ljungqvist, A., Lu, Y., 2007. Whom you know matters: venture capital networks and investment performance. J. Financ. 62, 251–301.

Hornbeck, R., Naidu, S., 2014. When the levee breaks: black migration and economic development in the American south. Am. Econ. Rev. 104 (3), 963–990. http://dx.doi.org/10.1257/aer.104.3.963.

Hsieh, C.T., 2003. Do consumers react to anticipated income changes? Evidence from the Alaska permanent fund. Am. Econ. Rev. 93 (1), 397–405.

Huntington, S.P., 1991. Democracy's third wave. J. Democr. 2 (2), 12–34.

Imbens, G.W., Rubin, D.B., Sacerdote, B.I., 2001. Estimating the effect of unearned income on labor earnings, savings, and consumption: evidence from a survey of lottery players. Am. Econ. Rev. 91 (4), 778–794.

Iyer, L., 2010. Direct versus indirect colonial rule in India: long-term consequences. Rev. Econ. Stat. 92 (4), 693–713.

Jancec, M., 2012. Do Less Stable Borders Lead to Lower Levels of Political Trust? Empirical Evidence from Eastern Europe. Mimeo, University of Maryland at College Park.

Jappelli, T., Pistaferri, L., 2010. The consumption response to income changes. Ann. Rev. Econ. 2, 479–506.

Johnson, D.S., Parker, J.A., Souleles, N.S., 2006. Household expenditure and the income tax rebates of 2001. Am. Econ. Rev. 96 (5), 1589–1610.

Jones, E., 2003. The European Miracle: Environments, Economies and Geopolitics in the History of Europe and Asia. Cambridge University Press, Cambridge, United Kingdom.

Juhasz, R., 2014. Temporary protection and technology adoption: evidence from the napoleonic blockade. Job Market Paper.

Kaplan, G., Violante, G.I., 2014a. A model of the consumption response to fiscal stimulus payments. Econometrica 82 (4), 1199–1239.

Kaplan, G., Violante, G.I., 2014b. A tale of two stimulus payments: 2001 vs. 2008. Am. Econ. Rev. 104 (5), 116–121.

Kline, P., Moretti, E., 2014. Local economic development, agglomeration economies, and the big push: 100 years of evidence from the Tennessee valley authority. Q. J. Econ. 129 (1), 275–331.

Knack, S., Keefer, P., 1997. Does social capital have an economic payoff? A cross-country investigation. Q. J. Econ. 112 (4), 1251–1288.

Kraay, A., 2012. How large is the government spending multiplier? Evidence from world bank lending. Q. J. Econ. 127 (2), 829–887.

Kraay, A., 2014. Government spending multipliers in developing countries: evidence from lending by official creditors. Am. Econ. J. Macroecon. 6 (4), 170–208.

Kreinin, M.E., 1961. Windfall income and consumption: additional evidence. Am. Econ. Rev. 51 (3), 388–390.

Krugman, P., 1991. Increasing returns and economic geography. J. Polit. Econ. 99 (3), 483–499.

Kueng, L., 2015. Explaining consumption excess sensitivity with near-rationality: evidence from large predetermined payments. NBER Working Paper No. 21772.

Kuhn, P., Kooreman, P., Soetevent, A., Kapteyn, A., 2011. The effects of lottery prizes on winners and their neighbors: evidence from the Dutch postcode lottery. Am. Econ. Rev. 101 (5), 2226–2247.

Kuhnen, C.M., 2009. Business networks, corporate governance, and contracting in the mutual fund industry. J. Financ. 64 (5), 2185–2220.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R., 1997. Legal determinants of external finance. J. Financ 52, 1131–1150.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R., 1998. Law and finance. J. Polit. Econ. 106, 1113–1155.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., 2008. The economic consequences of legal origins. J. Econ. Lit. 46 (2), 285–332.

Landsberger, M., 1966. Windfall income and consumption: comment. Am. Econ. Rev. 56 (3), 534–540.

Leth-Petersen, S., 2010. Intertemporal consumption and credit constraints: does total expenditure respond to an exogenous shock to credit? Am. Econ. Rev. 100 (3), 1080–1103.

Lipset, S.M., 1959. Some social requisites of democracy: economic development and political legitimacy. Am. Polit. Sci. Rev. 53 (01), 69–105.

Loury, G.C., 1977. Women, minorities and employment discrimination. In: Wallace, P.A., LaMond, A. (Eds.), A Dynamic Theory of Racial Income Differences. Lexington Books, Lanham, MD, pp. 153–188.

Ludvigson, S.C., Michaelides, A., 2001. Does buffer-stock saving explain the smoothness and excess sensitivity of consumption? Am. Econ. Rev. 91 (3), 631–647.

Madestam, A., Shoag, D., Veuger, S., Yanagizawa-Drott, D., 2013. Do political protests matter? Evidence from the tea party movement. Q. J. Econ. 128 (4), 1633–1685.

Mankiw, N.G., 1985. Small menu costs and large business cycles: a macroeconomic model of monopoly. Q. J. Econ. 100 (2), 529–537. ISSN 00335533. http://www.jstor.org/stable/1885395.

Manski, C.F., 1993. Identification of endogenous social effects: the reflection problem. Rev. Econ. Stud. 60, 531–542.

Mastrobuoni, G., Weinberg, M., 2009. Heterogeneity in intra-monthly consumption patterns, self-control, and savings at retirement. Am. Econ. J. Econ. Pol. 1 (2), 163–189.

Mauro, P., 1995. Corruption and growth. Q. J. Econ. 110 (3), 681–712.

Mian, A., Sufi, A., 2012. The effects of fiscal stimulus: evidence from the 2009 cash for Clunkers Program. Q. J. Econ. 127 (3), 1107–1142.

Michalopoulos, S., Papaioannou, E., 2011. The long-run effects of the scramble for Africa. NBER Working Paper No. 17620.

Michalopoulos, S., Papaioannou, E., 2014. National institutions and subnational development in Africa. Q. J. Econ. 129 (1), 151–213.

Miguel, E., Roland, G., 2011. The long run impact of bombing Vietnam. J. Dev. Econ. 96 (1), 1–15.

Miguel, E., Satyanath, S., Sergenti, E., 2004. Economic shocks and civil conflict: an instrumental variables approach. J. Polit. Econ. 112 (4), 725–753. ISSN 00223808. http://www.jstor.org/stable/10.1086/421174.

Misra, K., Surico, P., 2014. Consumption, income changes, and heterogeneity: evidence from two fiscal stimulus programs. Am. Econ. J. Macroecon. 6 (4), 84–106.

Mountford, A., Uhlig, H., 2009. What are the effects of fiscal policy shocks? J. Appl. Econ. 24 (6), 960–992.

Munshi, K., 2003. Networks in the modern economy: Mexican migrants in the U. S. labor market. Q. J. Econ. 118 (2), 549–599. ISSN 00335533. http://www.jstor.org/stable/25053914.

Murphy, K.M., Shleifer, A., Vishny, R.W., 1989. Industrialization and the big push. J. Polit. Econ. 97, 1003–1026. Reprinted in Dilip Mookherjee and Debraj Ray eds., Readings in Theory of Economic Development, Blackwell Publishing, 2001.

Nakamura, E., Steinsson, J., 2014. Fiscal stimulus in a monetary union: evidence from US regions. Am. Econ. Rev. 104 (3), 753–792.

North, D.C., 1981. Structure and Change in Economic History. W.W. Norton & Company, New York, NY.

North, D.C., Thomas, R.P., 1973. The Rise of the Western World: A New Economic History. Cambridge University Press, Cambridge, United Kingdom.

Nunn, N., 2013. Historical development. Handb. Econ. Growth 2, 347.

Nunn, N., Wantchekon, L., 2011. The slave trade and the origins of mistrust in Africa. Am. Econ. Rev. 101 (7), 3221–3252.

Olsson, O., 2004. Unbundling Ex-Colonies: A Comment on Acemoglu, Johnson, and Robinson, 2001. Mimeo, Goteborg University.

Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge, United Kingdom.

Owyang, M.T., Ramey, V.A., Zubairy, S., 2013. Are government spending multipliers greater during periods of Slack? Evidence from twentieth-century historical data. Am. Econ. Rev. 103 (3), 129–134.

Parker, J.A., 1999. The reaction of household consumption to predictable changes in social security taxes. Am. Econ. Rev. 89 (4), 959–973.

Parker, J.A., 2014. Why don't households smooth consumption? Evidence from a 25 million dollar experiment. Mimeo.

Parker, J.A., Souleles, N.S., Johnson, D.S., McClelland, R., 2013. Consumer spending and the economic stimulus payments of 2008. Am. Econ. Rev. 103 (6), 2530–2553.

Parsons, C., Vezina, P.L., 2014. Migrant networks and trade: the vietnamese boat people as a natural experiment. University of Oxford.

Paxson, C.H., 1993. Consumption and income seasonality in Thailand. J. Polit. Econ. 101 (1), 39–72.

Persson, T., Tabellini, G., 2009. Democratic capital: the nexus of political and economic change. Am. Econ. J. Macroecon. 1 (2), 88–126. http://dx.doi.org/10.1257/mac.1.2.88.

Pinkovskiy, M.L., 2013. Economic Discontinuities at Borders: Evidence from Satellite Data on Lights at Night. Mimeo, Feseral Reserve Bank of New York.

Portes, R., Rey, H., 2005. The determinants of cross-border equity flows. J. Int. Econ. 65 (2), 269–296.

Putnam, R.D., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon and Schuster, New York, NY.

Putnam, R., Leonardi, R., Nanetti, R., 1993. Making Democracy Work. Simon & Schuster, New York, NY.

Ramey, V.A., 2011. Identifying government spending shocks: it's all in the Timing. Q. J. Econ. 126 (1), 1–50.

Ramey, V.A., Shapiro, M.D., 1998. Costly capital reallocation and the effects of government spending. Carn.-Roch. Conf. Ser. Public Policy 48 (1), 145–194.

Rauch, J.E., Trindade, V., 2002. Ethnic Chinese networks in international trade. Rev. Econ. Stat. 84 (1), 116–130.

Redding, S.J., Sturm, D.M., 2008. The costs of remoteness: evidence from German division and reunification. Am. Econ. Rev. 98 (5), 1766–1797.

Redding, S.J., Sturm, D., Wolf, N., 2011. History and industrial location: evidence from German airports. Rev. Econ. Stat. 93 (3), 814–831.

Reis, R., 2006. Inattentive consumers. J. Monet. Econ. 53 (8), 1761–1800.

Rice, T.W., Feldman, J.L., 1997. Civic Culture and democracy from Europe to America. J. Polit. 59 (4), 1143–1172. http://www.jstor.org/stable/2998596.

Roberts, B.E., 1990. A dead senator tells no lies: seniority and the distribution of federal benefits. Am. J. Polit. Sci. 34 (1), 31–58.

Rodrik, D., Subramanian, A., Trebbi, F., 2004. Institutions rule: the primacy of institutions over geography and integration in economic development. J. Econ. Growth 9 (2), 131–165.

Romer, C.D., Romer, D.H., 2010. The macroeconomic effects of tax changes: estimates based on a new measure of fiscal shocks. Am. Econ. Rev. 100 (3), 763–801.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.

Sacerdote, B., 2001. Peer effects with random assignment: results for Dartmouth roommates. Q. J. Econ. 116 (2), 681–704.

Sahm, C.R., Shapiro, M.D., Slemrod, J., 2010. Household response to the 2008 tax rebate: survey evidence and aggregate implications. In: Brown, J.R. (Ed.), Tax Policy and the Economy, vol. 24. The University of Chicago Press, Chicago, IL, pp. 69–110.

Sahm, C.R., Shapiro, M.D., Slemrod, J., 2012. Check in the mail or more in the paycheck: does the effectiveness of fiscal stimulus depend on how it is delivered? Am. Econ. J. Econ. Pol. 4 (3), 216–250.

Saxenian, A.L., 1999. Silicon Valley's New Immigrant Entrepreneurs. Public Policy Institute of California, San Francisco, CA.

Scholnick, B., 2013. Consumption smoothing after the final mortgage payment: testing the magnitude hypothesis. Rev. Econ. Stat. 95 (4), 1444–1449.

Serrato, J.C.S., Wingender, P., 2014. Estimating Local Fiscal Multipliers. Mimeo, Duke University.

Shapiro, J.M., 2005. Is there a daily discount rate? Evidence from the food stamp nutrition cycle. J. Publ. Econ. 89 (2-3), 303–325.

Shapiro, M.D., Slemrod, J., 2003. Consumer response to tax rebates. Am. Econ. Rev. 93 (1), 381–396.

Shea, J., 1995. Union contracts and the life-cycle/permanent-income hypothesis. Am. Econ. Rev. 85 (1), 186–200.

Shoag, D., 2013. Using state pension shocks to estimate fiscal multipliers since the great recession. Am. Econ. Rev. 103 (3), 121–124.

Shoag, D., 2015. The Impact of Government Spending Shocks: Evidence on the Multiplier from State Pension Plan Returns. Mimeo, Harvard University.

Shue, K., 2013. Executive networks and firm policies: evidence from the random assignment of MBA peers. Rev. Financ. Stud. 26 (6), 1401–1442.

Solow, R.M., 1956. A contribution to the theory of economic growth. Q. J. Econ. 70 (1), 65–94.

Souleles, N.S., 1999. The response of household consumption to income tax refunds. Am. Econ. Rev 89 (4), 947–958.

Souleles, N.S., 2000. College tuition and household savings and consumption. J. Publ. Econ. 77 (2), 185–207.

Souleles, N.S., 2002. Consumer response to the Reagan tax cuts. J. Publ. Econ. 85 (1), 99–120.

Stephens, M., 2003. "3rd of tha Month": do social security recipients smooth consumption between checks? Am. Econ. Rev. 93 (1), 406–422.

Stephens, M., 2006. Paycheque receipt and the timing of consumption. Econ. J. 116 (513), 680–701.

Stephens, M., 2008. The consumption response to predictable changes in discretionary income: evidence from the repayment of vehicle loans. Rev. Econ. Stat. 90 (2), 241–252.

Stephens, M., Unayama, T., 2011. The consumption response to seasonal income: evidence from Japanese public pension benefits. Am. Econ. J. Appl. Econ. 3 (4), 86–118.

Stiglitz, J.E., 1990. Peer monitoring and credit markets. World Bank Econ. Rev. 4 (3), 351–366. http://dx.doi.org/10.1093/wber/4.3.351.

Swan, T.W., 1956. Economic growth and capital accumulation. Econ. Rec. 32 (2), 334–361.

Tabellini, G., 2010. Culture and institutions: economic development in the regions of Europe. J. Eur. Econ. Assoc. 8 (4), 677–716.

Trezzi, R., Porcelli, F., 2014. Reconstruction multipliers. Board of Governors of the Federal Reserve System (U.S.) No. 2014-79. Finance and Economics Discussion Series.

Wilcox, D.W., 1989. Social security benefits, consumption expenditure, and the life cycle hypothesis. J. Polit. Econ. 97 (2), 288–304.

Wilson, D.J., 2012. Fiscal spending jobs multipliers: evidence from the 2009 American recovery and reinvestment act. Am. Econ. J. Econ. Pol. 4 (3), 251–282.

CHAPTER 13

Accounting for Business Cycles

P. Brinca[*,¶], V.V. Chari[†,‡], P.J. Kehoe[†,‡,§], E. McGrattan[†,‡]
[*]Nova School of Business and Economics, Lisboa, Portugal
[†]University of Minnesota, Minneapolis, MN, United States
[‡]Federal Reserve Bank of Minneapolis, Minneapolis, MN, United States
[§]University College London, London, United Kingdom
[¶]Centre for Economics and Finance, University of Porto, Porto, Portugal

Contents

Abstract

We elaborate on the business cycle accounting method proposed by Chari et al. (2006), clear up some misconceptions about the method, and then apply it to compare the Great Recession across OECD countries as well as to the recessions of the 1980s in these countries. We have four main findings. First,

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.05.002

with the notable exception of the United States, Spain, Ireland, and Iceland, the Great Recession was driven primarily by the efficiency wedge. Second, in the Great Recession, the labor wedge plays a dominant role only in the United States, and the investment wedge plays a dominant role in Spain, Ireland, and Iceland. Third, in the recessions of the 1980s, the labor wedge played a dominant role only in France, the United Kingdom, and Belgium. Finally, overall in the Great Recession, the efficiency wedge played a more important role and the investment wedge played a less important role than they did in the recessions of the 1980s.

Keywords

Great Recession, Labor wedge, Efficiency wedge, Investment wedge, Decomposition of variance

JEL Classification Codes:

E3, E32, F44

In this chapter, we elaborate on the business cycle accounting method proposed by Chari et al. (2006), henceforth CKM, clear up some misconceptions about the method, and then apply it to compare the Great Recession across OECD countries as well as to the recessions of the 1980s in these countries. The goal of the method is to help guide researchers' choices about where to introduce frictions into their detailed quantitative models in order to allow the models to generate business cycle fluctuations similar to those in the data.

The method has two components: an equivalence result and an accounting procedure. The *equivalence result* is that a large class of models, including models with various types of frictions, is equivalent to a *prototype model* with various types of time-varying wedges that distort the equilibrium decisions of agents operating in otherwise competitive markets. At face value, these wedges look like time-varying productivity, labor income taxes, investment taxes, and government consumption. We labeled these wedges *efficiency wedges*, *labor wedges*, *investment wedges*, and *government consumption wedges*.

The *accounting procedure* also has two components. It begins by measuring the wedges, using data together with the equilibrium conditions of a prototype model. The measured wedge values are then fed back into the prototype model, one at a time and in combinations, in order to assess how much of the observed movements of output, labor, and investment can be attributed to each wedge, separately and in combinations.

Here, we use this method to study the Great Recession in OECD countries. We also compare this recession with the recessions of the early 1980s. While the exact timing of the recessions of the early 1980s differs across countries in our OECD sample, most of the countries had a recession between 1980 and 1984. Throughout we refer to the recessions of the early 1980s as the *1982 recession*. We have four main findings. First, with the notable exception of the United States, Spain, Ireland, and Iceland, the Great Recession was driven primarily by the efficiency wedge. Second, in the Great Recession, the labor wedge plays a dominant role only in the United States, and the investment wedge plays a dominant role in Spain, Ireland, and Iceland. Third, in the recessions of the 1980s, the labor wedge played a dominant role only in France, the United Kingdom, and Belgium. Finally, overall in the Great Recession, the efficiency wedge played a much more important role and the investment wedge played a much less important role than they did in the recessions of the 1980s.

We now turn to the elaborating on the equivalence results in CKM that link the four wedges to detailed models. We begin by showing that a detailed economy with fluctuations in investment-specific technological change similar to that in Greenwood et al. (1997) maps into a prototype economy with investment wedges. This result makes clear that investment wedges are by no means synonymous with financial frictions, a point stressed by CKM.

We then consider an economy that blends elements of Kiyotaki and Moore (1997) with that of Gertler and Kiyotaki (2009). The economy has a representative household and heterogenous banks that face collateral constraints. We show that such an economy is equivalent to a prototype economy with investment wedges. This result makes clear that some ways of modeling financial frictions do indeed show up as investment wedges.

Finally, we turn to an economy studied by Buera and Moll (2015) consisting of workers and entrepreneurs. The entrepreneurs have access to heterogeneous production technologies that are subject to shocks to collateral constraints. We follow Buera and Moll (2015) in showing that this detailed economy is equivalent to a prototype model with a labor wedge, an investment wedge, and an efficiency wedge. This equivalence makes the same point as does the input-financing friction economy in CKM, namely that other ways of modeling financial frictions can show up as efficiency wedges and labor wedges.

The point of the three examples just discussed is to help clarify how the pattern of wedges in the data can help researchers narrow down the class of models they are considering. If, for example, most of the fluctuations are driven by the efficiency and labor wedges in the data, then of the three models just considered, the third one is more promising than the first two.

We then turn to models with search frictions. We use these models to make an important point. Researchers should choose the baseline prototype economy that provides the most insights for the research program of interest. In particular, when the detailed economies of interest are sufficiently different from the one-sector growth model, it is often more instructive to adjust the prototype model so that the version of it without wedges corresponds to the planning problem for the class of models at hand. For example, when we map the model with efficient search into the one-sector model, that model does have efficiency and labor wedges, but if we map it into a new prototype model with two capital-like variables, physical capital and the stock of employed workers, the new prototype model has no wedges.

We then consider a search model with an inefficient equilibrium. When we map this model into the new prototype model with two capital-like variables, then the prototype model has only labor wedges. But if we map it into the original prototype model, it has efficiency wedges and (complicated) labor wedges. These findings reinforce the point that it is often more instructive to adjust the prototype model so that the version of it without wedges corresponds to the planning problem for the class of models at hand.

Taken together, these equivalence results help clear up some common misconceptions. The first misconception is that efficiency wedges in a prototype model can only come from technology shocks in a detailed model. In our judgment, by far the least interesting interpretation of efficiency wedges is as narrowly interpreted shocks to the blueprints governing individual firm production functions. More interesting interpretations rest on frictions that deliver such high-frequency movements in this wedge. For example, the input-financing friction model in CKM shows how financial frictions in a detailed model can manifest themselves as efficiency wedges. Indeed, we think that exploring detailed models in which the sudden drops in efficiency wedges experienced in recessions come from frictions such as input-financing frictions is more promising than blaming these drops on abrupt negative shocks to blueprints for technologies. The second misconception is that labor wedges in a prototype model arise solely from frictions in labor markets in detailed economies. The Buera–Moll economy makes clear that this view is incorrect. The third misconception is that investment wedges arise solely due to financial frictions. Clearly, the detailed model with investment-specific technical change shows that this view is also incorrect.

We turn now to describing our procedure. This procedure is designed to answer questions of the following kind: How much would output fluctuate if the only wedge that fluctuated is the efficiency wedge and the probability distribution of the efficiency wedge is the same as in the prototype economy? If the wedges were independent at all leads and lags, the procedure can be implemented in a straightforward manner by letting only, say, the efficiency wedge fluctuate and setting all other wedges to constants. In the data, the wedges are correlated with each other, so the straightforward implementation does not answer our question.

Our implementation views the wedges as being functions of underlying abstract events. In practice, we assume that the dimension of the underlying events is the same as the dimension of the wedges, namely four, and identify each event with one of the wedges. We then use the data to estimate the stochastic process for the underlying events. Given this estimated stochastic process, we can then answer our question by letting the wedge of interest vary with the underlying events in the same way it did in the data but assuming that all other wedges are constant functions of the underlying events. The procedure ensures that the probability distribution over the wedge of interest is the same in the prototype economy with all wedges and in the experiment.

We then briefly discuss what at first seems to be an intuitive way to proceed in which the wedges are identified with the underlying event not only in the estimation but also in the thought experiment. The problem with this procedure is that it does not make clear the conceptual distinction between underlying events and wedges. This distinction is apparent when the wedges are correlated. Indeed, in this case, this procedure makes it impossible to hold all but one wedge constant without changing the probability distribution over the wedge of interest. We note that not keeping clear the conceptual distinction between underlying events and wedges has been the source of some confusion in the literature (see, for example, Christiano and Davis, 2006).

Our business cycle accounting method is intended to shed light on promising classes of mechanisms through which primitive shocks lead to economic fluctuations. It is not intended to identify the primitive sources of shocks. Many economists think, for example, that shocks to the financial sector drove the Great Recession in developed economies, but these economists disagree about the details of the driving mechanism. Our analysis suggests that the transmission mechanism from shocks to the financial sector to broader economic activity must be different in the United States, Spain, Ireland, and Iceland than in the rest of the countries in the OECD. More precisely, our analysis shows that these shocks must manifest themselves as labor wedges in the United States, as investment wedges in Spain, Ireland, and Iceland, and as efficiency wedges in the rest of the OECD.

As CKM argue, the equivalence results provide the logical foundation for the way our accounting procedure uses the measured wedges. At a mechanical level, the wedges represent deviations in the prototype model's first-order conditions, in its relationship between inputs and outputs, and in a variable in the resource constraint. One interpretation of these deviations, of course, is that they are simply errors, so that their size indicates the goodness-of-fit of the model. Under that interpretation, however, feeding the measured wedges back into the model makes no sense. Our equivalence result leads to a more economically useful interpretation of the deviations by linking them directly to classes of models; that link provides the rationale for feeding the measured wedges back into the model.

Also in terms of method, the accounting procedure goes beyond simply plotting the wedges. Such plots, by themselves, are not useful in evaluating the quantitative importance of competing mechanisms of business cycles because they tell us little about the equilibrium responses to the wedges. Feeding the measured wedges back into the prototype model and measuring the model's resulting equilibrium responses is what allows us to discriminate between competing mechanisms.

Related Literature

The chapter most closely related to ours is Ohanian and Raffo (2012), who use a methodology similar to ours to study the Great Recession in 14 OECD countries and compare the peak-to-trough declines in output and hours across countries and recessions. In part, our findings are the same in spirit: we both find that in the Great Recession, the labor wedge plays a dominant role in the United States.

In part our findings are in contrast: they find that in Korea the labor wedge plays a large role in the Great Recession. We instead find that in Korea the efficiency wedge does. We note that both Ohanian and Raffo (2012) and Lopez and Garcia (2014) find that the labor wedge rather than the investment wedge plays a dominant role in the Great

Recession in Spain. Our findings differ from both studies in part because of differences in the treatment of the data, including, for example, how we treat consumer durables and how we deflate nominal variables to make them real. We also differ from Ohanian and Raffo (2012) in terms of methodology: we fit stochastic processes for the wedges, whereas they focus on perfect foresight models. For some related studies, see Mulligan (2009) and Ohanian (2010).

The business cycle accounting methodology has been used for many countries and time periods. For example, it has been used for Portugal by Cavalcanti (2007), for the economies of Brazil, Russia, India, and China by Chakraborty and Otsu (2013), for India by Chakraborty (2006), for the East Asian economies by Cho and Doblas-Madrid (2013), for the United Kingdom by Kersting (2008), for Japan by Kobayashi and Inaba (2006), for Asian economies by Otsu (2010), and for monetary economies by Sustek (2011) and Brinca (2013), and for a variety of countries by Brinca (2014).

1. DEMONSTRATING THE EQUIVALENCE RESULT

Here, we show how various detailed models with underlying distortions are equivalent to a prototype growth model with one or more wedges.

1.1 The Benchmark Prototype Economy

The *benchmark prototype economy* that we use later in our accounting procedure is a stochastic growth model. In each period t, the economy experiences one of finitely many events s_t, which index the shocks. We denote by $s^t = (s_0, \ldots, s_t)$ the history of events up through and including period t and often refer to s^t as the *state*. The probability, as of period 0, of any particular history s^t is $\pi_t(s^t)$. The initial realization s_0 is given. The economy has four exogenous stochastic variables, all of which are functions of the underlying random variable s^t: the *efficiency wedge* $A_t(s^t)$, the *labor wedge* $1 - \tau_{lt}(s^t)$, the *investment wedge* $1/[1 + \tau_{xt}(s^t)]$, and the *government consumption wedge* $g_t(s^t)$.

In the model, consumers maximize expected utility over per capita consumption c_t and per capita labor l_t,

$$\sum_{t=0}^{\infty}\sum_{s^t}\beta^t\pi_t(s^t)U(c_t(s^t), l_t(s^t))N_t,$$

subject to the budget constraint

$$c_t + [1 + \tau_{xt}(s^t)]x_t(s^t) = [1 - \tau_{lt}(s^t)]w_t(s^t)l_t(s^t) + r_t(s^t)k_t(s^{t-1}) + T_t(s^t)$$

and the capital accumulation law

$$(1+\gamma_n)k_{t+1}(s^t) = (1-\delta)k_t(s^{t-1}) + x_t(s^t), \tag{1}$$

where $k_t(s^{t-1})$ denotes the per capita capital stock, $x_t(s^t)$ per capita investment, $w_t(s^t)$ the wage rate, $r_t(s^t)$ the rental rate on capital, β the discount factor, δ the depreciation rate of capital, N_t the population with growth rate equal to $1 + \gamma_n$, and $T_t(s^t)$ per capita lump-sum transfers.

The production function is $A(s^t)F(k_t(s^{t-1}), (1+\gamma)^t l_t(s^t))$, where $1 + \gamma$ is the rate of labor-augmenting technical progress, which is assumed to be a constant. Firms maximize profits given by $A_t(s^t)F\left(k_t(s^{t-1}), (1+\gamma)^t l_t(s^t)\right) - r_t(s^t)k_t(s^{t-1}) - w_t(s^t)l_t(s^t)$.

The equilibrium of this benchmark prototype economy is summarized by the resource constraint,

$$c_t(s^t) + x_t(s^t) + g_t(s^t) = y_t(s^t), \tag{2}$$

where $y_t(s^t)$ denotes per capita output, together with

$$y_t(s^t) = A_t(s^t)F\left(k_t(s^{t-1}), (1+\gamma)^t l_t(s^t)\right), \tag{3}$$

$$-\frac{U_{lt}(s^t)}{U_{ct}(s^t)} = [1 - \tau_{lt}(s^t)]A_t(s^t)(1+\gamma)^t F_{lt}, \text{ and} \tag{4}$$

$$U_{ct}(s^t)[1 + \tau_{xt}(s^t)] \tag{5}$$

$$= \beta \sum_{s^{t+1}} \pi_t(s^{t+1}|s^t)U_{ct+1}(s^{t+1})\{A_{t+1}(s^{t+1})F_{kt+1}(s^{t+1}) + (1-\delta)[1 + \tau_{xt+1}(s^{t+1})]\},$$

where, here and throughout, notations such as U_{ct}, U_{lt}, F_{lt}, and F_{kt} denote the derivatives of the utility function and the production function with respect to their arguments and $\pi_t(s^{t+1}|s^t)$ denotes the conditional probability $\pi_t(s^{t+1})/\pi_t(s^t)$. We assume that $g_t(s^t)$ fluctuates around a trend of $(1+\gamma)^t$.

Notice that in this benchmark prototype economy, the efficiency wedge resembles a blueprint technology parameter, and the labor and the investment wedges resemble tax rates on labor income and investment. Other more elaborate models could be considered, such as models with other kinds of frictions that look like taxes on consumption or capital income. Consumption taxes induce a wedge between the consumption-leisure marginal rate of substitution and the marginal product of labor in the same way as do labor income taxes. Such taxes, if time-varying, also distort the intertemporal margins in (5). Capital income taxes induce a wedge between the intertemporal marginal rate of substitution and the marginal product of capital, which is only slightly different from the distortion induced by a tax on investment. We experimented with intertemporal distortions that resemble capital income taxes rather than investment taxes and found that our substantive conclusions are unaffected. (For details, see the Appendix.)

We emphasize that each of the wedges represents the overall distortion to the relevant equilibrium condition of the model. For example, distortions to labor supply affecting consumers and to labor demand affecting firms both distort the static first-order condition

(4). Our labor wedge represents the sum of these distortions. Thus, our method identifies the overall wedge induced by both distortions and does not identify each separately. Likewise, liquidity constraints on consumers distort the consumer's intertemporal Euler equation, whereas investment financing frictions on firms distort the firm's intertemporal Euler equation. Our method combines the Euler equations for the consumer and the firm and therefore identifies only the overall wedge in the combined Euler equation given by (5). We focus on the overall wedges because what matters in determining business cycle fluctuations is the overall wedges, not each distortion separately.

For the equivalence results that follow, it is notationally convenient to work with the prototype model just described. For our quantitative results, we add investment adjustment costs by replacing the capital accumulation law (1) with

$$(1+\gamma_n)k_{t+1}(s^t) = (1-\delta)k_t(s^{t-1}) + x_t(s^t) - \phi\left(\frac{x_t(s^t)}{k_t(s^{t-1})}\right), \tag{6}$$

where ϕ represents the per unit cost of adjusting the capital stock. We follow the macroeconomic literature in assuming that the adjustment costs are parameterized by the function

$$\phi\left(\frac{x}{k}\right) = \frac{a}{2}\left(\frac{x}{k} - b\right)^2,$$

where $b = \delta + \gamma + \gamma_n$ is the steady-state value of the investment–capital ratio.

1.2 The Mapping—From Frictions to Wedges

Now we illustrate the mapping between detailed economies and prototype economies for several types of wedges. We show that investment-specific technical change in a detailed economy maps into investment wedges in our prototype economy. Likewise, bank collateral constraints also map into investment wedges in our prototype economy. We then consider an economy with heterogeneous productivity and collateral constraints and show that it maps into a prototype economy with efficiency, labor, and investment wedges. Finally, we consider a search model with efficient allocations and show it maps into a prototype economy with a labor wedge and an efficiency wedge but no investment wedge. The four economies we use to illustrate this mapping are closed economies for which the associated government consumption wedge in the prototype economy is identically zero. Hence, we focus on the other three wedges and make no mention of the government consumption wedge.

We choose simple models in order to illustrate how the detailed models map into the prototypes. Since many models map into the same configuration of wedges, identifying one particular configuration does not uniquely identify a model; rather, it identifies a whole class of models consistent with that configuration. This point is seen clearly when comparing the prototype model associated with the economy with investment-specific technical change to that for the economy with bank collateral constraints. In this sense, our method does not uniquely determine the model most promising to analyze business cycle fluctuations. It does, however, guide researchers to focus on the key margins that need to be distorted in order to capture the nature of the fluctuations.

1.2.1 An Equivalence Result for a Model with Investment-Specific Technical Change

We begin with a two-sector model with investment-specific technical change and show how it maps into a prototype economy with only investment wedges.

1.2.1.1 A Detailed Economy with Investment Specific Technical Change

The detailed economy has consumption $c_t(s^t)$ and investment $x_t(s^t)$ produced according to

$$c_t(s^t) = A_t(s^t)F(k_{ct}(s^t), l_{ct}(s^t)) \text{ and } x_t(s^t) = A_{xt}(s^t)A_t(s^t)F(k_{xt}(s^t), l_{xt}(s^t), \tag{7}$$

where $k_{ct}(s^t)$ and $l_{ct}(s^t)$ denote capital and labor used to produce consumption goods, $k_{xt}(s^t)$ and $l_{xt}(s^t)$ denote capital and labor used to produce investment goods, $A_t(s^t)$ is neutral technical change, $A_{xt}(s^t)$ denotes investment-specific technical change, and F satisfies constant returns to scale. The timing is that the (total) capital stock in use at period t is chosen at the end of period $t-1$ given the shock history s^{t-1}, whereas at the beginning of each period, after the current shock s_t is realized, labor and capital are allocated between sectors. This timing gives rise to a capital accumulation rule

$$k_{t+1}(s^t) = (1-\delta)k_t(s^{t-1}) + x_t(s^t) \tag{8}$$

and adding up constraints for sectoral capital allocation,

$$k_{ct}(s^t) + k_{xt}(s^t) \leq k_t(s^{t-1}), \tag{9}$$

and sectoral labor allocation,

$$l_{ct}(s^t) + l_{xt}(s^t) \leq l_t(s^t). \tag{10}$$

The planning problem is to choose allocations to solve

$$max \sum_{s^t} \beta^t \mu(s^t) U(c_t(s^t), l_t(s^t))$$

subject to (7)–(10). Using that the production function F has constant returns to scale, the first-order conditions imply that

$$\frac{k_{ct}(s^t)}{l_{ct}(s^t)} = \frac{k_{xt}(s^t)}{l_{xt}(s^t)} = \frac{k_t(s^{t-1})}{l_t(s^t)},$$

and hence

$$F_{kc}(k_{ct}(s^t), l_{ct}(s^t)) = F_{kx}(k_{xt}(s^t), l_{xt}(s^t)) \text{ and } F_{lc}(k_{ct}(s^t), l_{ct}(s^t)) = F_{lx}(k_{xt}(s^t), l_{xt}(s^t)),$$

and we can write these marginal products as $F_k(k(s^{t-1}), l(s^t))$ and $F_l(k(s^{t-1}), l(s^t))$. The Euler equation is

$$\frac{U_{ct}(s^t)}{A_{xt}(s^t)} = \sum_{s_{t+1}} \beta\mu(s^{t+1}|s^t)\left[U_{ct+1}(s^{t+1})A_{t+1}(s^{t+1})F_k(s^{t+1}) + (1-\delta)\frac{U_{ct}(s^{t+1})}{A_{xt+1}(s^{t+1})}\right],$$

and the static first-order condition for labor is given by

$$-\frac{U_{lt}(s^t)}{U_{ct}(s^t)} = A_t(s^t)F_l(s^t).$$

If we express output in current consumption units, we can write

$$A_t(s^t)F(k_{ct}(s^t), l_{ct}(s^t)) + q_t(s^t)A_{xt}(s^t)A_t(s^t)F(k_{xt}(s^t), l_{xt}(s^t) = A_t(s^t)F(k(s^{t-1}, l(s^t)))$$

since the relative price of investment to consumption goods is $q_t(s^t) = 1/A_{xt}(s^t)$.

1.2.1.2 The Associated Prototype Economy with Investment Wedges

Now consider a prototype economy with just investment wedges. This prototype economy has a productivity shock $A_t(s^t)$ equal to that in the consumption goods sector in the detailed economy, an investment wedge equal to the reciprocal of the level of investment-specific technical change, and no other wedges.

Proposition 1 *The aggregate allocations in the detailed economy with investment-specific technical change coincide with those of the prototype economy if the efficiency wedge in the prototype economy equals the productivity shock in the consumption goods sector, the investment wedge is given by*

$$1 - \tau_{xt}(s^t) = \frac{1}{A_{xt}(s^t)},$$

and the labor wedge is zero.

Note that if we measure output in the detailed economy at base period prices rather than at current prices, the map between the detailed economy and the prototype economy is more complicated.

1.2.2 An Equivalence Result for an Economy with Bank Collateral Constraints

Here, we show the equivalence between an economy with bank collateral constraints and a prototype economy with only investment wedges.

1.2.2.1 A Detailed Economy with Bank Collateral Constraints

Consider an infinite horizon economy that blends elements of Kiyotaki and Moore (1997) with that of Gertler and Kiyotaki (2009) and is composed of a household that works and operates financial intermediaries, referred to as *banks*, together with firms and a government. Households elastically supply labor and save by holding deposits in banks and government bonds and receive dividends. Banks raise deposits from households and use these deposits plus retained earnings to invest in capital as well as to pay dividends to consumers. Firms rent capital and labor and produce output. The government finances an exogenous stream of government spending by taxing labor income and the capital stock and by selling government bonds.

Let the state of the economy be $s_t \in S$ distributed according to $\pi(s_t|s_{t-1})$. Let $s^t = (s_0, \ldots, s_t)$. The resource constraint is given by

$$C_t(s^t) + K_{t+1}(s^t) = A_t(s^t)F\left(K_t\left(s^{t-1}\right), L_t(s^t)\right), \tag{11}$$

where C_t is aggregate consumption, K_{t+1} is the capital stock, L_t is aggregate labor, and F is a constant returns to scale production function that includes the undepreciated capital stock. Throughout we use the convention that uppercase letters denote aggregates and lowercase letters denote the decisions of individual households or banks.

We follow Gertler and Karadi (2011) and Gertler et al. (2012) in the formulation of households. The decision making in each household can be thought of as being made by different entities: a measure 1 of workers and a measure 1 of bankers. The workers supply labor and return their wages to the household while each banker manages a bank that transfers nonnegative dividends to the household. The household as a whole has preferences

$$\sum_{t=0}^{\infty} \sum_{s^t} \beta^t \pi(s^t|s_0) U(c_t(s^t), l_t(s^t)), \tag{12}$$

where c_t and l_t are an individual household's consumption and labor supply. Given initial asset holdings b_{H0} and d_0, the stand-in household in the economy maximizes this utility by choosing $\{c_t, l_t, d_{t+1}\}$ subject to the budget constraint

$$c_t(s^t) + \sum_{s^{t+1}} q_{t+1}\left(s^{t+1}\right) d_{t+1}\left(s^{t+1}\right) \le w_t(s^t) l_t(s^t) + d_t(s^t) + X_t(s^t) - \frac{1-\sigma}{\sigma}\bar{n}$$

and the restrictions that

$$d_{t+1}\left(s^{t+1}\right) \ge \bar{d}, \tag{13}$$

where $\bar{d}$ is a large negative number. Here, d_{t+1} is the amount of deposits made by households in banks and q_{t+1} is the corresponding price. Also, w_t is the real wage, X_t are dividends paid by banks, and $\bar{n}$ is the amount of initial equity given to each newly formed bank of which a measure $(1 - \sigma)/\sigma$ is formed each period.

The first-order conditions for the household's problem can be summarized by

$$-\frac{U_{Lt}(s^t)}{U_{Ct}(s^t)} = w_t(s^t) \tag{14}$$

$$q_{t+1}\left(s^{t+1}\right) = \frac{\beta\pi(s^{t+1}|s^t) U_{Ct+1}(s^{t+1})}{U_{Ct}(s^t)}. \tag{15}$$

A representative firm rents capital at rate R_t from banks and hires L_t units of labor to maximize profits

$$\max_{K_t, L_t} AF(K_t, L_t) + (1-\delta)K_t - R_t K_t - w_t L_t. \tag{16}$$

The first-order conditions to this problem imply

$$AF_K(s^t) + 1 - \delta = R_t(s^t) \text{ and } AF_L(s^t) = w_t(s^t). \tag{17}$$

Next consider the banks. At the beginning of each period, an idiosyncratic random variable is realized at each existing bank. With probability σ, the bank will continue in operation until the next period. With probability $1 - \sigma$, the bank ceases to exist and, by assumption, pays out all of its accumulated net worth as dividends to the household. Also at the beginning of each period, a measure $(1 - \sigma)/\sigma$ of new banks is born, each of which is given an exogenously specified amount of initial equity $\bar{n}$ from households. Since only a fraction σ of these newborn banks survive until the end of the period, the measure of surviving banks is always constant at 1. This device of having banks die is a simple way to ensure that they do not build up enough equity to make the financial constraints that we will next introduce irrelevant.

Turning to the budget constraint of an individual bank, note first that for any non-newborn bank the budget constraint at t is

$$x_t(s^t) + k_{t+1}(s^t) - \sum_{s^{t+1}} q_{t+1}\left(s^{t+1}\right) d_{t+1}\left(s^{t+1}\right) \le R_t(s^t) k_t\left(s^{t-1}\right) - d_t(s^t), \tag{18}$$

where R_t is the rental rate for capital. We will let $n_t(s^t) = R_t(s^t)k_t(s^{t-1}) - d_t(s^t)$ denote the right side of (18) and will refer to it as the *net worth* of the bank. For a bank that is newly born at t, the left side of the budget constraint is the same and the right side of (18) is replaced by initial net worth $\bar{n}$. Banks face a *collateral* constraint for each s^{t+1},

$$d_{t+1}\left(s^{t+1}\right) \le \gamma R_{t+1}\left(s^{t+1}\right) k_{t+1}(s^t), \tag{19}$$

where $0 < \gamma < 1$, as well as nonnegativity constraints on dividends and bond holdings,

$$x_t(s^t) \ge 0. \tag{20}$$

For notational simplicity only, consider the problem of a bank born in period 0. The bank chooses $\{k_{t+1}(s^t), d_t(s^t), x_t(s^t)\}$ to solve

$$\max \sum_t^{\infty} \sum_{s^t} Q(s^t)\sigma^t[\sigma x_t(s^t) + (1-\sigma) n_t(s^t)] \tag{21}$$

subject to (18)–(20) where $n_t(s^t) = R_t(s^t) k_t\left(s^{t-1}\right) - d_t(s^t)$, $n_0\left(s^0\right) = \bar{n}$, and $Q(s^t)$ is the price of a good at date t in units of a good at date 0 after history s^t. We assume that a bank that ceases to operate pays out its accumulated net worth as dividends. Since the bank is owned by the household, it values dividends at the marginal rates of substitution of consumers, so that

$$Q(s^t) = \beta^t \pi(s^t) U_C(s^t) / U_{C0}(s^0). \tag{22}$$

From the household's first-order condition, it follows that the discount factor used by the bank is consistent with the rate of return on deposits in that $Q(s^t) = q_0(s^0) \cdots q_t(s^t)$.

The first-order conditions to the bank's problem can be written as

$$Q(s^t)\sigma^{t+1} + \eta_{xt}(s^t) = \lambda_t(s^t)$$

$$\lambda_t(s^t) = \sum_{s^{t+1}} \left[Q(s^{t+1})\sigma^{t+1}(1-\sigma)R_{t+1}(s^{t+1}) + R_{t+1}(s^{t+1})\left(\lambda_{t+1}(s^{t+1}) + \gamma\mu_{t+1}(s^{t+1})\right)\right]$$

$$-Q(s^{t+1})\sigma^{t+1}(1-\sigma) + \lambda_t(s^t)q_{t+1}(s^{t+1}) = \lambda_{t+1}(s^{t+1}) + \mu_{t+1}(s^{t+1}), \tag{23}$$

where $\lambda_t(s^t)$, $\mu_t(s^t)$, and $\eta_{xt}(s^t)$ are the multipliers on the bank budget constraint, the collateral constraint, and the nonnegative dividend constraint. We can manipulate these constraints to obtain

$$1 = \sum_{s^{t+1}} \left[R_{t+1}(s^{t+1}) q_{t+1}(s^{t+1}) \left(1 - (1-\gamma)\frac{\mu_{t+1}(s^{t+1})}{\lambda_t(s^t) q_{Dt+1}(s^{t+1})} \right) \right]. \tag{24}$$

A competitive equilibrium is defined in the standard fashion.

1.2.2.2 The Associated Prototype Economy with Investment Wedges

Consider a version of the benchmark prototype economy that will have the same aggregate allocations as the banking economy just detailed. This prototype economy is identical to our benchmark prototype except that the new prototype economy has an investment wedge that resembles a tax on capital income rather than a tax on investment. Here, the government consumption wedge is set equal to zero.

In the prototype economy, the consumer's budget constraint is

$$C_t(s^t) + K_{t+1}(s^t) = (1 - \tau_{Kt}(s^t))R_t(s^t)K_t(s^{t-1}) + (1 - \tau_{Lt}(s^t))w_t(s^t)L_t(s^t) + T_t(s^t). \tag{25}$$

The first-order condition for the investment wedge in this economy is given by

$$U_{Ct}(s^t) = \sum_{s^{t+1}} \beta\mu(s^{t+1}|s^t)U_{Ct+1}(s^{t+1})[AF_{Kt+1}(s^{t+1}) + 1 - \delta](1 - \tau_{Kt+1}(s^{t+1}). \tag{26}$$

Comparing the first-order conditions in the detailed economy with bank collateral constraints to those of the associated prototype economy leads us to set

$$\tau_{Kt}(s^t) = (1-\gamma)\frac{\mu_{t+1}(s^{t+1})}{\lambda_t(s^t)q_{Dt+1}(s^{t+1})}. \tag{27}$$

We then have the following proposition.

Proposition 2 *The aggregate allocations in the detailed economy with bank collateral constraints coincide with those of the prototype economy if the efficiency wedge in the prototype economy $A_t(s^t) = A$, the labor wedge is zero, and the investment wedge is given by (27).*

Clearly, the efficiency wedge here is just the constant level of technology A in the detailed economy. To see why there is no labor wedge, note that combining (14) and (17) gives that

$$-\frac{U_L(s^t)}{U_C(s^t)} = AF_L(s^t).$$

To derive the expression for the investment wedge, substitute for $R_{t+1}(s^{t+1})$ from the firm's first-order condition (17) and for $q_{t+1}(s^{t+1})$ from the consumer's first-order condition (15) to obtain

$$1 = \sum_{s^{t+1}} \left[\beta\mu(s^{t+1}|s^t) \frac{U_{Ct+1}(s^{t+1})}{U_{Ct}(s^t)} [AF_{Kt+1}(s^{t+1}) + 1 - \delta] \left(1 - (1-\gamma)\frac{\mu_{t+1}(s^{t+1})}{\lambda_t(s^t)q_{Dt+1}(s^{t+1})}\right)\right]$$

and compare (26) to this equation.

1.2.3 *An Equivalence Result for an Economy with Heterogeneous Productivity and Collateral Constraints*

We use an example from Buera and Moll (2015) to illustrate how a model with fluctuations in financial frictions, modeled as shocks to a collateral constraint on entrepreneurs, is equivalent to a prototype model with a labor wedge, an investment wedge, and an efficiency wedge. We think of this example as making a point identical to that in proposition 1 of Chari et al. (2006) but in a different context. That proposition showed how a detailed model with financial frictions modeled as input-financing frictions is equivalent to a prototype economy with a labor wedge, an investment wedge, and an efficiency wedge.

1.2.3.1 A Detailed Economy with Heterogeneous Productivity and Collateral Constraints

We consider an economy with only idiosyncratic shocks and exogenous incomplete markets against these shocks. A unit mass of identical workers supply labor L_t at a wage w_t, who can neither borrow nor lend, maximize

$$\sum_{t=0}^{\infty} \beta^t [\log(C_{Wt}) - V(L_t)]$$

subject to

$$C_{Wt} = w_t L_t. \tag{28}$$

The economy has a unit mass of entrepreneurs indexed by $i \in [0, 1]$ and a unit mass of identical households. An entrepreneur of type i draws an idiosyncratic shock z_{it} which is i.i.d. over time and across entrepreneurs and has density $\psi(z)$. This entrepreneur has a technology to produce output of the form $y_{it} = z_{it}^{\alpha} k_{it}^{\alpha} l_{it}^{1-\alpha}$ where k_{it} and l_{it} are the amounts of capital invested and labor hired by entrepreneur i.

The timing is that an entrepreneur's productivity in period $t + 1$, namely z_{it+1}, is revealed at the end of period t, before the entrepreneur issues new debt d_{t+1}. Written in recursive form, an entrepreneur with utility function $\sum \beta^t \log(c_t)$ solves

$$V_t(k, d, z_{-1}, z) = \max_{c, d', k'} \log c + \beta E[V_{t+1}(k', d', z, z')]$$

subject to a budget constraint

$$c + k' - d' = \Pi(z_{-1}, w, k) + (1-\delta)k - (1 + r_t)d$$

and a collateral constraint

$$d' \leq \theta_t k' \text{ with } \theta_t \in [0, 1]. \tag{29}$$

Note that (29) restricts the amount of leverage d'/k' to be less than some exogenous amount, θ_t. We use the constant returns to scale production function and the multiplicative technology shock to write total profits $\mathit{\Pi}(z_{-1}, w, k)$ as linear functions of the technology shock and the capital stock so that

$$\mathit{\Pi}(z_{-1}, w, k) = z\pi(w)k = \max_{l} (zk)^{\alpha} l^{1-\alpha} - wl,$$

where $\pi(w) = \alpha\left(\dfrac{1-\alpha}{w}\right)^{(1-\alpha)/\alpha}$.

An equilibrium consists of sequences of prices $\{r_t, w_t\}$ and quantities such that the allocations solve both the entrepreneur problem and the household problem, and markets clear in that

$$\int d_{it} di = 0 \text{ and } \int l_{it} di = L_t \tag{30}$$

$$C_{Et} + C_{Wt} + X_t = Y_t$$

$$K_{t+1} = X_t + (1-\delta)K_t,$$

where X_t denotes aggregate investment. To characterize the equilibrium, we let m_{it} denote the entrepreneur's cash on hand given by

$$m_{it} \equiv z_{it}\pi_t k_{it} + (1-\delta)k_{it} - (1 + r_t)d_{it},$$

and we let a_{it} denote the net worth of the entrepreneur,

$$a_{it} \equiv k_{it} - d_{it}.$$

We can use this notation to rewrite the dynamic programming problem of the entrepreneur as a two-stage budgeting problem: first choose how much net worth d' to carry over to the next period, and then in the second stage, conditional on d', decide how to split this net worth between capital k' and bonds $-d'$. The two-stage problem is then to solve

$$v_t(m,z) = \max_{d'}\left[\log(m-d') + \beta E v_{t+1}(\tilde{m}_{t+1}(d',z),z')\right],$$

where

$$\tilde{m}_{t+1}(d',z) = \max_{k',d'} z\pi_{t+1}k' + (1-\delta)k' - (1+r_{t+1})d'$$

subject to

$$k' - d' = a',$$

and

$$k' \leq \lambda_t d' \text{ where } \lambda_t = \frac{1}{1-\theta_t} \in [1,\infty). \tag{31}$$

This formulation immediately implies the following result.

Lemma 1 *There is a productivity cutoff for being active* $\underline{z}_{t+1}$ *defined by* $\underline{z}_{t+1}\pi(w_{t+1}) = r_{t+1} + \delta$. *Given this cutoff, capital and debt holdings are given by*

$$k_{it+1} = \left\{\begin{matrix} \lambda_t a_{it+1} & \textit{for } z_{it+1} \geq \underline{z}_{t+1} \\ 0 & \textit{otherwise} \end{matrix}\right\}, d_{it+1} = \left\{\begin{matrix} (\lambda_t - 1)a_{it+1} & \textit{for } z_{it+1} \geq \underline{z}_{t+1} \\ -a_{it+1} & \textit{otherwise} \end{matrix}\right\}, \tag{32}$$

and entrepreneurs save a constant fraction of cash on hand $a_{it+1} = \beta m_{it}$.

Note that the optimal capital choice is always at one of two corners. Sufficiently unproductive entrepreneurs lend out all their net worth for use by other entrepreneurs and receive return $r_{t+1} + \delta$, whereas sufficiently productive entrepreneurs borrow the maximal amount allowed by the collateral constraint, $\lambda_t a_{it+1}$, and invest these funds in their own projects. The marginal entrepreneur has a productivity that makes the returns from investing in capital, $\underline{z}_{t+1}\pi_{t+1}$, just equal to the returns to lending out funds, $r_{t+1} + \delta$.

We can use this characterization of decision rules together with market clearing conditions to determine the cutoff $\underline{z}_{t+1}$ as a function of the parameters of the economy. To do so, we aggregate over entrepreneurs to obtain

$$K_{t+1} = \beta[\alpha Y_t + (1-\delta)K_t] \text{ and } Y_t = A_t K_t^{\alpha} L_t^{1-\alpha},$$

where

$$A_t = \left(\frac{\int_{\underline{z}_t} z\psi(z)dz}{1-\Psi(\underline{z}_t)}\right)^{\alpha} = (E[z|z \geq \underline{z}_t])^{\alpha}, \tag{33}$$

where $\underline{z}_t$ is given by the solution to

$$\lambda_{t-1}(1-\Psi(\underline{z}_t)) = 1. \tag{34}$$

To understand the determination of the cutoff productivity level, we use the results of Lemma 1 to obtain that

$$d_{it+1}=\left\{\begin{array}{c}(\lambda_t-1)\beta m_{it}\text{ for }z_{it+1}\geq\underline{z}_{t+1}\\ -\beta m_{it}\text{ otherwise}\end{array}\right\}.$$

Using the observation that m_{it} chosen before z_{it+1} is realized and is therefore independent of z_{it+1}, we can write the market clearing condition for debt given in (30) for period $t+1$ as

$$(\lambda_t-1)\int_{\underline{z}_{t+1}}^{\infty}\psi(z)dz=\int_0^{\underline{z}_{t+1}}\psi(z)dz,$$

which, when rearranged, yields (34).

1.2.3.2 The Associated Prototype Economy with Efficiency, Labor, and Investment Wedges

Consider a version of the benchmark prototype economy that will have the same aggregate allocations as the banking economy just detailed. This prototype economy is identical to our benchmark prototype except that the new prototype economy has an investment wedge that resembles a tax on capital income rather than a tax on investment.

This economy can be mapped into our prototype economy with a period utility function of the form $U(C_t,L_t)=\log C_t-V(L_t)$ as follows. The efficiency wedge is given by (33), the labor wedge is given by

$$\tau_{Lt}=-\frac{C_{Et}}{C_{Wt}}, \tag{35}$$

and the investment wedge is defined recursively from

$$\frac{U_{ct}}{U_{ct+1}}\tau_{xt}=\beta(1-\delta)\tau_{xt+1}+\frac{C_{Wt}}{C_t}\left(\frac{C_{Wt+1}}{C_{Wt}}-\frac{C_{Et+1}}{C_{Et}}\right) \tag{36}$$

with $\tau_{x0}=0$. To derive (35), note that the labor wedge in the prototype economy is given by

$$C_tV'(L_t)=(1-\tau_{Lt})F_{Lt}. \tag{37}$$

Next, note that in the detailed economy, the first-order condition for the worker can be manipulated to yield $L_tV'(L_t)=1$. Using this condition along with $w_t=F_{Lt}$, $C_t=C_{Wt}+C_{Et}$, and $C_{Wt}=w_tL_t$ in (37) yields (35). Note that (36) can be obtained by using the result that entrepreneurs save a constant fraction of their wealth.

Proposition 3 *The aggregate allocations in the detailed economy with heterogeneous productivity and a collateral constraint coincide with those of the prototype economy if the efficiency wedge in the prototype economy is given by (33), the labor wedge is given by (35), and the investment wedge is given by (36).*

1.2.4 An Equivalence Result for an Economy with Efficient Search

Consider the efficient outcomes from a standard search model. We will show that if we view the outcomes of this model through the lens of a prototype growth model, the prototype model has a labor wedge and an efficiency wedge but no investment wedge.

1.2.4.1 A Detailed Economy with Efficient Search

For simplicity, we focus on a version of the model without aggregate uncertainty. The technology is as follows. The population is normalized to 1. In each period, measure n_t of the population is employed and the rest are unemployed. Of the employed, measure v_t is used as recruiters and $n_t - v_t$ are used in producing the single consumption–investment good. The matching technology depends on the measure of recruiters and the measure of unemployed, $1 - n_t$. The measure of new matches m_t created in any period is given by the constant returns to scale function $G(v_t, 1 - n_t)$. Existing matches dissolve at an exogenous rate δ_n so that the law of motion for the measure of employed is given by

$$n_{t+1} \leq (1-\delta_n)n_t + m_t \tag{38}$$

and the resource constraint for goods is

$$c_t + k_{t+1} \leq y_t + (1-\delta)k_t, \tag{39}$$

where c_t is consumption, k_{t+1} is the capital stock, $y_t = A_tF(k_t, n_t - v_t)$, and δ is the depreciation rate. We assume that $F = k_t^{\alpha}(n_t - v_t)^{1-\alpha}$. The utility of the stand-in household is given by

$$\sum \beta^t U(c_t, n_t). \tag{40}$$

The social planner's problem is to choose $\{c_t, v_t, n_{t+1}, k_{t+1}\}$ to maximize utility subject to (38) and (39). We summarize the key first-order conditions as

$$U_{ct} = \beta U_{ct+1}\left[\frac{\alpha y_{t+1}}{k_{t+1}} + 1 - \delta\right], \tag{41}$$

$$U_{ct}\frac{F_{nt}}{G_{1t}} = \beta U_{ct+1}\left\{\frac{F_{nt+1}}{G_{1t+1}}[(1-\delta_n) - G_{2t+1}] + F_{nt+1} + \frac{U_{nt+1}}{U_{ct+1}}\right\}, \tag{42}$$

and

$$y_t = A_tF(k_t, n_t - v_t).$$

1.2.4.2 The Associated Prototype Economy with Efficiency and Labor Wedges

Consider a prototype economy in which the production function is $y_t = \hat{A}_t k_t^{\alpha} n_t^{1-\alpha}$ where

$$\hat{A}_t = A_t\left(\frac{n_t - v_t}{n_t}\right)^{\alpha} \tag{43}$$

and the resource constraint is the same as in (39). Lagging and manipulating (42) and using

$$A_tF_{nt} = (1-\alpha)\frac{y_t}{n_t - v_t} = (1-\alpha)\frac{y_t}{n_t}\frac{n_t}{n_t - v_t},$$

we obtain

$$\frac{U_{nt}}{U_{ct}(1-\alpha)y_t/n_t} = \left(\frac{n_t}{n_t - v_t}\right)\left[\frac{U_{ct-1}}{\beta U_{ct}}\frac{F_{nt-1}}{A_tF_{nt}}\frac{1}{G_{1t-1}} - \frac{1}{G_{1t}A_t}[(1-\delta_n) - G_{2t}] - \frac{1}{A_t}\right]. \tag{44}$$

Since the labor wedge in the prototype economy is given by the right side of (44), we have the following result.

Proposition 4 *The aggregate allocations in the efficient search economy coincide with those of the prototype economy if the efficiency wedge in the prototype economy is given by (43), the labor wedge $1 - \tau_{lt}$ is given by the right side of (44), and the investment wedge is zero.*

1.3 Adjusting the Prototype Economy

So far we have always established equivalence results between a given detailed economy and the prototype one-sector growth model. When using business cycle accounting logic, one can always do that. When the underlying economy is sufficiently different from the one-sector growth model, however, it is often more instructive to adjust the prototype model so that the version of it without wedges is the planning problem for the class of models at hand.

1.3.1 An Equivalence Result for an Economy with Inefficient Search

Here, we illustrate what we mean by considering a version of the search model in which search is inefficient in that the equilibrium of the economy does not solve the planning problem just discussed. One alternative is to keep the prototype model as the one-sector growth model, in which case the wedges will simply be more elaborate versions of those just discussed. Here, we illustrate an alternative: we now measure the wedges relative to a distorted version of the social planning problem just studied.

1.3.1.1 A Detailed Economy with Inefficient Search

Consider the decentralized equilibrium of a standard search model. The matching technology is as before: the measure of new matches m_t created in any period is given by the constant returns to scale function $G(v_t, 1 - n_t)$. Letting $\theta_t = v_t/(1 - n_t)$ be the number of recruiters per unemployed worker, each firm that uses the recruiting technology attracts $\lambda_f(\theta_t) = G(v_t, 1 - n_t)/v_t$ per recruiter to the firm. Thus, a measure of recruiters v_t attracts $v_t\lambda_f(\theta_t)$ workers to the firm. The probability that an unemployment worker finds a job is $\lambda_w(\theta) = G(v_t, 1 - n_t)/(1 - n_t)$. Note for later that under constant returns to scale, $\lambda_w(\theta) = \theta\lambda_f(\theta)$.

Here, as is standard, we imagine that workers are part of a family which has idiosyncratic risk across its members. Since we abstract from aggregate shocks, the law of large numbers implies that the family solves a deterministic problem. As we did earlier, we assume that productivity deterministically varies over time. To keep notation simple, we only index the value function and the prices by time. The problem of a family written in recursive form is

$$V_t(a_t, n_t) = \max_{c,a'} U(c,n) + \beta V_{t+1}(a', n')$$

subject to the household budget constraint and the transition law for employed workers,

$$c + q_{t+1} a' = a + wn, \tag{45}$$

$$n' = (1-\delta_n)n + \lambda_w(\theta)(1-n). \tag{46}$$

In (45), a' is the quantity of goods saved at t and q_{t+1} is the price at t per unit of goods delivered at $t+1$. In (46), δ_n is the separation rate of employed workers and $\lambda_w(\theta)(1-n)$ is the measure of workers that transit from unemployment to employment.

The first-order condition for a' is

$$q_{t+1} U_{ct} = \beta V_{at+1},$$

and using the envelope condition for a, namely $V_{at} = U_{ct}$, gives

$$q_{t+1} U_{ct} = \beta U_{ct+1}. \tag{47}$$

We can use the envelope condition to derive the marginal value to the household of an additional employed worker,

$$V_{nt} = U_{ct} w_t + U_{nt} + \beta[1-\delta_n - \lambda_w(\theta_t)] V_{nt+1}(a', n'), \tag{48}$$

at the equilibrium wage w_t where n' is given from (46). The first term gives the marginal increase in utility from the increased consumption due to having an additional worker earning w_t. The second term gives the decrease in utility from increased work. The third term is the increase in the present value of utility from entering the next period with an additional worker.

In order to determine the wages in Nash bargaining, it is useful to define the value to the family of having an additional employed worker at an arbitrary current wage w. This worker will receive the equilibrium wage in all future periods if employed. This value is

$$\widetilde{V}_{nt}(a,n,w) = U_c(w - w_t) + V_{nt}(a,n).$$

The problem of the firm with a current stock of employed workers n and a current stock of capital k can be written in recursive form as

$$J_t(n,k) = \max_{v,k'} \left\{ z_t F(k, n-v) - [k' - (1-\delta)k] - w_t n + q_{t+1} J_{t+1}((1-\delta_n)n + v\lambda_f(\theta_t), k') \right\},$$

where the transition law from workers employed at this firm is

$$n' = (1-\delta_n)n + v\lambda_f(\theta_t).$$

Here, the flow profits at t are output, $z_tF(k, n - v)$, minus investment, $[k' - (1-\delta)k]$, minus the wage bill, w_tn. The firm discounts the present value of future profits from $t + 1$ on by q_{t+1}. The first-order condition for capital is $q_{t+1}J_{kt+1}(n', k') = 1$. Using the envelope condition for k, $J_{kt}(n, k) = z_tF_{kt} + (1 - \delta)$ in this first-order condition gives

$$1 = q_{t+1}[z_{t+1}F_{kt+1} + (1-\delta)]. \tag{49}$$

The first-order condition for the mass of recruiters to deploy at t is

$$z_tF_{nt} = \lambda_f(\theta_t)q_{t+1}J_{nt+1}(n', k'). \tag{50}$$

Using the envelope condition for n,

$$J_{nt}(n,k) = z_tF_n - w_t + [1-\delta_n]q_{t+1}J_{nt+1}(n', k'), \tag{51}$$

in the first-order condition for recruiters (50) gives

$$J_{nt}(n,k) = z_tF_{nt} - w_t + [1-\delta_n]\frac{z_tF_n}{\lambda_f(\theta_t)}. \tag{52}$$

The value of having an additional worker employed at an arbitrary wage w in the current period, who will receive the equilibrium wage in all future periods, is

$$\tilde{J}_{nt}(n,k,w) = w_t - w + J_{nt}(n_t, k_t).$$

Wages are determined according to Nash bargaining with the bargaining parameter ϕ for the worker and $1 - \phi$ for the firm. The bargained wage w maximizes the asymmetric Nash product by solving

$$\max_w \phi \log\left[\tilde{V}_{nt}(a_t, n_t, w)\right] + (1-\phi)\log\left[\tilde{J}_{nt}(k_t, n_t, w)\right],$$

where the first term in brackets is the value to the family of having an additional worker employed rather than unemployed at an arbitrary wage w. The first-order condition is

$$\phi\frac{\tilde{V}_{nwt}}{\tilde{V}_{nt}} + (1-\phi)\frac{\tilde{J}_{nwt}}{\tilde{J}_{nt}} = 0.$$

Using $\tilde{V}_{nwt}(a_t, n_t, w) = U_{ct}$ and $\tilde{J}_{nwt}(k_t, n_t, w) = -1$ and evaluating this first-order condition at equilibrium with $w = w_t$ so that $\tilde{V}_{nt} = V_{nt}$ and $\tilde{J}_{nt} = J_{nt}$ gives

$$\phi\frac{U_{ct}}{V_{nt}} = (1-\phi)\frac{1}{J_{nt}}. \tag{53}$$

Substituting for V_{nt} and V_{nt+1} from (53) into (48) and replacing J_{nt} with the right side of (52) gives

$$\phi\left[\left(1+\frac{1-\delta_n}{\lambda_f(\theta_t)}\right)z_tF_{nt}-w_t\right]=(1-\phi)\left[w_t+\frac{U_{nt}}{U_{ct}}\right]+\phi[1-\delta_n-\lambda_w(\theta_t)]q_{t+1}J_{nt+1}.$$

Replacing J_{nt+1} using the first-order condition for recruiters (50), we can solve for the equilibrium wage,

$$w_t=\phi[1+\theta_t]z_tF_{nt}+(1-\phi)\left(-\frac{U_{nt}}{U_{ct}}\right). \tag{54}$$

Here, hiring an unemployed worker produces a marginal value to the firm that includes both the direct value of production and the savings on recruiters' time. The wage is a weighted average of this marginal value and the marginal rate of substitution between consumption and employment for the household. Substituting the wage equation into the recruiter's first-order condition (50) gives

$$z_tF_{nt}U_{ct}=\beta U_{ct+1}\lambda_{ft}\left\{z_{t+1}F_{nt+1}\left[1+\frac{1-\delta_n}{\lambda_{ft+1}}\right]-\phi[1+\theta_{t+1}]z_{t+1}F_{nt+1}+(1-\phi)\frac{U_{nt+1}}{U_{ct+1}}\right\}. \tag{55}$$

The corresponding first-order condition for recruiters for the planner (42) can be manipulated to be

$$z_tF_{nt}U_{ct}=\beta U_{ct+1}G_{1t}\left\{z_{t+1}F_{nt+1}\left[\frac{1-\delta_n}{G_{1t+1}}-\frac{G_{2t+1}}{G_{1t+1}}\right]+z_{t+1}F_{nt+1}+\frac{U_{nt+1}}{U_{ct+1}}\right\}. \tag{56}$$

With a Cobb–Douglas matching function $G(v, 1-n)=Bv^{1-\eta}(1-n)^{\eta}$, we have that $G_{1t}=(1-\eta)\lambda_{ft}$ and $G_{2t}=\eta\theta_t\lambda_{ft}$ so that (56) becomes

$$z_tF_{nt}U_{ct}=\beta U_{ct+1}\lambda_{ft}\left\{z_{t+1}F_{nt+1}\left[1+\frac{1-\delta_n}{\lambda_{ft+1}}\right]-\eta[1+\theta_{t+1}]z_{t+1}F_{nt+1}+(1-\eta)\frac{U_{nt+1}}{U_{ct+1}}\right\}. \tag{57}$$

Clearly, these first-order conditions coincide if the Mortensen–Hosios condition is satisfied in that the worker's bargaining weight ϕ equals the elasticity of the matching function with respect to unemployment η. We can decentralize the solution to the planning problem as an equilibrium with a wage of

$$w_t^p=\eta[1+\theta_t]z_tF_{nt}+(1-\eta)\left(-\frac{U_{nt}}{U_{ct}}\right). \tag{58}$$

Notice that even in an efficient equilibrium, the wage typically equals neither the marginal product of labor z_tF_{nt} nor the marginal rate of substitution $-U_{nt}/U_{ct}$. Furthermore, the marginal rate of substitution is not equal to the marginal product of labor. These considerations suggest a different notion of a wedge relative to that used in the one-sector model. To that end, write the equilibrium wage as $w_t = (1-\tau_{lt})w_t^p$ where w_t^p is the planner's wage. Hence,

$$1-\tau_{lt} = \frac{\phi[1+\theta_t]z_t F_{nt} + (1-\phi)\left(-\frac{U_{nt}}{U_{ct}}\right)}{\eta[1+\theta_t]z_t F_{nt} + (1-\eta)\left(-\frac{U_{nt}}{U_{ct}}\right)}. \tag{59}$$

Clearly, if the Mortensen–Hosios condition is satisfied, the wedge $\tau_{lt} = 0$.

1.3.1.2 The Associated Prototype Economy with Efficiency and Labor Wedges

Consider the following prototype model. In this model, the bargaining power of the worker is equal to the elasticity η, but workers have to pay a tax τ_{lt} on their wages, investment is taxed at rate τ_{xt}, and the productivity is given by $\hat{A}_t$. Next we compare the aggregate outcomes of the prototype model and the equilibrium search model. From (47) and (49), we immediately have that the Euler equation is undistorted so that the investment wedge $\tau_{xt} = 0$. Using the production function $y_t = A_t F(k_t, n_t - v_t)$, it is immediate that $\hat{A}_t = A_t$. Thus, we have the following proposition.

Proposition 5 *The aggregate allocations in the equilibrium search economy coincide with those of the prototype economy if the efficiency wedge in the prototype economy is given by $\hat{A}_t = A_t$, the labor wedge $1 - \tau_{lt}$ is given by (59), and the investment wedge is zero.*

Note that if search is efficient, the labor wedge is zero in the two-sector prototype economy.

2. THE ACCOUNTING PROCEDURE

Having established our equivalence result, we now describe our accounting procedure at a conceptual level, discuss a Markovian implementation of it, and distinguish our procedure from others.

Our procedure is designed to answer questions of the following kind: How much would output fluctuate if the only wedge that fluctuated is the efficiency wedge and the probability distribution of the efficiency wedge is the same as in the prototype economy? Critically, our procedure ensures that agents' expectations of how the efficiency wedge will evolve are the same as in the prototype economy. For each experiment, we compare the properties of the resulting equilibria to those of the prototype economy. These comparisons, together with our equivalence results, allow us to identify promising classes of detailed economies.

2.1 The Accounting Procedure at a Conceptual Level

Recall that the state s^t is the history of the underlying abstract events s_t. Suppose for now that the stochastic process $\pi_t(s^t)$ and the realizations of the state s^t in some particular episode are known. Recall that the prototype economy has one underlying (vector valued) random variable, the state s^t, which has a probability of $\pi_t(s^t)$. All of the other stochastic variables, including the four wedges—the efficiency wedge $A_t(s^t)$, the labor wedge $1 - \tau_{lt}(s^t)$, the investment wedge $1/[1 + \tau_{xt}(s^t)]$, and the government consumption wedge $g_t(s^t)$—are simply functions of this random variable. Hence, when the state s^t is known, so are the wedges.

To evaluate the effects of just the efficiency wedge, for example, we consider an economy, referred to as an *efficiency wedge alone economy*, with the same underlying state s^t and probability $\pi_t(s^t)$ and the same function $A_t(s^t)$ for the efficiency wedge as in the prototype economy, but in which the other three wedges are set to be constant functions of the state, in that $\tau_{lt}(s^t) = \overline{\tau}_l, \tau_{xt}(s^t) = \overline{\tau}_x$, and $g_t(s^t) = \bar{g}$. Note that this construction ensures that the probability distribution of the efficiency wedge in this economy is identical to that in the prototype economy.

We compute the decision rules for the efficiency wedge alone economy, denoted $y^e(s^t)$, $l^e(s^t)$, and $x^e(s^t)$. For a given initial value k_0, for any given sequence s^t, we refer to the resulting values of output, labor, and investment as the *efficiency wedge components* of output, labor, and investment.

In a similar manner, we define the *labor wedge alone economy*, the *investment wedge alone economy*, and the *government consumption wedge alone economy*, as well as economies with a combination of wedges, such as the *efficiency and labor wedge economy*.

2.2 A Markovian Implementation

So far we have described our procedure assuming that we know the stochastic process $\pi_t(s^t)$ and that we can observe the state s^t. In practice, of course, we need to either specify the stochastic process a priori or use data to estimate it, and we need to uncover the state s^t from the data. Here, we describe a set of assumptions that makes these efforts easy. Then we describe in detail the three steps involved in implementing our procedure.

We assume that the state s^t follows a Markov process $\pi(s_t|s_{t-1})$ and that the wedges in period t can be used to uniquely uncover the event s_t, in the sense that the mapping from the event s_t to the wedges $(A_t, \tau_{lt}, \tau_{xt}, g_t)$ is one to one and onto. Given this assumption, without loss of generality, let the underlying event $s_t = (s_{At}, s_{lt}, s_{xt}, s_{gt})$, and let $A_t(s^t) = s_{At}$, $\tau_{lt}(s^t) = s_{lt}$, $\tau_{xt}(s^t) = s_{xt}$, and $g_t(s^t) = s_{gt}$. Note that we have effectively assumed that agents use only past wedges to forecast future wedges and that the wedges in period t are sufficient statistics for the event in period t. This assumption is only to make our estimation easier, and it can be relaxed.

In practice, to estimate the stochastic process for the state, we first specify a vector autoregressive AR(1) process for the event $s_t = (s_{At}, s_{lt}, s_{xt}, s_{gt})$ of the form

$$s_{t+1} = P_0 + Ps_t + \varepsilon_{t+1}, \tag{60}$$

where the shock ε_t is i.i.d. over time and is distributed normally with mean zero and covariance matrix V. To ensure that our estimate of V is positive semidefinite, we estimate the lower triangular matrix Q, where $V = QQ'$. The matrix Q has no structural interpretation. (Attempting to give Q such a structural interpretation is part of the source of some of the conceptual confusion about our approach. See Christiano and Davis (2006) for one such attempt.)

The first step in our procedure is to use data on y_t, l_t, x_t, and g_t from an actual economy to estimate the parameters of the Markov process $\pi(s_t|s_{t-1})$. We can do so using a variety of methods, including the maximum likelihood procedure described later.

The second step in our procedure is to uncover the event s_t by measuring the realized wedges. We measure the government consumption wedge directly from the data as the sum of government consumption and net exports. To obtain the values of the other three wedges, we use the data and the model's decision rules. With $y_t^d, l_t^d, x_t^d, g_t^d$, and k_0^d denoting the data and $y(s_t, k_t)$, $l(s_t, k_t)$, and $x(s_t, k_t)$ denoting the decision rules of the model, the realized wedge series s_t^d solves

$$y_t^d = y(s_t^d, k_t), \quad l_t^d = l(s_t^d, k_t), \text{ and } x_t^d = x(s_t^d, k_t), \tag{61}$$

with $k_{t+1} = (1-\delta)k_t + x_t^d$, $k_0 = k_0^d$, and $g_t = g_t^d$. Note that we construct a series for the capital stock using the capital accumulation law (1), data on investment x_t, and an initial choice of capital stock k_0. In effect, we solve for the three unknown elements of the vector s_t using the three Eqs. (3)–(5) and thereby uncover the state. We use the associated values for the wedges in our experiments.

Note that the four wedges account for all of the movement in output, labor, investment, and government consumption, in that if we feed the four wedges into the three decision rules in (61) and use $g_t(s_t^d) = s_{gt}$ along with the law of motion for capital, we simply recover the original data.

Note also that, in measuring the realized wedges, the estimated stochastic process plays a role only in measuring the investment wedge. To see that the stochastic process does not play a role in measuring the efficiency and labor wedges, note that these wedges can equivalently be directly calculated from (3) and (4) without computing the equilibrium of the model. In contrast, calculating the investment wedge requires computing the equilibrium of the model because the right side of (5) has expectations over future values of consumption, the capital stock, the wedges, and so on. The equilibrium of the model depends on these expectations and, therefore, on the stochastic process driving the wedges.

The third step in our procedure is to conduct experiments to isolate the marginal effects of the wedges. To do that, we allow a subset of the wedges to fluctuate as they do in the data while the others are set to constants. To evaluate the effects of the efficiency wedge, we compute the decision rules for the efficiency wedge alone economy, denoted $y^e(s_t, k_t)$, $l^e(s_t, k_t)$, and $x^e(s_t, k_t)$, in which $A_t(s^t) = s_{At}, \tau_{lt}(s^t) = \overline{\tau}_l, \tau_{xt}(s^t) = \overline{\tau}_x$, and $g_t(s^t) = \overline{g}$. Starting from k_0^d, we then use s_t^d, the decision rules, and the capital accumulation law to compute the realized sequence of output, labor, and investment, y_t^e, l_t^e, and x_t^e, which we call the *efficiency wedge components* of output, labor, and investment. We compare these components to output, labor, and investment in the data. Other components are computed and compared similarly.

Notice that in this experiment, we computed the decision rules for an economy in which only one wedge fluctuates and the others are set to be constants in all events. The fluctuations in the one wedge are driven by fluctuations in a four-dimensional state s_t.

By distinguishing the events to which the wedges are indexed from the wedges themselves, we can separate out the direct effect and the forecasting effect of fluctuations in wedges. As a wedge fluctuates, it directly affects either budget constraints or resource constraints. Whenever a wedge is not set to a constant, the fluctuations in the underlying state that lead to the fluctuations in the wedges also affect the forecasts of that wedge as well as those of other wedges in the future. Our experiments are designed so that when we hold a particular wedge constant, we eliminate the direct effect of that wedge, but we retain the forecasting effect of the underlying state on the future evolution of the wedge. By doing so, we ensure that expectations of the fluctuating wedges are identical to those in the prototype economy.

2.3 Distinguishing Our Procedure from Others

Since this way of separating the direct and forecasting effects of wedges is critical to our procedure, here we describe an alternative procedure that might, at first, seem like the intuitive way to proceed but does not answer the question that interests us.

Consider a simple example with just two wedges, an efficiency wedge and a labor wedge, denoted $W_t = (A_t, \tau_{lt})'$. Suppose that we used our prototype model to estimate the following vector process for them of the form $W_{t+1} = PW_t + \varepsilon_{t+1}$ where $E\varepsilon_t\varepsilon_t' = V$:

$$\begin{bmatrix} A_{t+1} \\ \tau_{lt+1} \end{bmatrix} = \begin{bmatrix} P_{AA} & P_{Al} \\ P_{lA} & P_{ll} \end{bmatrix} \begin{bmatrix} A_t \\ \tau_{lt} \end{bmatrix} + \begin{bmatrix} \varepsilon_{At+1} \\ \varepsilon_{lt+1} \end{bmatrix}, \tag{62}$$

where we have suppressed the constant terms. Suppose also that we have decision rules of the form

$$y_t = y(W_t, k_t), \quad l_t = l(W_t, k_t), \text{ and } x_t = x(W_t, k_t) \tag{63}$$

and that we have recovered the realized wedge series W_t^d along with the realized innovation series ε_{t+1}^d.

Now suppose want to answer the question: How much would output fluctuate under the following three conditions? First, only the efficiency wedge fluctuates. Second, for the event, the realized sequence of the efficiency wedges coincides with that in the data. Third, the probability distribution of the efficiency wedge is the same as in the prototype economy.

A first attempt to answer this question is to simply feed a realized innovation series $\hat{\varepsilon}_{t+1} = (\varepsilon^d_{At+1}, 0)$ for the event and to simulate the resulting shocks using

$$\begin{bmatrix} \hat{A}_{t+1} \\ \hat{\tau}_{lt+1} \end{bmatrix} = \begin{bmatrix} P_{AA} & P_{Al} \\ P_{lA} & P_{ll} \end{bmatrix} \begin{bmatrix} \hat{A}_t \\ \hat{\tau}_{lt} \end{bmatrix} + \begin{bmatrix} \varepsilon^d_{At+1} \\ 0 \end{bmatrix}. \tag{64}$$

This attempt meets our first condition but does not meet our second condition if P or V has nonzero off-diagonal elements, as we show they do in the data. Indeed, with nonzero off-diagonal elements, this procedure will not even produce a simulated $\hat{A}_t$ series that agrees with A^d_t. Moreover, this attempt clearly does not meet our third condition.

For a second attempt, suppose we choose the sequence of innovations so that the first two conditions are met. That is, we choose the sequence $\{\hat{\varepsilon}_{t+1}\}$ so that, in the event, the realized value of the efficiency wedge coincides with that in the data and the labor wedge is constant at, say, its mean value $\overline{\tau}_l$. Specifically, we choose $\{\hat{\varepsilon}_{t+1}\}$ so that $(\hat{A}_t, \hat{\tau}_{lt}) = (A^d_t, \overline{\tau}_l)$ in the event. The problem with this procedure is that agents' forecasts about future efficiency wedges are different under this procedure from what they are in the prototype economy. Hence, this procedure meets our first two conditions but not our third. To see why, note that the expected value of A_{t+1} in this procedure is given from

$$E_t \begin{bmatrix} A_{t+1} \\ \tau_{lt+1} \end{bmatrix} = \begin{bmatrix} P_{AA} & P_{Al} \\ P_{lA} & P_{ll} \end{bmatrix} \begin{bmatrix} A^d_t \\ \overline{\tau}_l \end{bmatrix}$$

so that

$$E_t A_{t+1} = P_{AA} A^d_t + P_{Al} \overline{\tau}_l \text{ and } E_t \tau_{lt+1} = P_{lA} A^d_t + P_{ll} \overline{\tau}_l. \tag{65}$$

The expectation of the underlying state s_{t+1} in the prototype economy, however, is calculated from

$$E_t \begin{bmatrix} s_{At+1} \\ s_{lt+1} \end{bmatrix} = \begin{bmatrix} P_{AA} & P_{Al} \\ P_{lA} & P_{ll} \end{bmatrix} \begin{bmatrix} s^d_{At} \\ s^d_{lt} \end{bmatrix} \tag{66}$$

to be

$$E_t s_{At+1} = P_{AA} s^d_{At} + P_{Al} s^d_{lt} \text{ and } E_t s_{lt+1} = P_{lA} s^d_{At} + P_{ll} s^d_{lt}. \tag{67}$$

Since we have identified s_{At+1} with A_{t+1} and s_{lt+1} with τ_{lt+1}, then (67) gives the expectations of the efficiency wedge and the labor wedge in the prototype economy during the event. Clearly, (65) and (67) do not agree when P_{Al} is not zero, so the procedure does not meet our third condition. Note that in some preliminary notes for Chari et al. (2006), while we were aware of the flaws in the second attempt, we followed a version of this second attempt as a quick approximation to get an initial set of answers. Christiano and

Davis (2006), unfortunately did not realize that even in our NBER working paper version we followed the correct procedure. We view their paper as a valuable exposition of why the second attempt is incorrect and of the flaws that arise when one follows it.

Next we show that our procedure meets our three conditions. In the efficiency wedge alone economy, the first two conditions are clearly met: only the efficiency wedge fluctuates, and in the event the realized efficiency wedge coincides with the measured efficiency wedge in the data. To see that the third, and more subtle, condition is met, note from (60) the probability distribution over s_{t+1}, and therefore A_{t+1} is the same in both the prototype economy and the efficiency wedge alone economy.

3. APPLYING THE ACCOUNTING PROCEDURE

Now we demonstrate how to apply our accounting procedure to the Great Recession and postwar data for the United States and a group of other OECD countries. (In Appendix, we describe in detail our data sources, parameter choices, computational methods, and estimation procedures.)

3.1 Details of the Application

To apply our accounting procedure, we use functional forms and parameter values that are familiar from the business cycle literature. We assume that the production function has the form $F(k, l) = k^{\alpha}l^{1-\alpha}$ and the utility function the form $U(c,l) = \log c + \psi \log(1-l)$. We choose the capital share α to be one-third and the time allocation parameter $\psi = 2.5$. We choose the depreciation rate δ, the discount factor β, and growth rates γ and γ_n so that, on an annualized basis, depreciation is 5%, the rate of time preference 2.5%, and the population growth rate and the growth of technology are country-specific and computed using OECD data. The adjustment cost parameter $b = \delta + \gamma + \gamma_n$ is pinned down by the previous parameters and varies across countries. For the adjustment cost parameter a, we follow Bernanke et al. (1999) in choosing this parameter so that the elasticity, η, of the price of capital with respect to the investment–capital ratio is 0.25. In this setup, the price of capital $q = 1/(1-\phi')$, so that, evaluated at the steady state, $\eta = ab$. Given η and b, we then set a accordingly.

Our prototype economy is a closed economy. When confronting the data, we let government consumption in the model correspond to the sum of government consumption and net exports in the data. The rationale for this choice is given in Chari et al. (2005), where we prove an equivalence result between an open economy model and a closed economy model in which government consumption is treated in this fashion. We then use a standard maximum likelihood procedure to estimate the parameters P_0, P, and V of the vector AR(1) process for the wedges. In doing so, we use the log-linear decision rules of the prototype economy and data on output, labor, investment, and the sum of government consumption and net exports.

In confronting the theory with the data, we need to decide how to treat consumer durables and sales taxes. At a conceptual level, we think of current expenditures on consumer durables as augmenting the stock of consumer durables, which in turn provides a service flow of consumption to consumers. Based on this idea, we reallocate current expenditures of consumer durables from consumption to investment. We then add the imputed service flow from the stock of consumer durables to consumption and output. This imputed service flow is the rental rate on capital times the stock of durables. We assume that the stock of consumer durables depreciates at the same rate as the stock of physical capital. We also adjust the data to account for sales taxes. We assume that sales taxes are levied solely on consumption. This assumption leads us to subtract sales tax revenues from both consumption and measured output.

At a practical level, it turns out that while the U.S. NIPA accounts have quarterly data on consumer durable expenditures for the 1980:1–2014:4 sample we use, the OECD has more limited data. For some of the countries in our sample, data are only available annually or are missing. For countries for which we only have annual data, we fill in quarterly estimates using maximum likelihood estimates of a state space model. For countries for which we only have quarterly data for a subsample, we regress consumer durables on investment and output and use the coefficients to construct estimates of the missing data. Once we have the quarterly series on consumer durables, we construct estimates of the capital stock using the perpetual inventory method. The service flow of durables is assumed to be 4% of the stock of durables. (For details, see the Appendix.)

We express all variables in per capita form and deflate by the GDP deflator. We then estimate separate sets of parameters for the stochastic process for wedges (60) for each of the OECD countries after removing country-specific trends in output, investment, and government consumption. The other parameters are the same across countries. The stochastic process parameters for the Great Recession are estimated using quarterly data for 1980:1–2014:4. The stochastic process (60) with these values is used by agents in our economy to form their expectations about future wedges. In Appendix, we give the details of the estimated values of the stochastic processes for each of the countries.

3.2 Findings

Now we describe the results of applying our procedure to OECD countries for the Great Recession and the 1982 Recession. Here, we focus primarily on the fluctuations due to the efficiency, labor, and investment wedges.[a]

3.2.1 The Great Recession

Here, we discuss our findings for the 24 OECD countries. The main finding is that in terms of accounting for the downturn, in the United States the labor wedge is by far the

[a] We alert the reader that the quantitative results for Spain should be treated with caution. In some robustness analysis for Spain, we found that the nonlinear labor wedge computed directly from the consumer's first-order condition (4) and found that the nonlinear labor wedge moved substantially more than the labor wedge computed using our log-linearization procedure.

most important, in Spain, Ireland, and Iceland the investment wedge is the most important, and in the rest of the countries, the efficiency wedge is the most important.

3.2.1.1 Three Illustrative Recessions

Here, we illustrate our findings for one country for which the efficiency wedge, labor wedge, and investment wedge, respectively, is the most important. In reporting our findings, we remove a country-specific trend from output, investment, and the government consumption wedge. Both output and labor are normalized to equal 100 in the base period 2008:1. Here, we focus primarily on the fluctuations due to the efficiency, labor, and investment wedges. We discuss the government consumption wedge and its components in Appendix.

3.2.1.1.1 France: Primarily an Efficiency Wedge Recession We begin with France. In Fig. 1A, we see that from 2008:1 to 2009:3, output fell about 7% while labor fell about 3% and investment fell about 18%. In Fig. 1B, we see that the efficiency wedge worsened by about 5%, the labor wedge worsened by about 1%, and the investment wedge worsened by about 5%. In Fig. 1C, we see that the efficiency wedge accounts for the bulk of the decline in output, namely about 6% of the 7% decline. Fig. 1D–E show that the efficiency and investment wedges play the most important roles in accounting for the declines in labor and investment.

Overall, these results imply that the Great Recession in France should be thought of as primarily an efficiency wedge recession with some role for the labor and investment wedges in accounting for the decline in hours and investment. This finding implies that models that emphasize fluctuations in the labor wedge in France are less promising than those that emphasize fluctuations in the efficiency and investment wedges.

3.2.1.1.2 United States: Primarily a Labor Wedge Recession Next consider the United States. In Fig. 2A, we see that output and labor both fell about 7% from 2008:1 to 2009:3 while investment fell about 23%. In Fig. 2B, we see that the efficiency wedge fell very modestly by only about 1%, while the labor wedge and the investment wedge both worsened dramatically, by about 8% and 9%, respectively. In Fig. 2C–E, we see that the labor and investment wedges play the most important role in accounting for the downturn in output and labor, while the investment wedge accounts for the bulk of the downturn in investment.

Overall, considering the period from 2008 until the end of 2011, these results imply that the Great Recession in the United States should be thought of as primarily a labor wedge recession, with an important secondary role for the investment wedge. This finding implies that the most promising models must yield significant fluctuations in the labor wedge, with some role for the investment wedge. Models that emphasize the efficiency wedge are less promising.[b]

[b] In the Appendix, we show that if we estimated the stochastic process for the wedges from 1948 to 2015, the contribution of the labor wedge rises and that of the investment wedge falls. A similar change occurs if we decrease the investment adjustment cost parameter.

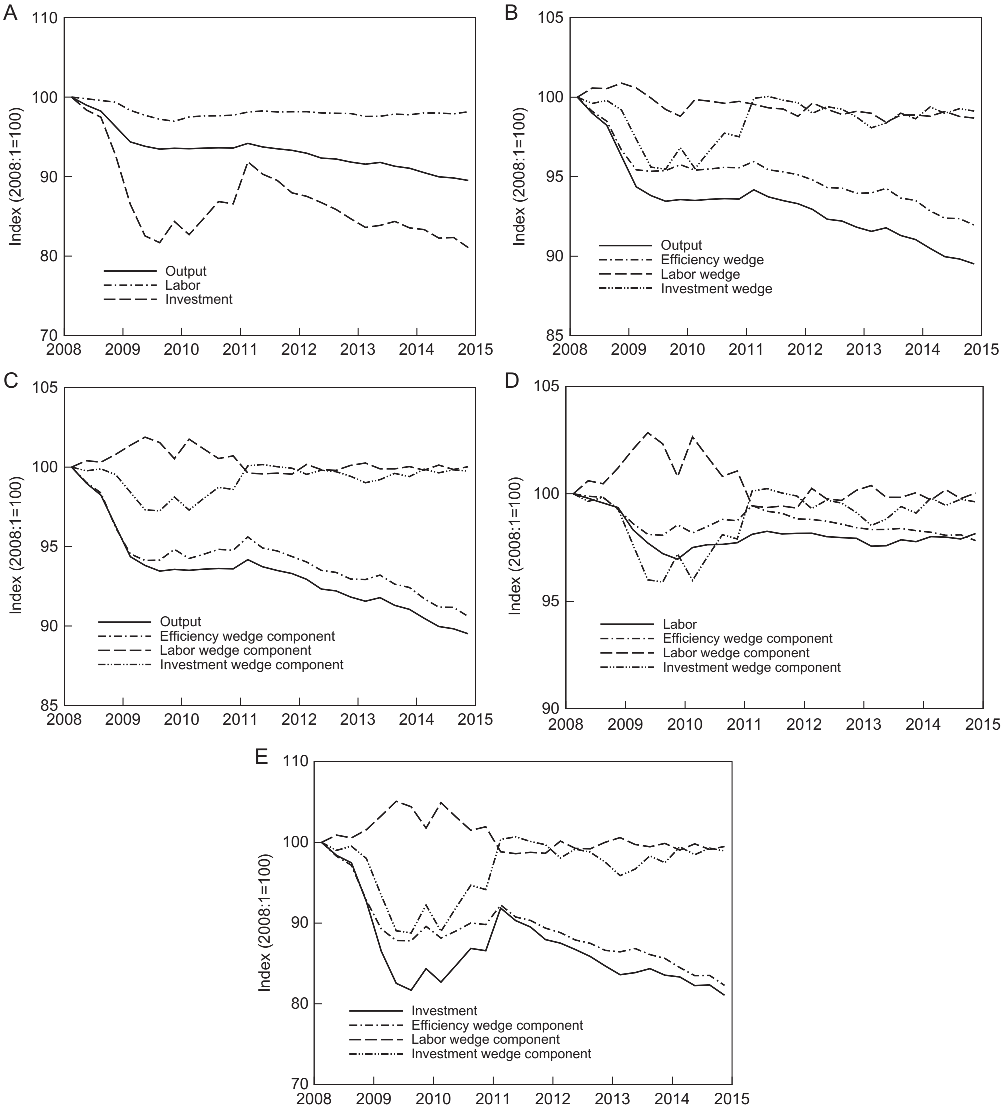


Fig. 1 (A) Output, labor, and investment for France, 2008:1–2014:4. (B) Output and three wedges for France, 2008:1–2014:4. (C) Output and output components for France, 2008:1–2014:4. (D) Labor and labor components for France, 2008:1–2014:4. (E) Investment and investment components for France, 2008:1–2014:4.

3.2.1.1.3 Ireland: Primarily an Investment Wedge Recession Finally consider Ireland. In Fig. 3A, we see that from 2008:1 to 2009:3, output fell about 13%, labor about 11%, and investment almost 50%. Fig. 3B shows that during this period, the efficiency wedge fell about 5%, the labor wedge worsened by about 10%, and the investment wedge worsened dramatically, that is, by about 20%.

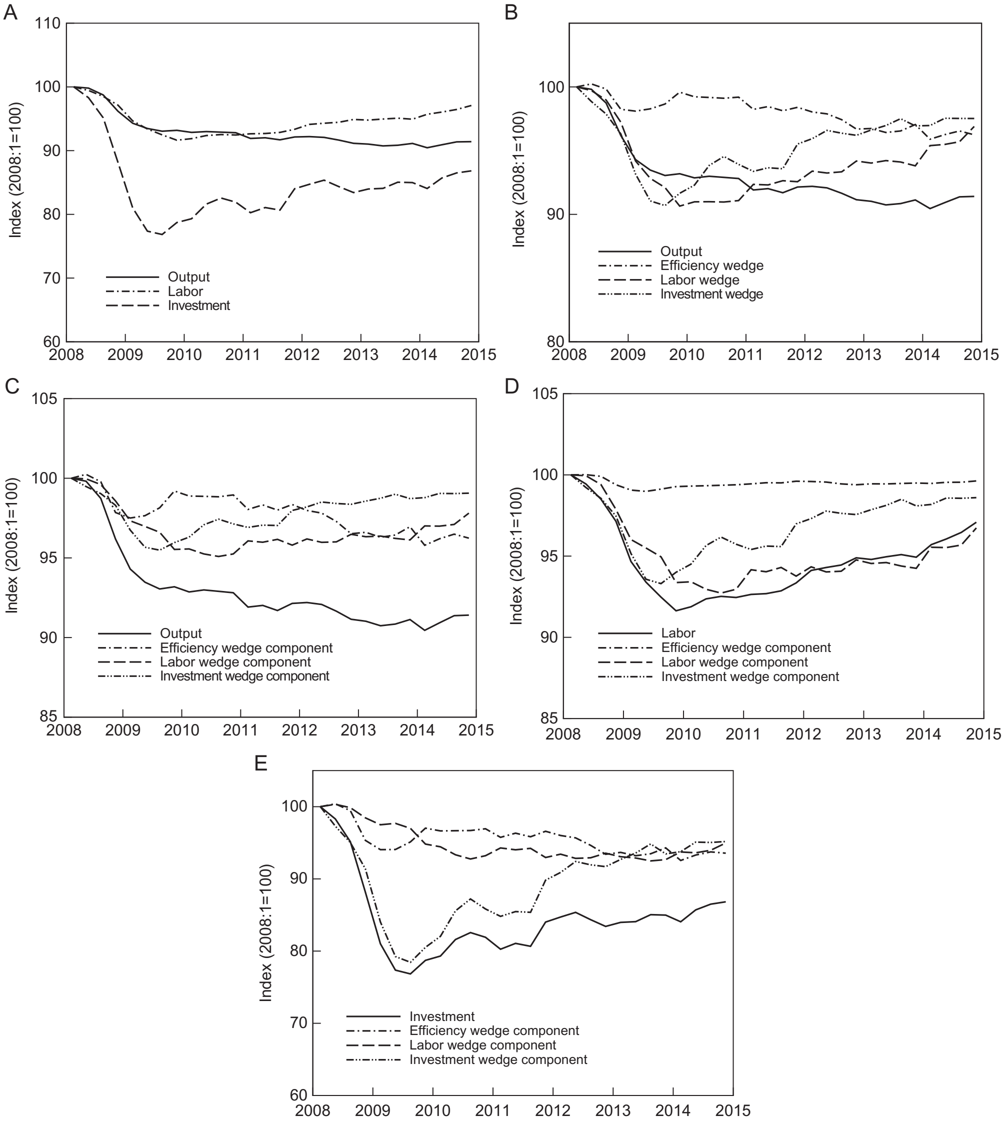


Fig. 2 (A) Output, labor, and investment for the United States, 2008:1–2014:4. (B) Output and three wedges for the United States, 2008:1–2014:4. (C) Output and output components for the United States, 2008:1–2014:4. (D) Labor and labor components for the United States, 2008:1–2014:4. (E) Investment and investment components for the United States, 2008:1–2014:4.

In Fig. 3C–E, we see that the investment wedge plays the largest role: it accounts for about half of the fall in output, about four-fifths of the fall in investment, and all of the fall in hours. Overall, these results imply that the Great Recession in Ireland should be thought of as primarily an investment wedge recession.

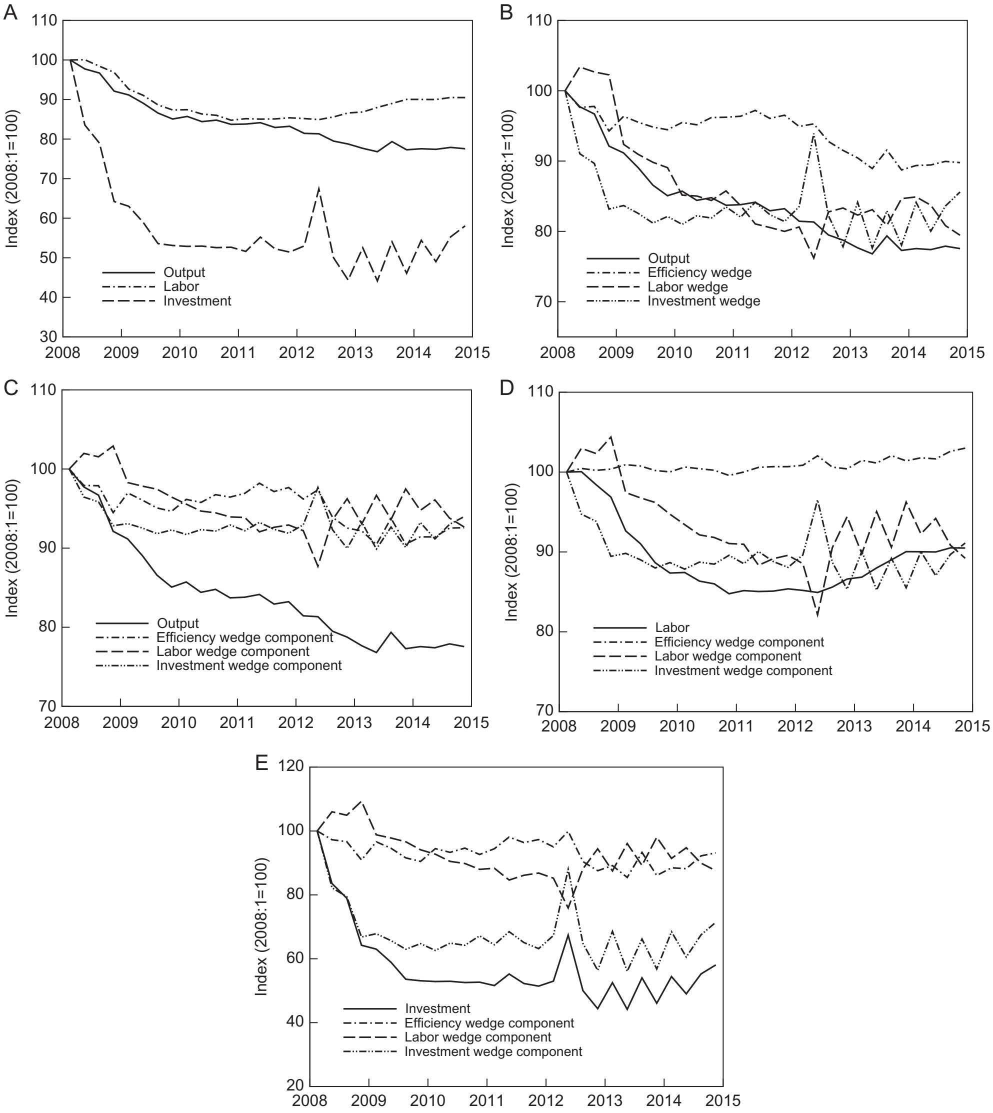


Fig. 3 (A) Output, labor, and investment for Ireland, 2008:1–2014:4. (B) Output and three wedges for Ireland, 2008:1–2014:4. (C) Output and output components for Ireland, 2008:1–2014:4. (D) Labor and labor components for Ireland, 2008:1–2014:4. (E) Investment and investment components for Ireland, 2008:1–2014:4.

3.2.1.2 Summary Statistics for Our OECD Countries

So far we have described the Great Recession in three countries. Here, we describe useful summary statistics over the period 2008:1–2011:3. One such statistic, referred to as the ϕ*statistic*, is intended to capture how closely a particular component, say, the output component due to the efficiency wedge, tracks the underlying variable, say, output. For our decomposition of output, we let

$$\phi_i^Y = \frac{1/\sum_t (y_t - y_{it})^2}{\sum_j \sum_t \left(1/(y_t - y_{jt})^2\right)},$$

where y_{it} is the output component due to wedge $i = (A, \tau_l, \tau_x, g)$. We compute similar statistics for labor and investment. The ϕ statistic has the desirable feature that it lies in [0, 1], sums to one across the four wedges, and when a particular output component tracks output perfectly, in that if $(y_t - y_{it}) = 0$ for all t, then $\phi_i^Y = 1$, that is, the ϕ statistic for the wedge reaches its maximum value of 1. Note that this statistic is the inverse of the mean-square error for each wedge appropriately scaled so that the sum across wedges adds to one.

Now consider our main finding. In Fig. 4A, we display the ϕ statistic for the efficiency wedge and labor wedge components of output. The downward-sloping lines represent

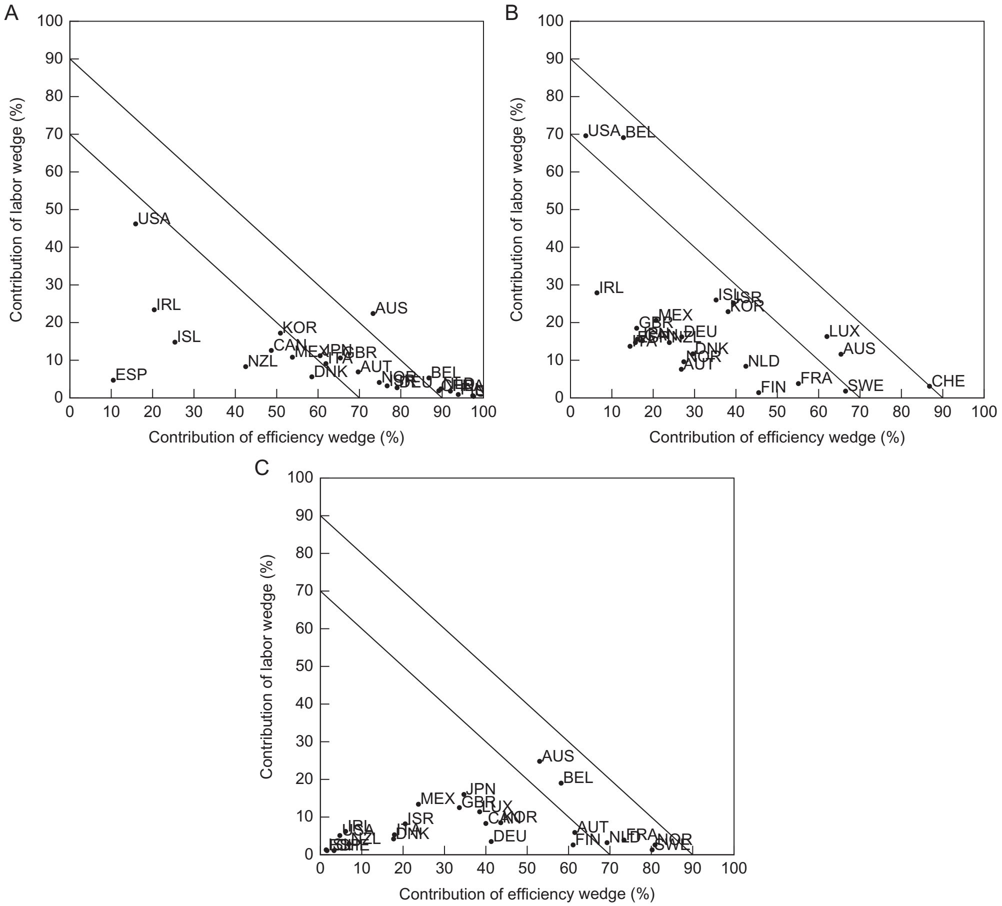


Fig. 4 (A) Decomposition of output, 2008:1–2011:3. (B) Decomposition of labor, 2008:1–2011:3. (C) Decomposition of investment, 2008:1–2011:3.

combinations for which the sum of the labor wedge and efficiency wedge components is constant at 70% and 90%, respectively. This figure shows that the United States stands out from the other countries in that the labor wedge accounts for a much greater fraction of the movements in output than it does in any other country. Specifically, the labor wedge accounts for about 46% of the movements in output in the United States but no more than 22% in any other country. In all other countries except Iceland, Ireland, New Zealand, and Spain, the efficiency wedge accounts for roughly 50% or more of the movements in output. In Table 1, we report the decompositions of output, labor, and investment for all countries. There we see that for Iceland, Ireland, New Zealand, and Spain, the investment wedge accounts for 51%, 48%, 42%, and 82% of the movements in output, respectively. In the other panels of Fig. 4, we display the ϕ statistics for the components of labor and investment.

Our main finding is also apparent if we use other ways to measure how important a given wedge is for the movements in output, labor and investment. When we discussed

Table 1 ϕ-Statistics for output, labor, and investment components, Great Recession

	Output components			Labor components			Investment components		
Countries	ϕ_A^Y	$\phi_{\tau_l}^Y$	$\phi_{\tau_x}^Y$	ϕ_A^L	$\phi_{\tau_l}^L$	$\phi_{\tau_x}^L$	ϕ_A^X	$\phi_{\tau_l}^X$	$\phi_{\tau_x}^X$
Australia	0.73	0.22	0.02	0.65	0.12	0.13	0.53	0.25	0.04
Austria	0.70	0.07	0.11	0.27	0.08	0.19	0.61	0.06	0.21
Belgium	0.87	0.05	0.05	0.13	0.69	0.14	0.58	0.19	0.15
Canada	0.49	0.13	0.18	0.17	0.15	0.32	0.40	0.08	0.47
Denmark	0.58	0.06	0.30	0.30	0.12	0.47	0.18	0.04	0.72
Finland	0.94	0.01	0.03	0.46	0.01	0.07	0.61	0.03	0.30
France	0.92	0.02	0.04	0.55	0.04	0.30	0.73	0.04	0.17
Germany	0.79	0.03	0.12	0.27	0.16	0.33	0.41	0.04	0.50
Iceland	0.25	0.15	0.51	0.35	0.26	0.27	0.01	0.01	0.95
Ireland	0.20	0.23	0.48	0.06	0.28	0.62	0.06	0.06	0.82
Israel	0.77	0.03	0.16	0.39	0.25	0.08	0.20	0.08	0.60
Italy	0.62	0.09	0.22	0.14	0.14	0.64	0.18	0.05	0.74
Japan	0.60	0.11	0.15	0.16	0.16	0.45	0.35	0.16	0.32
Korea	0.51	0.17	0.18	0.38	0.23	0.16	0.44	0.09	0.34
Luxembourg	0.97	0.01	0.01	0.62	0.16	0.15	0.39	0.11	0.07
Mexico	0.54	0.11	0.28	0.21	0.21	0.49	0.24	0.13	0.51
Netherlands	0.90	0.02	0.05	0.42	0.08	0.25	0.69	0.03	0.24
New Zealand	0.42	0.08	0.42	0.24	0.15	0.51	0.07	0.03	0.86
Norway	0.75	0.04	0.05	0.27	0.10	0.23	0.81	0.03	0.05
Spain	0.11	0.05	0.82	0.16	0.15	0.62	0.02	0.01	0.96
Sweden	0.98	0.00	0.01	0.67	0.02	0.17	0.80	0.01	0.17
Switzerland	0.89	0.02	0.07	0.87	0.03	0.03	0.03	0.01	0.94
United Kingdom	0.65	0.11	0.15	0.16	0.19	0.55	0.34	0.13	0.42
United States	0.16	0.46	0.32	0.04	0.70	0.25	0.05	0.05	0.88
Average	0.64	0.09	0.20	0.33	0.19	0.31	0.36	0.07	0.48

Table 2A Peak to trough declines in output and components, Great Recession

		Changes in output and its components			
Countries	**Trough**	ΔY	ΔY_A	ΔY_{τ_l}	ΔY_{τ_x}
Australia	2011:1	−5.6	−5.6	−2.0	0.7
Austria	2010:1	−9.2	−6.2	3.3	−4.7
Belgium	2010:1	−7.4	−5.8	−2.3	−0.1
Canada	2009:3	−6.5	−3.2	0.0	−2.1
Denmark	2009:4	−9.9	−6.9	1.4	−5.3
Finland	2010:1	−14.1	−12.5	4.2	−3.3
France	2009:3	−6.5	−5.9	1.5	−2.8
Germany	2009:2	−8.6	−7.2	2.3	−3.5
Iceland	2011:1	−14.3	−4.6	2.2	−15.5
Ireland	2009:4	−14.9	−5.3	−3.6	−7.7
Israel	2009:2	−4.8	−3.3	−1.6	−0.8
Italy	2010:1	−10.5	−6.7	−0.7	−3.5
Japan	2009:1	−10.0	−8.3	−0.4	0.4
Korea	2009:2	−7.4	−6.1	4.5	−5.6
Luxembourg	2009:4	−15.6	−16.5	1.2	5.9
Mexico	2009:2	−5.4	−4.7	0.5	−2.0
Netherlands	2010:3	−8.5	−7.4	1.2	−2.3
New Zealand	2010:4	−7.6	−5.3	−0.2	−2.2
Norway	2011:2	−11.9	−8.8	1.1	0.5
Spain	2013:4	−19.7	−9.2	−0.6	−10.8
Sweden	2009:4	−10.5	−9.5	2.9	−2.7
Switzerland	2009:2	−5.7	−5.7	3.3	−4.8
United Kingdom	2012:2	−14.8	−10.3	0.1	−2.9
United States	2009:3	−7.0	−1.9	−3.4	−4.5
Average		−9.9	−7.0	0.6	−3.3

Note: The date of the peak is 2008:1 for all countries.

the three illustrative recessions earlier, we compared simple peak-to-trough measures of output, labor, and investment to the corresponding measures for each of the components. In Tables 2A, 2B, and 2C, we report such measures for all of our countries. A quick perusal of these measures shows that they give the same message as the ϕ statistics do. Consider, for example, France. The ϕ statistic indicates that the efficiency wedge accounts for the bulk of the movements in output, namely about 92% of its decline. The peak-to-trough measure indicates that the efficiency wedge also accounts for the bulk of the peak-to-trough decline, namely about 5.9% of the 6.5% decline or about 91% of the decline.

3.2.2 Comparing the Great Recession with Recessions of the Early 1980s

The postwar era had essentially two periods during which most developed economies experienced recessions at roughly the same time: the early 1980s and the Great Recession

Table 2B Peak to trough declines in labor and components, Great Recession

		Changes in labor and its components			
Countries	**Trough**	ΔL	ΔL_A	ΔL_{τ_l}	ΔL_{τ_x}
Australia	2011:1	−0.5	−1.0	−3.1	1.1
Austria	2010:1	−4.9	−1.3	5.0	−7.0
Belgium	2010:1	−3.2	−0.8	−3.4	−0.1
Canada	2009:3	−5.7	−0.7	0.0	−3.1
Denmark	2009:4	−5.3	−1.1	2.2	−7.9
Finland	2010:1	−2.9	−1.3	6.3	−4.9
France	2009:3	−2.8	−1.9	2.3	−4.1
Germany	2009:2	−3.6	−2.0	3.5	−5.2
Iceland	2011:1	−9.1	1.0	3.4	−22.4
Ireland	2009:4	−12.6	0.0	−5.3	−11.3
Israel	2009:2	−1.6	0.1	−2.3	−1.3
Italy	2010:1	−5.2	−0.5	−1.0	−5.1
Japan	2009:1	−3.4	−1.2	−0.6	0.6
Korea	2009:2	−2.9	−1.7	6.8	−8.3
Luxembourg	2009:4	3.7	0.0	1.7	9.0
Mexico	2009:2	−2.5	−1.1	0.7	−3.0
Netherlands	2010:3	−1.1	−0.5	1.9	−3.4
New Zealand	2010:4	−3.3	−1.2	−0.3	−3.3
Norway	2011:2	−3.3	1.0	1.7	0.8
Spain	2013:4	−14.8	−3.7	−0.8	−15.7
Sweden	2009:4	−3.2	−2.0	4.3	−4.1
Switzerland	2009:2	−1.2	−1.3	5.1	−7.1
United Kingdom	2012:2	−3.8	−1.0	0.1	−4.2
United States	2009:3	−7.5	−0.9	−5.0	−6.7
Average		−4.2	−1.0	1.0	−4.9

Note: The date of the peak is 2008:1 for all countries.

of 2008. Here, we compare the recessions of the early 1980s with the Great Recession. For the United States, we use the NBER business cycle dates; for the OECD countries, we use the business cycle dates as estimated by ECRI when available and otherwise use the CEPR Euro Area Business Cycle Dates. We use the stochastic process for wedges estimated over the 1980–2014 period for both episodes. (See the Appendix for details.)

In Fig. 5A, we compare the ϕ statistics for the efficiency wedge component of output for the two recessions. This panel shows that for most of the countries, the efficiency wedge in the Great Recession played a more important role than it did during the recessions of the 1980s. In Fig. 5B, we compare the ϕ statistics for the labor wedge component of output for the two recessions. This panel shows that in the Great Recession, the labor wedge accounts for over 40% of the fluctuations in output only in the United States, while in the 1982 recession, it does so only in Belgium, the United Kingdom, and France.

Table 2C Peak to trough declines in investment and components, Great Recession

		Changes in investment and its components			
Countries	**Trough**	ΔX	ΔX_A	ΔX_{τ_l}	ΔX_{τ_x}
Australia	2011:1	−13.0	−9.8	−3.5	3.1
Austria	2010:1	−19.6	−10.2	8.4	−16.2
Belgium	2010:1	−21.8	−11.9	−10.1	−0.3
Canada	2009:3	−13.9	−7.0	−0.1	−9.7
Denmark	2009:4	−33.1	−14.7	5.0	−23.8
Finland	2010:1	−23.9	−19.7	8.2	−12.6
France	2009:3	−18.3	−12.2	4.4	−11.2
Germany	2009:2	−19.9	−14.2	4.6	−14.5
Iceland	2011:1	−56.6	−4.6	6.5	−55.0
Ireland	2009:4	−46.9	−9.5	−5.9	−35.3
Israel	2009:2	−14.9	−5.6	−4.8	−4.1
Italy	2010:1	−18.4	−9.6	−2.2	−12.7
Japan	2009:1	−15.4	−13.1	−2.2	1.7
Korea	2009:2	−23.2	−9.8	9.0	−20.5
Luxembourg	2009:4	−13.2	−28.2	−2.7	30.2
Mexico	2009:2	−18.3	−8.7	−0.4	−9.7
Netherlands	2010:3	−16.6	−13.4	4.2	−10.0
New Zealand	2010:4	−16.7	−9.8	1.5	−9.9
Norway	2011:2	−16.6	−14.4	4.5	2.0
Spain	2013:4	−47.9	−17.8	2.2	−38.9
Sweden	2009:4	−21.8	−21.4	8.4	−12.7
Switzerland	2009:2	−18.7	−10.4	8.6	−21.8
United Kingdom	2012:2	−28.0	−17.5	−1.5	−12.7
United States	2009:3	−23.2	−4.9	−3.0	−21.6
Average		−23.3	−12.4	1.6	−13.2

Note: The date of the peak is 2008:1 for all countries.

In Fig. 5C, we compare the ϕ statistics for the investment wedge component of output for the two recessions. This panel shows that in most of the countries, the investment wedge played a larger role in the recessions of the 1980s than it did during the Great Recession.

In Table 3, we report the ϕ statistics for the 1982 recessions. This table shows that the efficiency wedge played the most important role for ten countries, the labor wedge for three countries, and the investment wedge for seven countries. Together with Table 1, this table broadly reinforces our two main findings for the comparison. First, the labor wedge played an important role for output in the Great Recession only for the United States, and in the 1982 recession it played a dominant role only in Belgium, France, and the United Kingdom. Second, for most countries, in the Great Recession the efficiency wedge played a more important role and the investment wedge played a less important role than they did in the recessions of the 1980s.

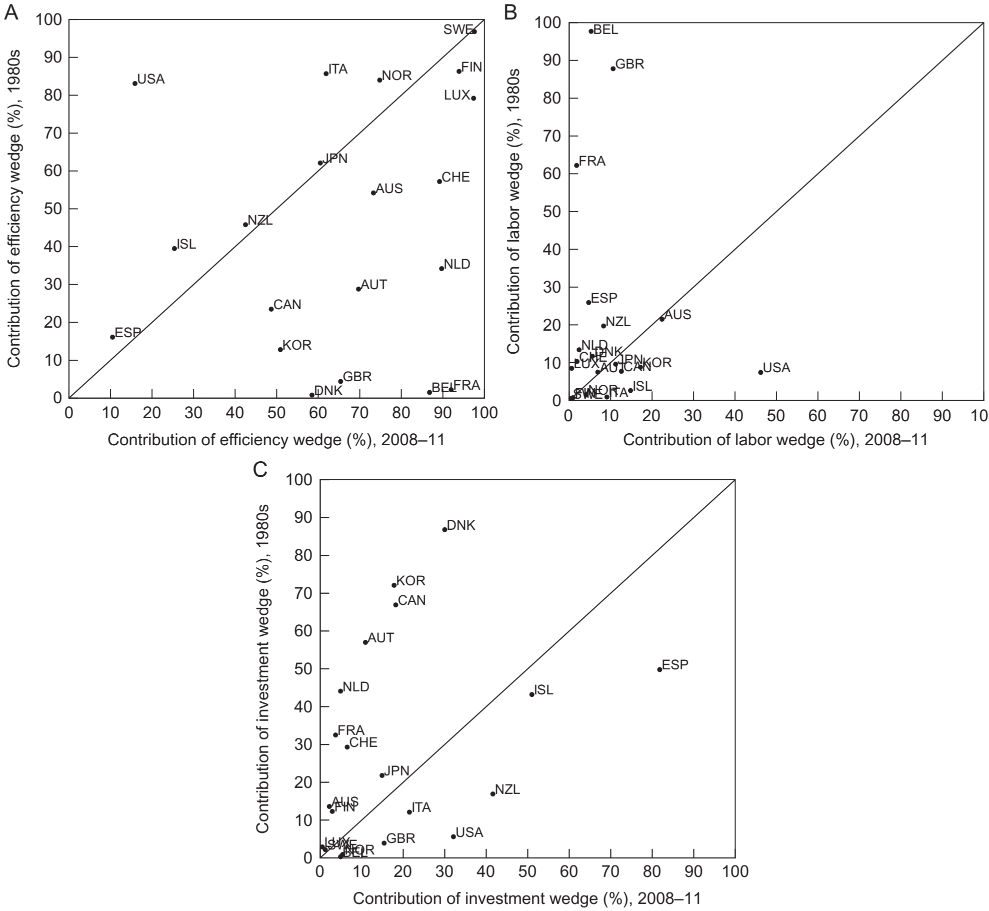


Fig. 5 (A) Efficiency component of output for two recessions. (B) Labor component of output for two recessions. (C) Investment component of output for two recessions.

In Tables 4A–4C, we report peak-trough results for the 1982 recession. Comparing Table 3 with the Tables 4A–4C, we see that the peak-trough results present the same overall picture as our ϕ statistics do. If we compare the classification of the most important wedge for each country using ϕ statistics for output to that using the peak-trough decline for output, we see that they agree in all but three cases.

3.2.3 Summary Statistics for the Entire Period

In Tables 5A–5C, we present some summary statistics for the entire period 1980:1–2014:3 about the importance of the various wedges in accounting for the movements in output, labor, and investment. In Table 5A, for example, we report the standard deviation of the output component due to each wedge relative to the standard deviation

Table 3 ϕ-Statistics for output, labor, and investment components, 1982 Recession

	Output components			Labor components			Investment components		
Countries	ϕ_A^Y	$\phi_{\tau_l}^Y$	$\phi_{\tau_x}^Y$	ϕ_A^L	$\phi_{\tau_l}^L$	$\phi_{\tau_x}^L$	ϕ_A^X	$\phi_{\tau_l}^X$	$\phi_{\tau_x}^X$
Australia	0.54	0.22	0.14	0.22	0.39	0.23	0.43	0.18	0.24
Austria	0.29	0.07	0.57	0.19	0.13	0.59	0.04	0.02	0.91
Belgium	0.01	0.98	0.00	0.09	0.82	0.03	0.04	0.91	0.01
Canada	0.23	0.08	0.67	0.11	0.07	0.82	0.13	0.04	0.80
Denmark	0.01	0.12	0.87	0.02	0.28	0.68	0.01	0.02	0.96
Finland	0.86	0.01	0.12	0.87	0.06	0.02	0.04	0.01	0.94
France	0.02	0.62	0.33	0.07	0.63	0.25	0.04	0.10	0.82
Iceland	0.40	0.03	0.43	0.41	0.11	0.07	0.13	0.03	0.77
Italy	0.86	0.01	0.12	0.97	0.01	0.01	0.10	0.02	0.85
Japan	0.62	0.10	0.22	0.15	0.15	0.62	0.29	0.13	0.47
Korea	0.13	0.09	0.72	0.09	0.12	0.72	0.02	0.02	0.94
Luxembourg	0.79	0.09	0.03	0.05	0.72	0.01	0.10	0.73	0.02
Netherlands	0.34	0.13	0.44	0.13	0.28	0.50	0.03	0.01	0.94
New Zealand	0.46	0.20	0.17	0.22	0.47	0.09	0.17	0.06	0.61
Norway	0.84	0.01	0.01	0.66	0.23	0.07	0.11	0.34	0.10
Spain	0.16	0.26	0.50	0.12	0.28	0.54	0.04	0.04	0.90
Sweden	0.97	0.01	0.02	0.85	0.04	0.06	0.52	0.05	0.34
Switzerland	0.57	0.10	0.29	0.22	0.59	0.16	0.04	0.03	0.92
United Kingdom	0.04	0.88	0.04	0.06	0.85	0.05	0.17	0.49	0.15
United States	0.83	0.07	0.06	0.21	0.54	0.19	0.64	0.14	0.15
Average	0.42	0.22	0.29	0.28	0.33	0.30	0.17	0.16	0.59

of output during entire period, along with the correlation of each such output component with output. In Tables 5B and 5C, we report similar statistics for labor and its components and for investment and its components.

Using these statistics to infer the importance of various wedges is more subtle than using the ϕ statistics. The ϕ statistic captures in one statistic how much the component due to a wedge moves, as well as how closely this component tracks the underlying variable. Instead, to evaluate the importance of a wedge using the statistics in this table, we need to jointly consider the relative standard deviations and the correlations.

Consider, for example, France. Viewing the relative standard deviations alone suggests that the labor and investment wedges play roughly the same role in accounting for the movement in output. Indeed, the relative standard deviations of the labor and investment components of output are 93% and 92%, respectively. But the correlations of these variables with output suggest that the investment wedge plays a much more important role. Indeed, the labor component of output comoves negatively with output, whereas the investment component of output comoves positively with output.

With this perspective in mind, the averages across countries show that the efficiency wedge plays the most important role in accounting for output. The standard deviation of

Table 4A Peak to trough declines in output and components, 1982 Recession

			Changes in output and its components			
Countries	**Peak**	**Trough**	ΔY	ΔY_A	ΔY_{τ_l}	ΔY_{τ_x}
Australia	1981:3	1983:2	−10.4	−5.9	−1.4	−3.6
Austria	1980:1	1983:1	−7.2	−2.0	0.5	−6.4
Belgium	1980:1	1983:2	−8.6	−3.6	−7.9	2.4
Canada	1981:2	1982:4	−8.7	−5.1	−1.0	−6.5
Denmark	1980:1	1981:2	−5.4	0.4	−2.7	−4.8
Finland	1980:3	1984:2	−8.3	−7.0	0.9	−5.4
France	1982:1	1984:4	−4.4	1.5	−3.5	−2.5
Iceland	1980:1	1983:4	−10.5	−13.2	7.3	−5.5
Italy	1980:2	1983:2	−9.2	−8.3	6.1	−9.2
Japan	1991:2	1995:1	−5.8	−3.7	−0.9	−1.9
Korea	1997:3	1998:3	−11.5	−3.4	−2.2	−7.1
Luxembourg	1980:1	1983:1	−13.2	−9.7	−3.7	3.4
Netherlands	1980:1	1982:3	−11.2	−5.2	−3.0	−3.9
New Zealand	1981:3	1983:1	−5.1	−3.3	−0.9	−1.9
Norway	1980:1	1982:3	−7.7	−6.7	0.3	3.6
Spain	1980:1	1984:2	−13.9	0.3	−5.9	−10.8
Sweden	1980:1	1983:1	−6.3	−6.2	1.4	−2.3
Switzerland	1981:3	1982:4	−6.6	−6.2	−0.3	−2.6
United Kingdom	1980:1	1982:2	−8.7	−1.1	−8.9	1.7
United States	1980:1	1982:4	−9.1	−6.8	−1.3	−1.6
Average			−8.1	−4.2	−1.7	−3.2

Table 4B Peak to trough declines in labor and components, 1982 recession

			Changes in labor and its components			
Countries	**Peak**	**Trough**	ΔL	ΔL_A	ΔL_{τ_l}	ΔL_{τ_x}
Australia	1981:3	1983:2	−7.7	−1.1	−2.1	−5.3
Austria	1980:1	1983:1	−6.3	−0.1	0.7	−9.4
Belgium	1980:1	1983:2	−8.9	−1.4	−11.6	3.7
Canada	1981:2	1982:4	−8.4	−3.6	−1.6	−9.7
Denmark	1980:1	1981:2	−8.0	0.2	−4.0	−7.1
Finland	1980:3	1984:2	−1.2	0.3	1.3	−8.0
France	1982:1	1984:4	−7.3	−0.6	−5.2	−3.7
Iceland	1980:1	1983:4	3.9	−1.0	11.2	−8.2
Italy	1980:2	1983:2	−2.7	−2.9	9.4	−13.4
Japan	1991:2	1995:1	−4.8	−0.8	−1.3	−2.9
Korea	1997:3	1998:3	−11.8	−0.1	−3.2	−10.4
Luxembourg	1980:1	1983:1	−5.1	−0.3	−5.5	5.1
Netherlands	1980:1	1982:3	−7.0	1.0	−4.5	−5.8
New Zealand	1981:3	1983:1	−3.9	−0.6	−1.4	−2.8
Norway	1980:1	1982:3	−0.5	1.3	0.4	5.5
Spain	1980:1	1984:2	−15.8	1.3	−8.7	−15.7
Sweden	1980:1	1983:1	−0.4	−0.6	2.0	−3.4
Switzerland	1981:3	1982:4	−1.5	−1.1	−0.5	−3.9
United Kingdom	1980:1	1982:2	−9.5	−0.1	−13.0	2.5
United States	1980:1	1982:4	−4.2	−1.0	−2.0	−2.4
Average			−5.7	−0.6	−2.4	−4.6

Table 4C Peak to trough declines in investment and components, 1982 Recession

			Changes in investment and its components			
Countries	**Peak**	**Trough**	ΔX	ΔX_A	ΔX_{τ_l}	ΔX_{τ_x}
Australia	1981:3	1983:2	−25.1	−10.3	−2.1	−14.3
Austria	1980:1	1983:1	−21.0	−2.9	2.8	−21.5
Belgium	1980:1	1983:2	−29.9	−9.2	−25.4	12.8
Canada	1981:2	1982:4	−35.4	−15.5	0.3	−27.7
Denmark	1980:1	1981:2	−25.6	1.2	−4.4	−21.7
Finland	1980:3	1984:2	−23.3	−9.8	4.6	−20.0
France	1982:1	1984:4	−14.2	1.4	−5.3	−10.3
Iceland	1980:1	1983:4	−25.7	−20.7	14.9	−23.6
Italy	1980:2	1983:2	−32.2	−15.1	11.5	−30.9
Japan	1991:2	1995:1	−17.1	−6.4	−2.6	−8.2
Korea	1997:3	1998:3	−31.1	−3.9	−1.3	−25.2
Luxembourg	1980:1	1983:1	−9.5	−17.6	−7.9	16.6
Netherlands	1980:1	1982:3	−23.2	−7.3	−2.9	−16.8
New Zealand	1981:3	1983:1	−14.4	−6.0	−0.4	−8.4
Norway	1980:1	1982:3	−1.2	−10.4	1.2	15.4
Spain	1980:1	1984:2	−39.5	3.0	−8.0	−38.9
Sweden	1980:1	1983:1	−20.8	−13.2	4.7	−10.8
Switzerland	1981:3	1982:4	−13.2	−10.6	1.0	−12.5
United Kingdom	1980:1	1982:2	−10.9	−2.0	−17.8	8.1
United States	1980:1	1982:4	−20.2	−12.2	−3.2	−8.1
Average			−20.4	−7.4	−2.7	−11.8

the efficiency component of output is 92% of output, and its correlation with output is 0.77. Even though the labor component of output is 89% as variable as output itself, it is essentially uncorrelated with output. In this sense, the labor wedge does not account for much of the movements in output.

3.2.4 The Importance of the Classification of Consumer Durables

Macroeconomists have long argued that theory implies it is appropriate to treat the expenditures on consumer durables as a form of investment that yields a flow of consumption services. This treatment requires adjustments to the national income account classification of consumption and investment to make them consistent with the theory.

Here, we show that while this adjustment is quantitatively important for some countries, for most countries it does not change the overall findings. In Fig. 6A, we contrast the ϕ statistic for the efficiency wedge component of output when this consistent adjustment is made and when it is not. Clearly, the countries with statistics most affected by this adjustment are Iceland and Spain. In Iceland, for example, the contribution of the efficiency wedge falls from 26% when durables are correctly accounted for to 12% when

Table 5A Properties of the output components, entire sample

	Standard deviations			Correlations		
Countries	σ_{Y_A}/σ_Y	$\sigma_{Y_{\tau_l}}/\sigma_Y$	$\sigma_{Y_{\tau_x}}/\sigma_Y$	$\rho_{Y_A,Y}$	$\rho_{Y_{\tau_l},Y}$	$\rho_{Y_{\tau_x},Y}$
Australia	0.92	0.94	0.85	0.67	−0.10	0.71
Austria	1.06	0.98	1.05	0.82	−0.32	0.37
Belgium	0.77	1.00	0.44	0.72	0.68	−0.34
Canada	0.67	0.42	0.63	0.89	−0.03	0.79
Denmark	1.18	0.95	0.89	0.58	−0.15	0.72
Finland	0.74	0.72	0.89	0.80	−0.33	0.71
France	1.11	0.93	0.92	0.88	−0.45	0.64
Germany	0.74	0.34	0.61	0.87	0.02	0.69
Iceland	0.97	1.19	1.44	0.75	−0.15	0.27
Ireland	0.84	0.92	0.92	0.62	−0.02	0.53
Israel	0.83	0.58	0.59	0.92	0.08	0.40
Italy	0.99	1.03	1.39	0.85	−0.32	0.51
Japan	0.97	0.48	0.46	0.85	0.01	0.35
Korea	1.04	0.99	0.90	0.69	−0.12	0.58
Luxembourg	1.14	1.01	1.14	0.95	−0.18	−0.20
Mexico	0.97	0.69	0.68	0.91	0.15	0.21
Netherlands	0.99	0.87	1.06	0.72	−0.27	0.50
New Zealand	1.06	0.83	0.88	0.66	−0.14	0.58
Norway	1.08	2.15	1.35	0.71	−0.21	0.24
Spain	0.72	1.15	1.29	0.34	0.35	0.35
Sweden	0.93	0.53	0.40	0.93	−0.28	0.84
Switzerland	1.13	1.15	1.32	0.90	−0.25	0.35
United Kingdom	0.73	0.85	0.55	0.61	0.50	0.43
United States	0.60	0.58	0.61	0.76	0.64	0.74
Average	0.92	0.89	0.89	0.77	−0.04	0.46

Notes: The entire sample is 1980:1–2014:4. Series are first logged and detrended with the filter of Hodrick and Prescott (1997).

they are not. In Spain, the contribution of the efficiency wedge increases from 11% when durables are correctly accounted for to 29% when they are not.

In Fig. 6B and C, we contrast the analogous ϕ statistics for the labor wedge component of output and for the investment wedge component of output. In panel C, we see that in Iceland and Spain, the contribution of the investment wedge to output is 51% and 82% when durables are correctly accounted for and 65% and 35% when they are not.

3.2.5 Comparing Our Procedure with a Perfect Foresight Procedure

Some authors implement a perfect foresight version of our procedure in which agents have perfect foresight about the future evolution of the wedges. The equilibrium conditions for the deterministic version of our prototype model are

Table 5B Properties of the labor components, entire sample

	Standard deviations			Correlations		
Countries	σ_{L_A}/σ_L	$\sigma_{L_{\tau_l}}/\sigma_L$	$\sigma_{L_{\tau_x}}/\sigma_L$	$\rho_{L_A,L}$	$\rho_{L_{\tau_l},L}$	$\rho_{L_{\tau_x},L}$
Australia	0.27	1.20	1.08	0.39	0.42	0.50
Austria	0.28	1.77	1.90	−0.14	0.36	0.20
Belgium	0.26	1.40	0.61	0.36	0.95	−0.50
Canada	0.39	0.66	0.99	0.75	0.36	0.82
Denmark	0.23	1.10	1.03	−0.44	0.73	0.53
Finland	0.16	1.25	1.56	0.19	0.05	0.61
France	0.63	1.90	1.87	0.25	0.20	0.38
Germany	0.27	0.63	1.13	0.40	0.31	0.78
Iceland	0.22	2.05	2.47	−0.33	0.29	0.37
Ireland	0.21	1.23	1.24	0.30	0.53	0.39
Israel	0.09	1.69	1.74	−0.88	0.38	0.33
Italy	0.55	2.15	2.90	0.07	0.15	0.29
Japan	0.49	1.06	1.02	−0.05	0.46	0.51
Korea	0.45	1.48	1.35	−0.28	0.49	0.34
Luxembourg	0.46	3.22	3.63	−0.18	0.39	0.08
Mexico	0.38	1.64	1.62	0.17	0.39	0.29
Netherlands	0.39	1.45	1.76	−0.35	0.39	0.41
New Zealand	0.28	1.16	1.23	−0.43	0.47	0.55
Norway	0.58	3.49	2.20	−0.13	0.31	0.25
Spain	0.31	1.19	1.33	0.10	0.49	0.42
Sweden	0.75	0.93	0.69	0.83	0.16	0.70
Switzerland	0.38	2.62	3.00	−0.03	0.30	0.13
United Kingdom	0.12	1.16	0.75	−0.27	0.81	0.29
United States	0.14	0.84	0.89	0.64	0.83	0.75
Average	0.35	1.55	1.58	0.04	0.43	0.39

Notes: The entire sample is 1980:1–2014:4. Series are first logged and detrended with the filter of Hodrick and Prescott (1997).

$$c_t + x_t + g_t = y_t, \tag{68}$$

$$y_t = A_t F\left(k_t, (1+\gamma)^t l_t\right), \tag{69}$$

$$-\frac{U_{lt}}{U_{ct}} = [1-\tau_{lt}]A_t(1+\gamma)^t F_{lt}, \text{ and} \tag{70}$$

$$U_{ct}[1+\tau_{xt}] = \beta U_{ct+1}\{A_{t+1}F_{kt+1} + (1-\delta)[1+\tau_{xt+1}]\}. \tag{71}$$

Clearly, the efficiency wedge, the labor wedge, and the government consumption wedge can be recovered from the static relationships in (68), (69), and (70). Recovering the investment wedge, however, requires solving the difference equation implied by the Euler equation (71). To do so, we need to impose either an initial condition or a terminal condition. In practice, we imposed an initial condition that the investment wedge begins at zero.

Table 5C Properties of the investment components, entire sample

	Standard deviations			Correlations		
Countries	σ_{X_A}/σ_X	$\sigma_{X_{\tau_l}}/\sigma_X$	$\sigma_{X_{\tau_x}}/\sigma_X$	$\rho_{X_A,X}$	$\rho_{X_{\tau_l},X}$	$\rho_{X_{\tau_x},X}$
Australia	0.38	0.38	0.77	0.78	−0.31	0.87
Austria	0.62	0.71	1.35	0.53	−0.71	0.89
Belgium	0.39	0.76	0.47	0.84	0.91	−0.69
Canada	0.43	0.16	0.75	0.89	−0.28	0.97
Denmark	0.54	0.42	0.86	0.44	−0.32	0.97
Finland	0.34	0.39	0.95	0.73	−0.66	0.98
France	0.63	0.58	0.97	0.90	−0.72	0.91
Germany	0.53	0.22	0.93	0.58	−0.12	0.96
Iceland	0.29	0.40	1.12	−0.17	−0.36	0.93
Ireland	0.34	0.40	0.92	0.49	−0.36	0.95
Israel	0.39	0.33	0.79	0.69	−0.03	0.83
Italy	0.47	0.48	1.37	0.54	−0.73	0.90
Japan	0.70	0.37	0.74	0.65	−0.01	0.71
Korea	0.56	0.50	1.01	0.57	−0.59	0.93
Luxembourg	0.58	0.58	1.33	0.23	−0.92	0.87
Mexico	0.50	0.36	0.92	0.67	−0.12	0.72
Netherlands	0.60	0.54	1.30	0.20	−0.70	0.96
New Zealand	0.51	0.40	0.96	0.36	−0.47	0.94
Norway	0.48	0.88	1.05	−0.06	0.20	0.44
Spain	0.38	0.49	1.24	0.13	−0.36	0.90
Sweden	0.74	0.32	0.51	0.94	−0.36	0.97
Switzerland	0.35	0.41	1.10	0.27	−0.81	0.99
United Kingdom	0.39	0.50	0.73	0.42	0.23	0.84
United States	0.35	0.29	0.92	0.79	0.15	0.94
Average	0.48	0.45	0.96	0.52	−0.31	0.82

Notes: The entire sample is 1980:1–2014:4. Series are first logged and detrended with the filter of Hodrick and Prescott (1997).

In Fig. 7A–C, we plot the ϕ statistics for the perfect foresight procedure against the same statistics for our procedure. These panels show that for a significant number of the countries, the ϕ statistics are very different. In particular, the perfect foresight procedure greatly exaggerates the importance of the labor wedge for the United States and Spain. Under perfect foresight, the labor wedge accounts for 92% and 72% of the movements in output for the United States and Spain, while under the standard business cycle accounting procedure, the labor wedge accounts for only 46% and 5%, respectively.

We highlight two important sources for these differences. One is that in the perfect foresight procedure, private agents anticipate the evolution of future wedges perfectly and thus react in the current period to actual future worsening or improvement of the wedges. In this sense, the perfect foresight procedure brings with it all the undesirable properties of the simple "news" models by which an anticipated worsening of, say, the

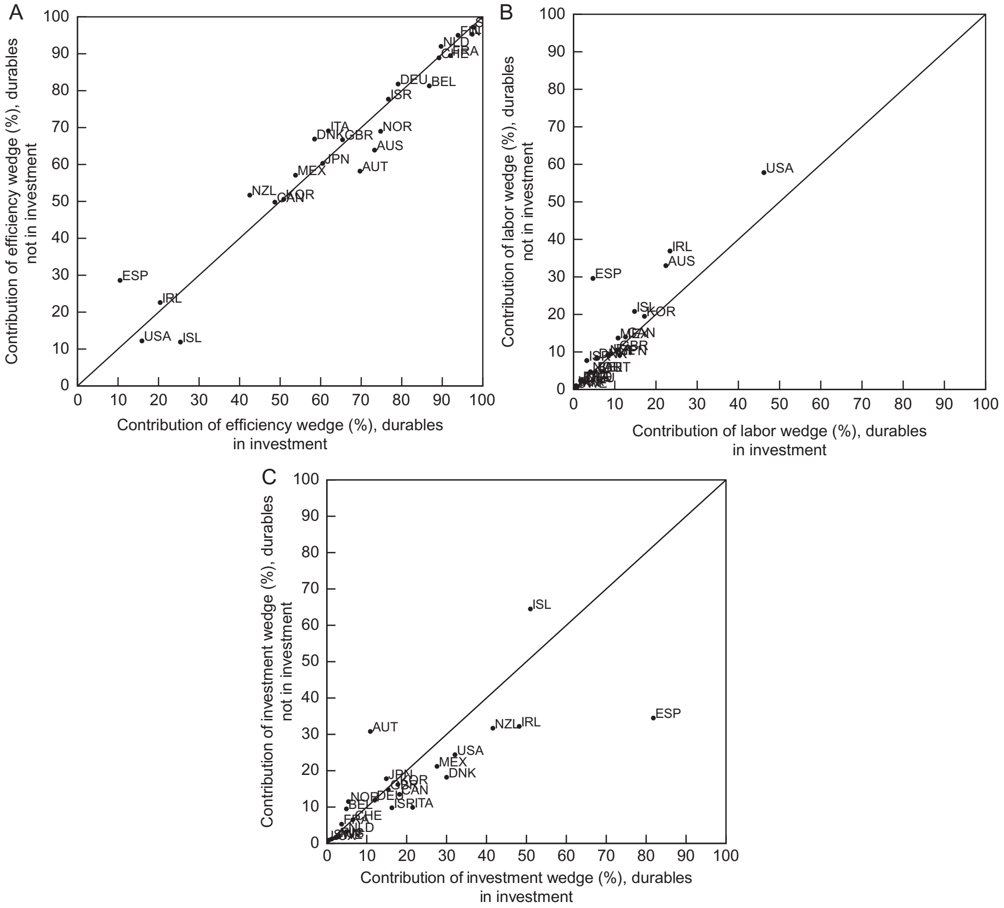


Fig. 6 (A) Efficiency component of output for two investment measures. (B) Labor component of output for two investment measures. (C) Investment component of output for two investment measures.

labor wedge leads to a current boom as households choose to increase labor supply before times worsen. The other is that, as we noted earlier, the perfect foresight procedure uses the nonlinear version of the first-order conditions (68)–(71) to compute the wedges while our procedure uses log-linearized versions of these conditions.

4. CONCLUSION

We have elaborated on the business cycle accounting method proposed by CKM, cleared up some misconceptions about the method, and applied it to compare the Great Recession across OECD countries as well as to the recessions of the 1980s in these countries.

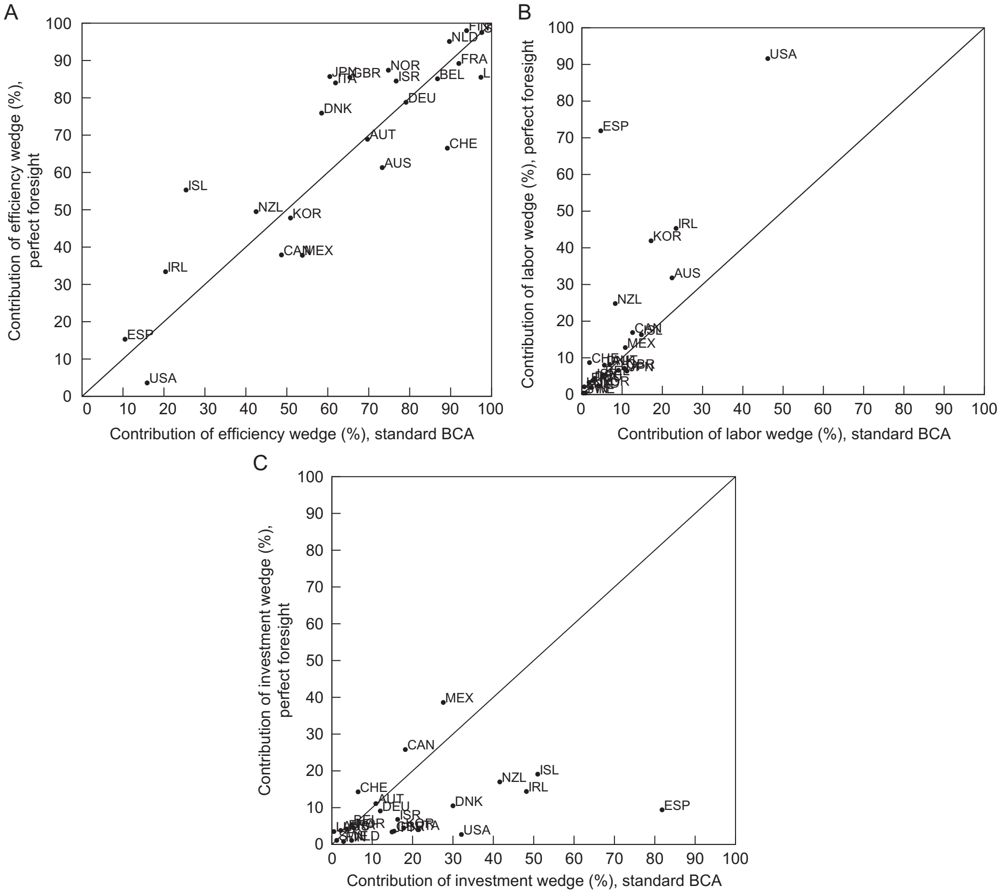


Fig. 7 (A) Efficiency component of output for two expectational assumptions. (B) Labor component of output for two expectational assumptions. (C) Investment component of output for two expectational assumptions.

We documented four findings. First, with the notable exception of the United States, Spain, Ireland, and Iceland, the Great Recession was driven primarily by the efficiency wedge. Second, in the Great Recession, the labor wedge plays a dominant role only in the United States, and the investment wedge plays a dominant role in Spain, Ireland, and Iceland. Third, in the recessions of the 1980s, the labor wedge played a dominant role only in France, the United Kingdom, and Belgium. Finally, overall in the Great Recession, the efficiency wedge played a much more important role and the investment wedge played a much less important role than they did in the recessions of the 1980s.

APPENDIX

A.1 Data and Sources

The data used for the business cycle accounting exercises throughout the chapter come mainly from OECD (variable codes in parenthesis). The time span is from 1980 to the end of 2014 and, unless mentioned otherwise, at the quarterly frequency. For some countries (such as Germany, Ireland, Israel, and Mexico), data for most series were only available starting later than 1980Q1 and thus the business cycle accounting exercises were performed for shorter samples.

- Economic Outlook 98
  - Gross domestic product, value, market prices (GDP)
  - GDP deflator, market prices (PGDP)
  - Gross capital formation, current prices (ITISK)
  - Government final consumption expenditures, value, expenditure approach (CG)
  - Exports of goods and services, value, national accounts basis (XGS)
  - Imports of goods and services, value, national accounts basis (MGS)
  - Hours worked per employee, total economy (HRS)
  - Total employment (ET)
- System of Quarterly National Accounts
  - Durable goods (subcategory of CQRsa: private final consumption expenditure by durability, national currency, current prices)
- Tax on goods and services
  - Taxes on goods and services as a share of GDP, annual (TAXGOODSERV, PCGDP)
- Population and Labor Force
  - Population 15–64, persons, annual

All data are deflated by the GDP deflator. Data on durables are available for different time spans and frequency. When data were available at a quarterly frequency, the series of durables were computed by regressing durables on a constant, gross capital formation (ITISK) and gross domestic product (GDP) in logs, for the available time span, and then using the coefficient estimates to compute the series for durables from the beginning of sample. When data on durables were only available at the annual frequency, quarterly observations were estimated using maximum likelihood estimates of a state space model and, as before, series on gross capital formation and gross domestic product. Once we get durables at the quarterly frequency, we extend the series to the beginning of sample by the method described above. Population data are available at annual frequency and thus is interpolated to quarterly frequency using cubic splines. All other transformations are standard and described in detail below:

- per capita output (y): real GDP $-$ sales taxes $+$ services from consumer durables (with annualized return at 4% $+$ depreciation of durables at an annualized rate of 25%) deflated by the GDP deflator and divided by population 16–64.
- per capita hours (h): hours worked*total employment, divided by population 16–64.

- per capita investment (x): gross capital formation + personal consumption expenditures on durables net of sales taxes, all deflated by the GDP deflator and divided by population 16–64.
- per capital government consumption (g): government final consumption expenditures + Exports of goods and services − Imports of goods and services, all deflated by the GDP deflator and divided by population 16–64.

A.2 Parametrization and Calibration

Table A.1 Parameters held fix across countries

β	δ	ψ	σ	θ
0.975	0.05	2.5	1	0.33

where β is the (annualized) and δ the (annualized) depreciation rate of capital.

Other parameters are specific to each country and shown in the Table A.2 below, where γ_n is the average growth rate of population, γ the growth rate of labor augmenting technology and a the adjustment costs coefficient. To compute γ, we set it so that detrended log output is mean zero over the sample period.

Table A.2 Parameters that are specific to each country

Country	γ_n	γ	a
Australia	0.014	0.022	11.550
Austria	0.005	0.023	12.602
Belgium	0.003	0.021	13.348
Canada	0.011	0.017	13.308
Denmark	0.003	0.021	13.515
Finland	0.002	0.031	11.956
France	0.005	0.018	13.563
Germany	−0.001	0.021	14.159
Ireland	0.014	0.047	9.370
Iceland	0.012	0.025	11.320
Israel	0.020	0.023	10.740
Italy	0.002	0.018	14.206
Japan	−0.001	0.021	14.189
Korea	0.013	0.054	8.600
Luxembourg	0.013	0.037	9.896
Mexico	0.018	0.007	13.223
Netherlands	0.005	0.024	12.539
Norway	0.008	0.024	12.106
New Zealand	0.012	0.018	12.963
Spain	0.007	0.024	12.177
Sweden	0.004	0.022	13.078
Switzerland	0.009	0.014	13.600
United Kingdom	0.003	0.025	12.745
United States	0.010	0.019	12.574

Note also that there are other parameters which are country specific, namely the elements of the P_0, P and Q matrices that result from the maximum likelihood estimation procedure that models expectations. These estimates are available at http://pedrobrinca.pt/2016-accounting-for-business-cycles/

A.3 Replication Instructions

Replication files are available at http://pedrobrinca.pt/2016-accounting-for-business-cycles/. We also make available an extensive Appendix that includes all the tables and figures in the chapter and country reports which include additional tables and figures regarding each of the business cycle accounting exercises performed, for both the Great Recession period and the recessions in the 1980s. The Appendix also includes the elements of the P_0, P, and Q matrices that result from the maximum likelihood estimation procedure that models expectations for each country.

ACKNOWLEDGMENTS

We thank Peter Klenow for useful comments and Francesca Loria for excellent research assistance. V.V.C. and P.J.K. thank the National Science Foundation for financial support. P.B. is grateful for financial support from the Portuguese Science and Technology Foundation, grants number SFRH/BPD/99758/2014, UID/ECO/00124/2013, and UID/ECO/04105/2013. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

REFERENCES

Bernanke, B., Gertler, M., Gilchrist, S., 1999. The financial accelerator in a quantitative business cycle framework. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics, vol. 1C. North-Holland, Amsterdam, pp. 1341–1393.

Brinca, P., 2013. Monetary business cycle accounting for Sweden. BE J. Macroecon. 13 (1), 1085–1119.

Brinca, P., 2014. Distortions in the neoclassical growth model: a cross-country analysis. J. Econ. Dyn. Contr. 47, 1–19.

Buera, F.J., Moll, B., 2015. Aggregate implications of a credit crunch. Am. Econ. J. Macroecon. 7 (3), 1–42.

Cavalcanti, T.V., 2007. Business cycle and level accounting: the case of Portugal. Portug. Econ. J. 6 (1), 47–64.

Chakraborty, S., 2006. Business cycle accounting of the Indian economy. Ind. Econ. J. 54 (2), 117.

Chakraborty, S., Otsu, K., 2013. Business cycle accounting of the BRIC economies. BE J. Macroecon. 13 (1), 381–413.

Chari, V.V., Kehoe, P.J., McGrattan, E.R., 2005. Sudden stops and output drops. Am. Econ. Rev. 95 (2), 381–387.

Chari, V., Kehoe, P., McGrattan, E., 2006. Business cycle accounting. Econometrica 75 (3), 781–836.

Cho, D., Doblas-Madrid, A., 2013. Business cycle accounting East and West: Asian finance and the investment wedge. Rev. Econ. Dyn. 16 (4), 724–744.

Christiano, L., Davis, J., 2006. Two flaws in business cycle accounting. NBER Working Paper 12647.

Gertler, M., Karadi, P., 2011. A model of unconventional monetary policy. J. Monet. Econ. 58 (1), 17–34.

Gertler, M., Kiyotaki, N., 2009. Financial intermediation and credit policy in business cycle analysis. In: Friedman, B.M., Woodford, M. (Eds.), Handbook of Monetary Economics. North-Holland, Amsterdam, pp. 547–599.

Gertler, M., Kiyotaki, N., Queralto, A., 2012. Financial crises, bank risk exposure and government financial policy. J. Monet. Econ. 59, S17–S34.

Greenwood, J., Hercowitz, Z., Krusell, P., 1997. Long-run implications of investment-specific technical change. Am. Econ. Rev. 87 (3), 342–362.

Hodrick, R.J., Prescott, E.C., 1997. Postwar U.S. business cycles: an empirical investigation. J. Money Credit Bank. 29 (1), 1–16.

Kersting, E.K., 2008. The 1980s recession in the UK: a business cycle accounting perspective. Rev. Econ. Dyn. 11 (1), 179–191.

Kiyotaki, N., Moore, J., 1997. Credit cycles. J. Polit. Econ. 105, 211–248.

Kobayashi, K., Inaba, M., 2006. Business cycle accounting for the Japanese economy. Jap. World Econ. 18 (4), 418–440.

Lopez, J.R., Garcia, M.S., 2014. Accounting for Spanish business cycles. Macroecon. Dyn. 1–30.

Mulligan, C., 2009. What caused the recession of 2008? Hints from labor productivity.

Ohanian, L., 2010. The economic crisis from a neoclassical perspective. J. Econ. Perspect. 24 (4), 45–66.

Ohanian, L., Raffo, A., 2012. Aggregate hours worked in OECD countries: new measurement and implications for business cycles. J. Monet. Econ. 59 (1), 40–56.

Otsu, K., 2010. A neoclassical analysis of the Asian crisis: business cycle accounting for a small open economy. BE J. Macroecon. 10 (1), 1–39.

Sustek, R., 2011. Monetary business cycle accounting. Rev. Econ. Dyn. 14 (4), 592–612.

CHAPTER 14

Incomplete Information in Macroeconomics: Accommodating Frictions in Coordination

G.-M. Angeletos[*,†], **C. Lian**[*]
[*]MIT, Cambridge, MA, United States
[†]NBER, Cambridge, MA, United States

Contents

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.04.011

Abstract

This chapter studies how incomplete information helps accommodate frictions in coordination, leading to novel insights on the joint determination of expectations and macroeconomic outcomes. We review and synthesize recent work on global games, beauty contests, and their applications. We elaborate on the distinct effects of strategic uncertainty relative to fundamental uncertainty. We demonstrate the potential fragility of workhorse macroeconomic models to relaxations of common knowledge; the possibility of operationalizing the notions of "coordination failure" and "animal spirits" in a manner that unifies unique- and multiple-equilibrium models; and the ability of incomplete information to offer a parsimonious explanation of important empirical regularities. We provide a general treatment of these ideas, as well as specific applications in the context of business cycles, financial crises, and asset pricing.

Keywords

Informational frictions, Higher-order beliefs, Strategic uncertainty, Coordination failure, Animal spirits, Aggregate demand, Business cycles, Financial crises, Global games, Beauty contests

JEL Classification Codes:

C7, D8, E1, E3, E4, G1

1. INTRODUCTION

Modern economies are vastly complex networks, in which the decisions of any given agent are highly dependent on expectations of the decisions of many other agents. For instance, how much a firm wants to hire and invest depends on the firm's expectations of the future demand for its product, which in turn depends on the spending plans of the consumers who are likely to buy the product of that firm; but consumer plans themselves depend on expectations of income and labor-market conditions, which in turn depend on the decisions of other firms and other consumers, and so on. Markets, the government, and other social institutions help facilitate the coordination of these expectations, and of the associated decisions, but only up to a point.

The limits to such coordination are rarely acknowledged in macroeconomics. Workhorse models are constructed as if all economic agents could confer with one another, reach a unanimous consensus on the current state and future prospects of the economy, and effortlessly coordinate on a commonly known profile of actions, or allocation, at all times.[a] Importantly, this is not just about theory: such models form the basis for both the interpretation of the data and the guiding of policy.

How would the predictions of these models differ if there were nontrivial frictions in the agents' ability to reach such a consensus and to coordinate their actions? What are useful ways for formalizing and operationalizing such frictions in the first place? And how does the accommodation of such frictions affect our interpretation of the observed phenomena and the guidance we may offer to policy makers?

This chapter reviews and synthesizes recent advances made in addressing these questions. We argue that incomplete information offers a useful method for introducing frictions in coordination and for enriching the dynamics of expectations in macroeconomic models. This enrichment leads to the questioning of existing interpretations of, and helps shed new light on, important phenomena such as business cycles and crises.

[a] This allocation need not be a socially desirable one; it only has to be consistent with equilibrium.

1.1 Background and Key Ideas

Before reviewing the contents of this chapter, we need to clarify the benchmark from which we depart, and the nature of the departure.

It is useful to start with the textbook RBC model. This model, like other workhorse models in macroeconomics and finance, is based on the Arrow–Debreu paradigm. In this paradigm, agents attain perfect communication: the general equilibrium is akin to the Nash equilibrium of a game of complete information, in which the players attain common knowledge, not only about exogenous fundamentals such as preferences and technology, but also about endogenous outcomes such as employment, output, and asset prices.[b]

While modern macroeconomic models often depart from the Arrow–Debreu paradigm by allowing for monopolistic power, sticky prices, credit constraints, search frictions, and so forth, they typically impose that all agents share the same information. This assumption, together with standard equilibrium concepts, tends to trivialize the problem of forecasting the actions of other agents and the associated macroeconomic outcomes: in equilibrium, all agents share the same "state of mind" about where the economy is heading to. It is in this sense that perfect coordination of beliefs and actions is imposed.

This property is evident in unique-equilibrium DSGE models that pin down equilibrium outcomes as functions of commonly known fundamentals such as preferences and technologies. But it also applies to certain multiple-equilibrium models that aimed to capture the idea that recessions, bank runs, and liquidity crises are the product of "coordination failures," such as Diamond (1982), Bryant (1983), Cooper and John (1988), Diamond and Dybvig (1983), Calvo (1988), and Obstfeld (1986, 1996).

Let us explain. The models in the aforementioned papers boil down to static coordination games that admit multiple equilibria. Some of these equilibria can be interpreted as coordination failures because they are welfare-inferior to other equilibria. Nevertheless, the following property holds in *any* of the equilibria of these models: the agents know, not only which strategy profile has been selected, but also the particular actions played by all other agents. This represents a certain oxymoron: although a coordination failure obtains *across* equilibria, coordination remains flawless *along* any given equilibrium.[c]

[b] Whenever we say "fundamentals," the game theorist can read "payoff types;" and whenever we say "endogenous outcomes" or "economic activity," the game theorist can read "actions."

[c] These points are exact in static coordination models, such as those studied in the aforementioned works, but become fussy in dynamic sunspot models. See the remark at the end of Section 3.4; see also Aumann et al. (1988) and Peck and Shell (1991) on the relation between sunspot equilibria in dynamic macroeconomic models and correlated equilibria in games.

The oxymoron gets worse: once we allow the coordination that obtains along any given equilibrium to be imperfect, in the manner that we formalize in the sequel, the multiplicity of equilibria may vanish; and when this happens, the conventional formalizations of "coordination failures" and "animal spirits" fall apart. This underscores the potential fragility of the predictions of standard macroeconomic models—and of the associated interpretations of macroeconomic phenomena—to the kind of friction we are after.

Nonetheless, we also show that this friction offers a new, refined, version of the aforementioned concepts. Interestingly, this new version helps unify unique- and multiple-equilibria models. Most importantly, it leads to new insights into the nature of important phenomena, including financial crises, business cycles, and asset-price fluctuations; and to new lessons for policy.

But how can one operationalize the sought-after friction in coordination? We propose that incomplete information offers a useful method for doing so.[d]

We thus depart from standard macroeconomic models in one crucial respect: we let agents have differential information about the aggregate shocks hitting the economy. This transforms the macroeconomic models of interest into games of incomplete information, thereby also opening the door to rich higher-order beliefs.

Once we look at the equilibria of the modified models, we find that there is no more a unanimous consensus about where the economy is heading: agents fail to reach the same equilibrium beliefs about one another's actions and the associated macroeconomic outcomes. In this sense, coordination has become imperfect.

Shedding light on how this imperfection affects the predictions of macroeconomic models and our interpretation of macroeconomic phenomena is the goal of this chapter.

1.2 Preview and Main Lessons

Our analysis alternates between two modes: one abstract, seeking to highlight general properties; and another applied, spelling out concrete lessons in the context of specific applications.

In Section 2, we set up our baseline framework. The framework is highly stylized: it boils down to a static game with a continuum of agents. Yet, the framework is sufficient for our purposes. First, it encapsulates the fixed-point relation between expectations of economic outcomes and actual behavior that is at the core of most macroeconomic models. Second, it facilitates a sharp analysis of how the theoretical properties of this fixed point, and the implied restrictions on the observables of the model, vary with the information structure. Last but not least, it helps clarify the language employed throughout this chapter.

[d] By "incomplete information" we refer to situations in which agents have dispersed private information about, and lack common knowledge of, aggregate shocks. Our preferred definitions of this and other key concepts are provided in Section 3, after the introduction of our framework.

In Section 3, we offer our preferred definitions of key concepts such as "fundamentals," "incomplete information," and "imperfect coordination." We next distinguish the uncertainty that the agents face about the behavior of others and the associated macroeconomic outcomes, hereafter referred to as "strategic uncertainty," from the uncertainty they face about fundamentals such as preferences and technologies. We finally elaborate on how incomplete information helps disentangle these two types of uncertainty and how it helps formalize the sought-after friction in coordination.

For the remainder of the chapter, we concentrate on two special cases of our framework, each of which corresponds to a different class of models. The first class, which is introduced in Section 4, is known in the literature as "global games." second class, which is introduced in Section 7, is often referred to as "beauty contests." classes exhibit strategic complementarity and use incomplete information to accommodate a friction in coordination; what differentiates them is the strength of the coordination motives and the potential impact of the friction on the determinacy of the equilibrium.

In Section 4, we thus review the following result from the literature on global games, which was initiated by Carlsson and Van Damme (1993a,b) and was further advanced by Morris and Shin (1998, 2001, 2003). Take a classic coordination model, as in Diamond (1982), Diamond and Dybvig (1983), Calvo (1988), and Obstfeld (1986, 1996). Modify it by introducing idiosyncratic noise in the agents' observation of the fundamentals. In the limit as this noise vanishes, the modified model is indistinguishable from the original model in the sense that every agent has nearly perfect knowledge of the underlying fundamentals. And yet, the modified model admits a unique equilibrium.

At the core of this result is the subtle difference between perturbations of mutual knowledge (everybody knows A) and perturbations of common knowledge ("everybody knows A," "everybody knows that everybody knows A," … ad infinitum). The friction may be small in the sense that every agent faces little uncertainty about the underlying fundamentals; and yet it can have a sizable effect on equilibrium outcomes insofar as it shatters common knowledge.

Notwithstanding this subtlety, the result reveals a discontinuity in the predictions of a large body of applied work to perturbations of the informational assumptions. At first glance, this discontinuity is quite troubling: the interpretations of certain phenomena and the related policy implications that macroeconomists have developed on the basis of multiple-equilibria models appear to be in jeopardy.

We argue that this worry is largely not warranted.

First, equilibrium multiplicity is no more needed in order for economic outcomes to be fragile, for self-fulfilling beliefs to matter, and for coordination failures to occur. For instance, we show that small exogenous shocks can still trigger large changes in equilibrium outcomes because, and only because, they trigger large changes in the equilibrium expectations that agents hold about one another's choices.

Second, the determinacy of the equilibrium under incomplete information is sensitive to the particular way that the information structure is modeled. For instance, if market signals such as prices create sufficient correlation in beliefs, or approximate common knowledge, multiplicity may survive the kind of perturbations that would have induced equilibrium uniqueness in the absence of such endogenous signals.

These observations help resuscitate the spirit of the earlier, complete-information models. But this is not the endgame. By offering a more structured way for modeling expectations, the global-games methodology qualifies existing applied lessons and leads to new ones.

In Section 5, we expand on these lessons within the context of several applications. Some of these applications regard bank runs, currency attacks, and sovereign debt crises; others are in the context of business cycles, market freezes, and bubbles. Inter alia, we highlight a number of novel lessons regarding the adverse effects of big speculators, the effectiveness of lenders of last resort such as the IMF, and the role of bank regulation in preventing coordination failures.

We also discuss a few extensions that endogenize the information structure in global games. This permits us to elaborate on the robustness and the empirical content of the uniqueness result. More importantly, it leads to new applied lessons, such as the possibility that an informed policy maker finds herself trapped in a situation where her policy actions are shaped by market expectations rather than the other way around.

Summing up, although the global-game methodology does not offer a panacea for equilibrium uniqueness, it helps open "the black box" of expectations and coordination. This leads to useful applied lessons that were not possible in the context of the earlier, complete-information, literature on coordination failures.

In Section 6, we review two contributions that shift attention to a dynamic issue, the ability of the agents to *synchronize* their choices. The first contribution, which is by Frankel and Pauzner (2000) and Burdzy et al. (2001), uses a friction as in Calvo (1983) to introduce asynchronous choice in a dynamic coordination game and shows that this helps select a unique equilibrium, even when the Calvo-like friction is vanishingly small. The second contribution, which is by Abreu and Brunnermeier (2003), is also concerned with a dynamic coordination game; but instead of adding a Calvo friction, it introduces asynchronous awareness of a shock in the environment; it is then shown that such asynchronous awareness results, not only in asynchronous responses to the aforementioned shock, but also in significant delay of these responses. We discuss how these contributions connect to the global-games literature and how they shed light on the subtle relation between synchronization and coordination.

In Section 7, we shift gears and study a class of models that feature a weaker form of coordination motives than that in the global-games literature. These games, which are often described as beauty contests, have linear best-response conditions and admit a unique equilibrium, features shared by the type of (log)linear DSGE models commonly used to study business cycles and macroeconomic policy, as well as by certain asset-pricing models.

In this class of models, equilibrium determinacy is no longer an issue and the key preoccupation is with the stochastic properties of the equilibrium. Accordingly, the questions we pose are the following: How does an economy respond to innovations to fundamentals? How does this response depend on strategic uncertainty as opposed to fundamental uncertainty? Do coordination motives matter for level and the nature of aggregate volatility? Can one discern a role for "animal spirits" and "self-fulling beliefs"?

The core of Section 7 is devoted to answering the above questions in a relatively abstract manner. The analysis builds heavily on Morris and Shin (2002b), Angeletos and Pavan (2007), and Bergemann and Morris (2013), although we adapt and extend the results of those papers in ways that suit our purposes. Inter alia, we establish the following three general points.

First, incomplete information can help generate significant rigidity, or inertia, in the response of macroeconomic outcomes to aggregate shocks to fundamentals. This inertia can be quantitatively significant even if the noise in the observation of the fundamentals is small, a finding that underscores the distinct role played by strategic uncertainty, or imperfect coordination, relative to fundamental uncertainty.

Second, inertia may obtain at the macro level even if it is largely absent at the micro level. This finding may help liberate macroeconomists from ad hoc adjustment costs that DSGE models have to use in order to match the inertia in the estimated responses of macroeconomic variables to identified shocks to fundamentals. It also highlights a subtle difference between incomplete information and the formalizations of "inattention" found in Sims (2003), Reis (2006), and Gabaix (2014): whereas these formalizations are designed so as to introduce inertia at the microeconomic level, the friction we accommodate in this chapter is uniquely positioned to generate inertia at the macroeconomic level.

Third, this friction can also manifest as extrinsic volatility, that is, as volatility in equilibrium outcomes that is not spanned by the volatility in either the fundamentals or beliefs thereof. It follows that the traditionally dichotomy between models that admit multiple equilibria and those that do not may no longer be central to the understanding of "animal spirits" and "self-fulfilling beliefs": the same type of volatility can now obtain in either class of models.

In Section 8, we consider several applications. In our first application, which is based on Angeletos and La'O (2010), we show how incomplete information can induce rigidity in the response of employment and output to technology shocks. This helps reconcile the RBC paradigm with facts that have so far been considered prima-facia evidence against that model and in favor of the New-Keynesian model (Gali, 1999). We also discuss other works that use informational frictions to generate real rigidities in other contexts.

Our second application concerns nominal rigidity. In particular, we review Woodford (2003), who shows how inertia in higher-order beliefs can be a potent, if not superior, substitute to sticky prices. We also discuss the connection to the complementary works of Mankiw and Reis (2002) and Maćkowiak and Wiederholt (2009), as well as to the seminal contribution by Lucas (1972).

We proceed to review some recent empirical work that uses surveys of economic forecasts to measure the role of frictions in information, most notably by Coibion and Gorodnichenko (2012). We argue that this work provides important evidence in support for the main ideas reviewed in this chapter; but we also highlight certain caveats in the mapping from this evidence to the theory.

In another application, we argue that incomplete information helps accommodate the notion of demand-driven fluctuations without the need for either nominal rigidity or constraints on monetary policy. We also discuss how the incorporation of plausible higher-order belief dynamics in DSGE models can help match salient features of the business-cycle data.

We then turn to a review of applications in finance, which have indicated how incomplete information can also offer a parsimonious explanation to asset-pricing puzzles such as momentum, excess volatility, and the disconnect between exchanges rates and macroeconomic fundamentals.

Section 9 concludes our analysis by touching upon certain normative questions. We first review a notion of constrained efficiency that helps dissect the normative properties of the equilibrium in the class of incomplete-information models we are concerned with. We then discuss two sets of applications: a literature that characterizes optimal monetary policy in the presence of informational frictions; and a literature that studies the welfare effects of the information disseminated by markets, policy makers, or the media.

1.3 Additional Points and Position in the Literature

We now offer a few additional remarks about the theme of our chapter and its position in the literature.

1. The hallmark of the friction we accommodate in this chapter is the inability of agents to reach a consensus about the state of the economy. Along with it comes a certain disentanglement of the uncertainty agents face about endogenous economic outcomes such as employment, output, and asset prices, from their uncertainty about exogenous fundamentals such as preferences and technologies. This disentanglement is made possible with the help of certain types of private information, which permit higher-order beliefs of fundamentals to diverge from the corresponding first-order beliefs. This divergence, in turn, is instrumental to understanding the equilibrium properties of the models we study. However, once equilibrium is imposed, higher-order beliefs become a sideshow: all that matters for the behavior of an agent is the equilibrium expectations (first-order beliefs) she forms about the relevant economic outcomes. Furthermore, for the applications we have in mind, it seems easier

to detect and quantify this kind of expectations, as opposed to the underlying belief hierarchies. These observations explain why the focal point of our analysis is the joint determination of equilibrium outcomes and expectations thereof: higher-order beliefs is the machinery, not the essence, of what we are after.

2. The uncertainty agents face about economic outcomes is tightly connected to the one they face about the behavior of other agents, which is often referred to as "strategic uncertainty" in game theory. Starting from a complete-information model, one notes that the equilibria of such a model rule out any uncertainty about the behavior of others conditional on payoffs.[e] One may then seek to accommodate such uncertainty by relaxing the solution concept. Alternatively, one may maintain the solution concept and instead engineer the sought-after uncertainty with the help of incomplete information. The approach taken in this chapter is the latter. Whenever we talk about strategic uncertainty in the sequel, we therefore refer to the uncertainty that agents face about the actions of others *on* equilibrium, due to incompleteness of information.
3. The frictionless coordination of beliefs and actions that is embedded in standard macroeconomic models hinges on the rational-expectations equilibrium concept just as much as it hinges on the conventional assumption of complete information. It follows that the relaxation of the latter assumption—the approach favored in this chapter—can also be seen as a substitute for relaxing the solution concept. We elaborate on this point in due course.
4. The preceding remarks help distinguish our chapter from the recent literature on "news" and "noise shocks" that has followed Beaudry and Portier (2006).[f] This literature extends workhorse models by letting agents observe noisy signals of future fundamentals, but typically maintains the assumption that all agents share the same information. In so doing, it fails to accommodate the type of friction we are after. In short, it enriches the stochasticity in expectations of fundamentals, but does not disentangle uncertainty of endogenous economic outcomes from uncertainty of exogenous fundamentals. The same point applies to Bloom (2009) and the subsequent literature on "uncertainty shocks."
5. Similar observations help position our chapter also vis-a-vis a strand of the literature that studies various forms of inattention. Consider, in particular, Sims (2003), Reis (2006), and Gabaix (2014). These papers study single-agent problems in which it is

[e] Strictly speaking, this statement applies only to Nash equilibria in pure strategies. However, mixed strategies are not relevant in our context, because we study settings in which there is no loss in ruling out mixed strategies. Also, the statement need not apply to the broader class of *correlated* equilibria, but this precisely because this solution concept departs from complete information.

[f] See, inter alia, Barsky and Sims (2011, 2012), Blanchard et al. (2013), Christiano et al. (2008), Jaimovich and Rebelo (2009), and the baseline model in Lorenzoni (2009). See also the related work by Collard and Dellas (2010) on imperfect observability of current fundamentals.

costly for an agent to acquire information about an exogenous payoff-relevant variable (or otherwise make her action covary appropriately with that variable). In so doing, these papers provide useful decision-theoretic foundations of informational frictions. But they also bypass the distinction between fundamental and strategic uncertainty, for such a distinction is meaningful only in settings in which agents interact with one another.

6. Applied work often confounds the aforementioned two types of uncertainty. To some extent, this is unavoidable: in many applications, accommodating uncertainty about the actions of others *requires* uncertainty about fundamentals. Yet, not only are the two types of uncertainty conceptually distinct, but they can also have different implications. For instance, in the class of global games we study in Sections 4 and 5, what matters for equilibrium determinacy is the strategic uncertainty that is induced by incomplete information, not the underlying fundamental uncertainty. Furthermore, as we elaborate in Sections 7 and 8, incomplete information and the resulting strategic uncertainty can help operationalize the notions of coordination failure and animals spirits within unique-equilibrium models, in a manner that no kind of fundamental uncertainty alone can do. An integral part of our analysis is therefore to disentangle the two types of uncertainty as much as possible. We hope that this will help clarify, not only the theoretical underpinnings, but also the empirical implications of the blossoming macroeconomics literature on informational frictions.
7. The aforementioned literature contains many strands. This chapter occupies only the subspace in this literature that relates to coordination, higher-order beliefs, and strategic uncertainty. For complementary reviews of the literature, see Sims (2010), Mankiw and Reis (2011), and Veldkamp (2011).
8. Models with search and trading frictions are occasionally interpreted as models with imperfect coordination; see, eg, Diamond (1982) and Shimer (2005). This is not what *we* have in mind. To the extent that these models maintain common knowledge of aggregate fundamentals and aggregate outcomes, for our purposes they impose the same level of coordination as the Arrow–Debreu framework.[g]
9. We are more sympathetic to the notion that asynchronous choice impedes coordination. We expand on this point in Section 6 by uncovering a subtle connection between the role of incomplete information and that of a Calvo-like friction in settings with strategic complementarities.
10. Our treatment of incomplete information is consistent with standard treatments in Bayesian games. In particular, our baseline framework specifies information in terms of Harsanyi types: the signal of an agent encodes, not only her beliefs about payoffs (fundamentals), but also her beliefs, or "state of mind," regarding the types of other

[g] That said, trading frictions can be instrumental for sustaining lack of common knowledge: accommodating incomplete information *requires* a departure from the Arrow–Debreu framework.

agents. We nevertheless adjust some of the definitions in manners that suit our applied purposes. We also bypass a number of deeper theoretical issues that are beyond the scope of this chapter, such as the equivalence between Harsanyi types and belief hierarchies (Mertens and Zamir, 1985) and the approximation of common knowledge with common beliefs (Monderer and Samet, 1989; Morris and Shin, 1997), as well as all the concerns of epistemic game theory (Dekel et al., 2015).

11. In our baseline framework, as well as in various applications, we favor a broad interpretation of incomplete information: signals are allowed to capture subjective states of mind about the beliefs and actions of others. For certain questions, however, a more narrow interpretation may be appropriate: when we study the aggregation of information through markets or other channels, or when we touch upon the desirability of central-bank communication, it is safer to assume that signals represent hard information.

2. FRAMEWORK

In this section we introduce our main framework. Although it is static and abstract, the framework is quite flexible and stylizes role of general-equilibrium interactions, and of coordination, in a variety of applications. Furthermore, it facilitates the formalization of a number of key concepts in the next section; it helps illustrate how the predictions of standard macroeconomic models hinge on strong assumptions about the determination of beliefs; and it paves the way to a number of applications presented later on.

2.1 Actions and Payoffs

Consider a one-shot game, or "economy," with a large number of players, or "agents," indexed by $i \in [0,1]$. Each agent i chooses an action $k_i \in D_k \subseteq \mathbb{R}^n$, with $n \geq 1$. His payoff is given by

$$u_i = U(k_i, \mathbf{K}, \theta_i), \tag{1}$$

where $\mathbf{K} \in D_{\mathbf{K}}$ denotes the distribution of actions in the cross section of the population (the aggregate economic activity), $\theta_i \in D_\theta \subseteq \mathbb{R}^m$, $m \geq 1$, summarizes any exogenous variable that is payoff-relevant to agent i (her fundamental), and U is a function, which will be further specified as we proceed.

As we move on, we will discuss a number of applications that can be nested either directly in the above framework or in appropriate extensions of it. In some of the applications, the players are speculators deciding whether to attack a currency, or depositors deciding whether to run against a bank. In others, they are firms making production and pricing choices, or consumers choosing how much to spend. Also note that the specification assumed in condition (1) allows an agent's payoff to depend on the entire distribution of actions in the cross-section of the population. In applications, however, the first and second moments (ie, the mean and the dispersion) often suffice. Furthermore, although we momentarily allow for both the actions and the fundamentals to be multidimensional, we will soon restrict attention to settings where $n = m = 1$. In such a case, we let $K \equiv \int k d\mathbf{K}(k)$ and $\sigma_k \equiv \sqrt{\int (k-K)^2 d\mathbf{K}(k)}$ denote, respectively, the mean action and the cross-sectional standard deviation of actions.

Remark 1 One may find it more appealing to represent the macroeconomy as a network where each agent is connected to a small set of other agents as opposed to the entire population. The ideas we develop in this chapter apply in such settings as well. For our purposes, the fact that the economy is modeled as a complete and symmetric network is mostly a simplification. For other purposes, however, it may be important to abandon this simplification. For instance, an important question that we will not address is how the effects of strategic uncertainty on observables depend on the network structure.

Remark 2 Recall that the question of interest for us is the equilibrium expectations of economic outcomes. In the static framework we introduce in this section, this question is reduced to the question of the equilibrium expectations of the *contemporaneous* actions of others. In dynamic applications in which decisions are forward-looking, the relevant expectations also regard the actions of others in *future* periods.

Remark 3 Notwithstanding the previous remark, there is an important multiperiod setting that defies the aforementioned distinction and can be readily nested in our static framework: an Arrow–Debreu economy, with a complete market for time- and state-contingent goods. In such an economy, θ_i captures the preferences and endowments of agent i, k_i captures her net demand of the various goods, and the dependence of the agent's utility on $\mathbf{K}$ emerges from the dependence of equilibrium prices on the net demands of other agents. Importantly, because in a complete Arrow–Debreu market all agents get to observe all the relevant prices, it is *as if* they also observe the actions of other agents. This underscores that the Arrow–Debreu framework and many workhorse macroeconomic models that are based on it are, in effect, a particular class of static, complete-information games. Conversely, accommodating incomplete information in macroeconomic models requires some kind of market incompleteness (such as missing forward markets).

2.2 Examples

We now use two simple examples to illustrate how the abstract payoff structure assumed above corresponds to a reduced-form representation of more concrete applications. The first example is a neoclassical economy similar to those studied in Angeletos and La'O (2010, 2013), Benhabib et al. (2015b), Sockin and Xiong (2015), and Huo and Takayama (2015a). The second is a monetary economy similar to those studied in Woodford (2003), Mankiw and Reis (2002), and Maćkowiak and Wiederholt (2009). Later on, we will explore the implications of a certain kind of informational friction in these and other examples. For now, we abstract for the informational friction and focus on demonstrating how these examples can be nested in our framework. Apart from providing concrete interpretations of the framework, these examples help clarify that, for our purposes, strategic interaction is often synonymous to market-based general-equilibrium effects.

2.2.1 An Neoclassical Economy

There is a continuum of "farmers," which can be interpreted as both firms and consumers, and a continuum of differentiated consumption goods. Each farmer specializes in the production of a single good but consumes all the goods in the economy. We let $i \in [0, 1]$ index both the farmer and the good she produces. There are two stages. Production takes places in stage 1; trading and consumption take place in stage 2. Preferences are given by

$$u_i = v(c_i) - n_i, \tag{2}$$

where v is a strictly increasing and strictly concave function, n_i denotes labor, and c_i is a CES aggregator of the farmer's consumption of all the goods. That is,

$$c_i = \left[\int c_{ij}^{1-\eta} dj \right]^{\frac{1}{1-\eta}}$$

where c_{ij} denotes the consumption of good j by household i and $\frac{1}{\eta}$ is the elasticity of substitution across goods. The budget constraint of any farmer i is given by $\int p_j c_{ij} dj = p_i q_i$, where p_j denotes the (nominal) price of good j and q_i the farmer's output. Finally, production is given by

$$q_i = A_i n_i,$$

where A_i is the farmer's exogenous productivity.

As is well known, the CES specification implies that the optimal consumption bundle of farmer i satisfies

$$\int p_j c_{ij} dj = P c_i \quad \text{and} \quad \frac{c_{ij}}{c_i} = \left(\frac{p_j}{P} \right)^{-1/\eta} \quad \forall j,$$

where $P \equiv \left[\int p_j^{1-1/\eta} dj \right]^{\frac{1}{1-1/\eta}}$ is the ideal price index. Market clearing imposes $\int c_{ji} dj = q_i \;\; \forall i$. Using this yields the following relation between the market-clearing prices and the quantities produced:

$$p_i = P \left(\frac{q_i}{Q} \right)^{-\eta} \quad \forall j, \tag{3}$$

where $Q \equiv \left[\int q_j^{1-\eta} dj\right]^{\frac{1}{1-\eta}}$ measures aggregate output. The budget constraint, on the other hand, gives $c_i = \frac{p_i q_i}{P} = Q^\eta q_i^{1-\eta}$. It follows that the utility of farmer i reduces to the following:

$$u_i = v(c_i) - n_i = v\left(Q^\eta q_i^{1-\eta}\right) - \frac{q_i}{A_i}.$$

We conclude that the example we have introduced here can be readily nested in our framework if we let

$$U(k_i, \mathbf{K}, \theta_i) = v\left(\left[\int \exp(x)^{1-\eta} d\mathbf{K}(x)\right]^{\frac{\eta}{1-\eta}} \exp(k_i)^{1-\eta}\right) - \exp(k_i - \theta_i), \tag{4}$$

with $k_i \equiv \log q_i$ and $\theta_i \equiv \log A_i$. That is, the economy under consideration can be interpreted as a game in which the players are the farmers, the actions are the quantities these farmers produce, the fundamentals are their exogenous productivities, and the payoffs are given by (4).[h]

2.2.2 A Monetary Economy

The structure of the economy is as above, except for the following two modifications. First, the farmers set nominal prices in stage 1 and commit to accommodate any quantity demanded in stage 2 at these prices. Second, the level of nominal GDP is given by

$$PQ = M, \tag{5}$$

where M is an exogenous variable. In line with much of the related literature, one may interpret the variation in M as monetary shocks.

Following similar steps to those above, we can solve for quantities as functions of prices (rather than the other way around). In particular, using (3), (5), and the budget constraint, we get

$$q_i = Q\left(\frac{p_i}{P}\right)^{-1/\eta} = MP^{1/\eta - 1} p_i^{-1/\eta} \quad \text{and} \quad c_i = \frac{p_i q_i}{P} = MP^{1/\eta - 2} p_i^{1-1/\eta}.$$

[h] Note that k_i and θ_i could have been defined as the absolute levels of q_i and A_i, rather than their logarithms. We opt for the logarithmic transformation because of the following reason: when we assume that productivities are log-normally distributed and that the available information is Gaussian, the power specification of preferences and technologies in this example guarantees the existence of a unique equilibrium in which $k_i \equiv \log q_i$ can be expressed as a linear function of the firm's expectation of $\theta_i \equiv \log A_i$ and K, with K itself being equal to $\log Q$ plus a constant. This in turn means that the neoclassical economy under consideration maps into the specific class of games we study in Sections 7 and 8.

The present example is nested in our framework by letting $k_i \equiv \log p_i$, $\theta_i \equiv (\theta_{i1}, \theta_2) \equiv (\log A_i, \log M)$, and[i]

$$U(k_i, \mathbf{K}, (\theta_{i1}, \theta_2)) = v\left(\exp(\theta_2)\left[\int \exp(x)^{1-1/\eta} d\mathbf{K}(x)\right]^{\frac{1/\eta-2}{1-1/\eta}} \exp(k_i)^{1-1/\eta} \right) \\ - \exp\left(\theta_2 - \theta_{i1} - \frac{1}{\eta} k_i\right)\left[\int \exp(x)^{1-1/\eta} d\mathbf{K}(x)\right]^{-1}. \tag{6}$$

That is, the actions now are the nominal prices set by the farmers and the fundamentals are the productivity and monetary shocks. Note that in this example, productivity is allowed to have both an idiosyncratic and an aggregate component, whereas the monetary shock has only an aggregate component. Woodford (2003) and Mankiw and Reis (2002) are then nested by shutting down all the productivity shocks, whereas Maćkowiak and Wiederholt (2009) is nested by letting the productivity shocks be purely idiosyncratic.

Remark 4 The above two examples can be thought as representing two diametrically opposite cases: in the first, firms set quantities and let prices adjust; in the second, firms set prices and let quantities adjust. This difference is inconsequential when the firms face no uncertainty, but matters for both positive and normative issues if the firms are uncertain about either the relevant exogenous fundamentals or the choices of other firms. We elaborate on these points in Sections 8 and 9.

Remark 5 Although our examples are populated by "farmers," one should of course not take this too literally. Starting with Lucas (1972), various authors have considered models that feature more appealing micro-foundations along with informational frictions. The key question for our purposes is whether these models make room for nontrivial strategic, or higher-order, uncertainty. This is what distinguishes the more recent works mentioned above from Lucas' earlier contribution: as further explained in Section 8.3, the specific model used in Lucas (1972) shuts down the effects we are after in this chapter.

2.3 Shocks and Information

So far, we have specified actions and payoffs. To complete the description of the framework, we must specify the exogenous stochastic structure (shocks and information) and must also pick a solution concept. We complete these two tasks in, respectively, the present section and Section 2.4.

We let $\omega_i \in D_\omega \subseteq \mathbb{R}^l$ denote the signal received by agent i, where by "signal" we mean the entire information set of the agent, or equivalently her Harsanyi type. We denote by $\mathbf{\Omega}$ the distribution of ω_i in the cross-section of the population, and by $\mathbf{\Theta}$ the cross-sectional distribution of θ_i. We let $D_{\mathbf{\Omega}}$ and $D_{\mathbf{\Theta}}$ denote the sets of the possible

[i] The logarithmic transformation is used for the same reason as in footnote h.

values for, respectively, $\mathbf{\Omega}$ and $\mathbf{\Theta}$. We let $\mathbf{S}$ denote the distribution that Nature uses to draw, in an i.i.d. fashion, a pair $s_i = (\omega_i, \theta_i)$ for each i, and let $D_{\mathbf{S}}$ denote the set of possible values for $\mathbf{S}$. We assume that a version of the law of large numbers applies, so that $\mathbf{S}$ is also the cross-sectional distribution of s_i in the population. It follows that $\mathbf{\Theta}$ and $\mathbf{\Omega}$ are also the marginals of $\mathbf{S}$ in, respectively, the θ and ω direction. We introduce aggregate uncertainty by allowing $\mathbf{S}$ (and hence also $\mathbf{\Theta}$ and $\mathbf{\Omega}$) to be random, drawn from the set $D_{\mathbf{S}}$ according to some fixed distribution $\mathcal{P}$, which constitutes the common prior.

One can thus describe the "moves of Nature" as follows. First, Nature draws $\mathbf{S}$ according to $\mathcal{P}$. Next, Nature uses $\mathbf{S}$ to draw, in an i.i.d. manner, a pair $s_i = (\theta_i, \omega_i)$ for each agent i. Finally, Nature reveals ω_i (and only that) to agent i.

The objects $\{U, D_k, D_{\mathbf{K}}, \mathcal{P}, D_\theta, D_\Theta, D_\omega, D_\Omega, D_{\mathbf{S}}\}$ and all the aforementioned facts are common knowledge. What may not be common knowledge is the realization of $\mathbf{S}$, and therefore also the realizations of $\mathbf{\Theta}$ and $\mathbf{\Omega}$. Different specifications of the stochastic structure—that is, of the prior $\mathcal{P}$ and the associated domains for the fundamentals and the signals—can accommodate different scenarios about how much each agent knows, not only about her own fundamentals, but also about the fundamentals and the information of other agents. Importantly, we can accommodate aggregate and idiosyncratic shocks, not only to fundamentals, but also to information.

To give a concrete example, consider the neoclassical economy introduced earlier on and let $\theta_i \equiv \log A_i = \overline{\theta} + \xi_i$, where $\overline{\theta}$ and ξ_i are independent Normally distributed variables with mean zero and variances, respectively, σ_θ^2 and σ_ξ^2, and ξ_i is i.i.d. across i. Then, variation in $\overline{\theta}$ represents aggregate TFP shocks, whereas variation in ξ_i represents idiosyncratic TFP shocks. Concerning the information structure, we could then let $\omega_i = (x_i, z)$, where $x_i = \theta_i + \epsilon_i$, $z = \overline{\theta} + \zeta$, and ϵ_i and ζ are Normally distributed, independent of one another, of $\overline{\theta}$, and of ξ_i for all i, with mean zero and fixed variances σ_ϵ^2 and σ_ζ^2. This would mean that the information set of an agent is given by the pair of two signals, a private signal of her own TFP and a public signal of the aggregate TFP. Finally, because all the relevant cross-sectional distributions, namely $\mathbf{\Theta}$, $\mathbf{\Omega}$, and $\mathbf{S}$, are now Normal distributions with fixed variances, we can think of the aggregate shocks in terms of the pair $(\overline{\theta}, \zeta) \in \mathbb{R}^2$ rather than of the high-dimensional objects $(\mathbf{\Theta}, \mathbf{\Omega})$ or $\mathbf{S}$.

In the example described above, x_i is a signal, not only of θ_i, but also of $\overline{\theta}$; and since $\overline{\theta}$ is the mean of the realized distribution of x_j for $j \neq i$, it follows that x_i is also signal of the signals of others. More generally, note that different realizations of $\mathbf{S}$ correspond to different conditional distributions for ω_i, which means that ω_i is a signal of $\mathbf{S}$ and hence also of $\mathbf{\Omega}$.[j] This underscores that ω_i shapes the agent's belief, not only about her own

[j] Note that ω_i is also a signal of $\mathbf{\Theta}$, the fundamentals of others. This, however, is not relevant per se. An agent cares only about her own fundamental and the actions of others. In equilibrium, the latter are pinned down by their information. What is relevant is therefore only the information that ω_i contains about the agent's own fundamental and the information of others.

payoff, but also about the *information* of others. Once we impose equilibrium, the latter will mean that ω_i shapes the beliefs about the *actions* of others.

We can now offer the following definitions, which help fix the language we use in this chapter.

Definition 1 The realized *fundamentals* are given by $\mathbf{\Theta}$, the distribution of the exogenous payoff characteristic θ_i in the cross-section of the population. The realized *information* is given by $\mathbf{\Omega}$, the corresponding distribution of the signal ω_i. The *state of Nature* is given by $\mathbf{S}$, the joint distribution of (ω_i, θ_i).[k]

Note that the above objects, as many others we introduce below, refer to the cross-section of the population, as opposed to any specific agent. This reflects merely our focus on macroeconomic outcomes, as opposed to the behavior of each individual player.

We now introduce our notation for first- and higher-order beliefs. Let b_i denote the belief of agent i about her own fundamental and the aggregate fundamentals, that is, the posterior of $(\theta_i, \mathbf{\Theta})$ conditional on ω_i. This belief encapsulates the agent's information about the payoff structure of the environment and is also known as her *first-order* belief (of the underlying fundamental). Clearly, b_i is a function of ω_i. Let $\mathbf{B}$ denote the realized distribution of b_j in the cross-section of the population; this is a function of $\mathbf{\Omega}$. Define the *second-order* belief of agent i as her posterior about $\mathbf{B}$ and denote it by b_i^2. Since $\mathbf{B}$ is a function of $\mathbf{\Omega}$ and since ω_i pins down i's belief about $\mathbf{\Omega}$, ω_i pins down b_i^2 as well. Let $\mathbf{B}^2$ denote the cross-sectional distribution of second-order beliefs. We can then iteratively define, for any $h \geq 3$, the hth order belief of i, b_i^h, as her belief about $\mathbf{B}^{h-1}$; and the object $\mathbf{B}^h$ as the cross-sectional distribution of b_j^h. To ease notation, we also let $b_i^1 \equiv b_i$ and $\mathbf{B}^1 \equiv \mathbf{B}$.

Definition 2 The (aggregate) *beliefs of fundamentals* are given by $\mathbf{B}$, the cross section of the first-order beliefs of own and aggregate fundamentals. The corresponding belief hierarchy is given by $\left\{\mathbf{B}^h\right\}_{h=1}^{\infty}$. An agent faces *higher-order uncertainty* (about the fundamentals) if she is uncertain about $\left\{\mathbf{B}^h\right\}_{h=1}^{\infty}$.

Because $\left\{\mathbf{B}^h\right\}_{h=1}^{\infty}$ is pinned down by $\mathbf{\Omega}$, we can think of $\mathbf{\Omega}$ interchangeably either as the cross-sectional profile of information or as a summary statistic of the belief hierarchy. If $\mathbf{\Omega}$ is known to an agent, then so is $\left\{\mathbf{B}^h\right\}_{h=1}^{\infty}$. Conversely, uncertainty about $\mathbf{\Omega}$ helps accommodate higher-order uncertainty about fundamentals. Importantly, once we impose equilibrium, such higher-order uncertainty helps induce *first-order* uncertainty about the actions of others, ultimately capturing the coordination friction we are after.

Remark 6 For the rest of this chapter, and unless otherwise stated, the terms "higher-order beliefs" and "higher-order uncertainty" refer to the kind of higher-order beliefs of fundamentals defined above, as opposed to higher-order beliefs of either Harsanyi types (beliefs of beliefs of... $\mathbf{\Omega}$) or actions (beliefs of beliefs of... $\mathbf{K}$). In general, these three

[k] Note that $\mathbf{S}$ is the joint distribution of $s_i = (\theta_i, \omega_i)$, not just the pair of the corresponding marginals: $\mathbf{S}$ contains more information than the pair $(\mathbf{\Theta}, \mathbf{\Omega})$.

types of higher-order beliefs are distinct from one another. For instance, in Aumann (1974, 1987), private information about payoff-irrelevant variables (correlation devices) helps sustain higher-order uncertainty about the information and the actions of others, while maintaining common knowledge of fundamentals.[1] Furthermore, the distinction is central to epistemic game theory. For the purposes of this chapter, however, one can think of the three types of higher-order uncertainty as different facets of one and the same departure from workhorse macroeconomic models. Indeed, in the applications we study in this chapter, and insofar as the equilibrium is unique, the equilibrium beliefs of actions is pinned down by the hierarchy of beliefs of fundamentals. It follows that the friction we are after can be accommodated only by introducing higher-order uncertainty about fundamentals.

Remark 7 Whenever we use the terms "expectations," "beliefs," or "uncertainty" without the explicit qualifier "higher-order" in front of these terms, we refer to first-order beliefs.

Remark 8 In game theory, it is a standard practice to write the payoff of a player as function of her Harsanyi type. This could be done here by recasting payoffs as $V(k_i, \mathbf{K}, \omega_i) \equiv \int U(k_i, \mathbf{K}, \theta_i) db(\theta_i|\omega_i)$, where $b(\theta_i|\omega_i)$ is the posterior belief about θ_i conditional on ω_i. We have opted to express payoffs as functions of θ_i rather than ω_i in order to disentangle fundamentals from information sets and to accommodate the possibility that the econometrician observes θ_i but not ω_i.

2.4 Equilibrium

Let us now turn to the solution concept. In line with the vast majority of macroeconomic research, we assume that agents play according to Rational-Expectations Equilibrium. In the present framework, this means the following.

Definition 3 A *rational-expectations equilibrium* (or, simply, an equilibrium) is a strategy $k^* \in D_\omega \to D_k$ and a mapping $\mathbf{K}^* : D_\Omega \to D_\mathbf{K}$ such that:

(i) the strategy k^* is the best response to the mapping $\mathbf{K}^*$, that is,

$$k^*(\omega) \in \arg\max_{k \in D_k} \mathbb{E}[U(k, \mathbf{K}^*(\Omega), \theta)|\omega] \quad \forall \omega; \tag{7}$$

(ii) for any $\mathbf{\Omega}$, $\mathbf{K}^*(\mathbf{\Omega})$ is the distribution of k that obtains when the distribution of ω is $\mathbf{\Omega}$ and $k = k^*(\omega)$.

This definition is essentially the same as the definition of Bayesian-Nash Equilibrium in games. This, however, does not mean that the agents themselves engage in strategic reasoning. The context we have in mind is a large market-based economy, in which there

[1] With our notation, this corresponds to situations where $\mathbf{B}$ is commonly known, but $\mathbf{\Omega}$ is not. This is because $\mathbf{\Omega}$ contains correlation devices about which agents have private information. By the same token, although higher-order beliefs collapse to first-order beliefs in the case of fundamentals, the same is not true in the case of actions.

is myriad of firms and consumers, each one trying to predict economic outcomes such as aggregate employment and output. This is quite different from an arms race or other contexts in which it may be more appealing to think that each player is explicitly trying to second guess the moves of other players. We thus envision that each agent treats $\mathbf{K}$ as an exogenous stochastic process and we seek to understand the predictions that any given model makes about this process.

Under this interpretation, the adopted solution concept imposes that all agents perceive the same stochastic process for $\mathbf{K}$, as well as that this commonly perceived process coincides with the actual one generated by the joint behavior of all agents. Furthermore, ω_i becomes a signal of the realized value of $\mathbf{K}$. By the same token, we can think of ω_i as the agent's "state of mind" about what is going on in the economy. Importantly, insofar as different agents do not share the same ω_i, they also do not need to share the same expectations about the endogenous outcomes of interest.

Summing up, we have that ω_i can play any subset of the following three modeling roles. First, it shapes the agent's beliefs about her own fundamental. Second, it shapes the agent's beliefs about the information of others and thereby also her higher-order beliefs of the fundamentals. Third, it shapes the agent's equilibrium expectations about the actions of others and thereby about the endogenous outcomes of interest. The first two modeling roles are by construction; the third rests on imposing the adopted solution concept.

Some of the applied literature on informational frictions relies on the first modeling role: it removes knowledge of θ_i and lets ω_i be a noisy signal of θ_i, often in settings that rule out strategic interactions. In this chapter, by contrast, we are primarily concerned with the other two modeling roles, which are themselves relevant only in the presence of strategic interactions.[m]

Remark 9 Once the equilibrium concept has been imposed, higher-order beliefs become a sideshow: in equilibrium, an agent's signal ω_i pins down her posterior about the joint distribution of θ_i and K, which is all that matter for her behavior. Therefore, the equilibrium can be characterized without ever invoking higher-order beliefs. However, understanding the structure of higher-order beliefs turns out to be instrumental to understanding the structure of the entire equilibrium set and its sensitivity to variations of the information structure. This will become evident as we proceed.

Remark 10 When we move on to dynamic settings, the equilibrium concept we use is essentially the same as Perfect Bayesian Equilibrium. The only difference—and a useful simplification—is that we do not always need to specify out-of-equilibrium beliefs. This is because, in the typical macroeconomic setting, each private agent is too small to trigger

[m] With this point in mind, in the sequel we will often consider specifications of the information structure that help isolate the stochastic variation in the equilibrium expectations of economic outcomes from the stochastic variation in beliefs of fundamentals.

deviations that are detectable by, and payoff-relevant to, any other agent. An important exemption to this rule, however, is settings with big players, such as the government.

Remark 11 While we have introduced the notation and the concepts needed for our purposes, we have not spelled out the mathematical structure of the underlying probability space and have swept under the rug technical issues regarding measurability and existence. We hope that this simplifies the exposition without obstructing the understanding of the more substantive issues. If the reader is annoyed by this imprecision, he/she can read the next section as if the sets of actions, fundamentals, and signals were finite. In subsequent sections, we will impose enough assumptions on the stochastic structure and the function U to guarantee the existence of an equilibrium, as well as its tractable characterization.

3. IMPERFECT COORDINATION

In this section, we use our framework to define certain key notions and to elaborate on the central theme of this paper. We also discuss how equilibrium expectations and outcomes are determined under incomplete information, and contrast them to their complete-information counterparts. The complete-information case defines the benchmark relative to which equilibrium expectations are enriched.

3.1 Some Key Notions

Macroeconomic models are typically used to provide a structural interpretation of the data, or to deliver counterfactuals that can help guide policy. To clarify what this means within our framework, we start by distinguishing the exogenous objects that the theorist can specify "at will" from the endogenous objects that she wishes to predict. The exogenous objects are the payoff function U together with the stochastic structure of the fundamentals and the information sets introduced in the previous section. The endogenous objects are the actions of the agents and the realized payoffs. Because the payoffs are themselves pinned down by the fundamentals and the actions, the only "true" endogenous objects are the actions. With this in mind, we adopt the following definition.

Definition 4 The *economy's outcome*, or the economy's endogenous state, is given by $\mathbf{K}$, the distribution of actions in the cross-section of the population.

In the context of an application, $\mathbf{K}$ may represent the hiring choices of the firms, the spending choices of consumers, or the capital stock of the economy (which is itself the product of past investment choices by firms and consumers). What a model ultimately does is to impose certain restrictions on the joint distribution of these endogenous objects with the exogenous shocks to fundamentals and information sets.

Suppose that the data that are available for testing or quantifying the model regard *at most* the following objects: the fundamentals, the agents' expectations of their fundamentals, their expectations of the endogenous state, and their actions. Then, the testable implications, or predictions, of the model are the restrictions that it imposes on the joint distribution of these objects. Of course, one can expand this list of objects by including the information sets and/or the higher-order beliefs. For applied purposes, however, this seems redundant, because data on information sets and higher-order beliefs are hardly available, in contrast to data on choices and on forecasts of economic outcomes. With these points in mind, we let π_i denote the expectation of agent i about $\mathbf{K}$; we let $\mathbf{\Pi}$ denote the distribution of such expectations in the population; and we adopt the following definition.

Definition 5 The *predictions* of the model are the restrictions it imposes on the joint distribution of $(\theta_i, b_i, \pi_i, k_i, \mathbf{\Theta}, \mathbf{B}, \mathbf{\Pi}, \mathbf{K})$.

One can thus think of any model as a "box" that takes the joint distribution of $(s_i, \mathbf{S})$ as an input and delivers a joint distribution for $(\theta_i, b_i, \pi, k_i, \mathbf{\Theta}, \mathbf{B}, \mathbf{\Pi}, \mathbf{K})$ as an output. Depending on the question of interest and/or the available data, one can then look at certain conditionals or marginals of the aforementioned distribution, or certain moments of it. One can thus also answer the following type of applied questions: How much does individual activity respond to idiosyncratic shocks to fundamentals? What is the corresponding response of aggregate activity to aggregate shocks to fundamentals? Are expectations of economic outcomes pinned down by expectations of fundamentals? What is the volatility in aggregate economic activity and how much of it is explained by volatility in fundamentals?

To illustrate, consider our earlier neoclassical economy from Section 2.2, in which case θ_i captures the exogenous TFP of an agent (farmer, firm, or island) and k_i captures her production level. Assume that TFP is log-normally distributed in the cross section, let $\overline{\theta}$ denote the aggregate log-TFP, and assume $\omega_i = (\theta_i, \overline{\theta})$; the latter means that every agent knows perfectly both her own TFP level and the aggregate TFP level. These assumptions, in conjunction with the power-form specification of preferences and technologies, imply that the equilibrium levels of output at the individual and the aggregate level are given by, respectively,

$$\log q_i \equiv k_i = \kappa_1(\theta_i - \overline{\theta}) + \kappa_2\overline{\theta} \quad \text{and} \quad \log Q \equiv K = \kappa_2\overline{\theta},$$

where κ_1 and κ_2 are fixed scalars, pinned down by the underlying preference and technology parameters. Equilibrium outcomes are therefore pinned down by fundamentals—and so are equilibrium expectations. Furthermore, the joint distribution of $(\theta_i, b_i, \pi_i, k_i, \mathbf{\Theta}, \mathbf{B}, \mathbf{\Pi}, \mathbf{K})$ is now conveniently summarized by the distribution of $(\theta_i, k_i, \overline{\theta}, K)$, Turning to some of the applied questions raised above, note that the "micro" elasticity of the response of individual output to idiosyncratic TFP shocks is given by κ_1, whereas the corresponding "macro" elasticity is given by κ_2. Finally, note that

$$\kappa_1 = \frac{Cov(k_i, \theta_i | \overline{\theta})}{Var(\theta_i | \overline{\theta})} \quad \text{and} \quad \kappa_2 = \frac{Cov(K, \overline{\theta})}{Var(\overline{\theta})},$$

which illustrates how the joint distribution of $(\theta_i, k_i, \overline{\theta}, K)$ contains the model's predictions about the two elasticities of interest.[n]

Later on, we will explore how this kind of predictions change as one departs from standard informational assumptions. Importantly, we will draw a distinction between two kinds of departure: those that regard the information, or beliefs, agents have about their fundamentals; and those regard the information, or beliefs, agents have about one another's beliefs and the associated friction in coordination (or lack thereof).

To do so, we need to introduce a few additional definitions. We start by distinguishing the two kinds of uncertainty mentioned in the Introduction.

Definition 6 An agent faces *fundamental uncertainty* if and only if, conditional on her information, she is uncertain about the value of $(\theta_i, \mathbf{\Theta})$.

Definition 7 An agent faces *uncertainty about the actions of others*, or about the economy's outcome, if and only if, conditional on her information, she is uncertain about the value of $\mathbf{K}$.

As noted in the Introduction, the latter type of uncertainty is connected to the notion of strategic uncertainty in games. In what follows, we use the term "strategic uncertainty" to refer to the type of uncertainty defined in Definition 7 and proceed to study how incomplete information helps accommodate such uncertainty in equilibrium. This is consistent with Morris and Shin (2002a, 2003). Note, however, that the same term is often used in game theory to refer to a distinct object, namely the uncertainty that agents may face about the strategies of others outside equilibrium.

Also note that Definitions 6 and 7 refer to the uncertainty the agents face in the interim stage, after they have received their information, as opposed to the stochasticity that exists at the ex ante stage. For instance, consider the neoclassical economy introduced in Section 2.2 and suppose that TFP is random but perfectly observed. We would say that the agents in this economy face no fundamental uncertainty.[o]

The aforementioned definitions invite us to put ourselves in the shoes of a particular agent and to examine the uncertainty the agent faces conditional on her information. The next definition, instead, invites us to inspect the cross-section of agents conditional

[n] Macroeconomists are primarily interested in the macro elasticity, κ_2. Empirical studies that use macro data, such as the paper by Gali (1999) reviewed in Section 8.1, seek to estimate this elasticity directly. By contrast, studies based on cross-sectional data concentrate on the identification of the micro-level elasticity, κ_1 (where "micro" may mean either at the level of an individual firm/household, or at the level of a region). The difference between the two types of elasticities reflects general-equilibrium effects that operate at the aggregate level and can thus not be identified from the cross section. See, however, Section 8.2 for a discussion of how incomplete information can help reduce this difference in the short run.

[o] One has to be careful with the extension of this notion to dynamic applications: in the RBC model (with capital), we would say that fundamental uncertainty is absent when the agents know, not on current TFP, but also future TFP.

on the state of nature. It then asks the following question: have the agents been able to reach a common belief about the relevant economic outcomes?

Definition 8 *Coordination is imperfect* in state **S** if and only if, in that state, a positive measure of agents fail to reach the same belief about **K**.

To see what motivates this definition, consider the following hypothetical scenario. Prior to playing an equilibrium, the agents get together in the same room and communicate with one another until they reach common knowledge, not only of the fundamentals, but also of their intended actions and the relevant economic outcomes (ie, of **K** in our setting). We think of this situation as being conducive to "perfect coordination." Conversely, we say that there is a friction in coordination if agents are unable to reach a common belief about one another's actions and thereby about the relevant economic outcomes.

The following remark helps refine the friction we are after. In our framework, once we impose equilibrium, reaching the same belief about the choices of others means facing no uncertainty about them. The last property, however, is an artifact of the static nature of our framework. In dynamic settings, agents may face uncertainty about the *future* choices of others, even if they do not face any uncertainty about the *current* choices of others. Reaching a common belief about the relevant economic outcomes is therefore distinct from facing no uncertainty about them. In particular, standard macroeconomic models accommodate for uncertainty in future economic activity, due to uncertainty in future fundamentals; they nevertheless impose that agents always hold exactly the same equilibrium beliefs about all future economic outcomes. It is this restriction—the unanimous consensus about the current state and the future prospects of the economy—that we interpret as perfect coordination and that we seek to depart from.

We can now also elaborate on our earlier claim that some of the related applied literature either confounds the roles of payoff and strategic uncertainty, or is exclusively concerned with fundamental uncertainty. When Sims (2003), Reis (2006), and Gabaix (2014) introduce different (but also complementary) formalizations of "inattention," they focus on single-agent decision problems, thus precluding a meaningful distinction between the two types of uncertainty.[p] When Bloom (2009), Bloom et al. (2014), and Arellano et al. (2012) study the role of "uncertainty shocks," or when Jaimovich and Rebelo (2009), Christiano et al. (2008), Beaudry and Portier (2006), Barsky and Sims (2011, 2012), and Blanchard et al. (2013) study the role of "news" and "noise shocks," they consider different facets of fundamental uncertainty,

[p] The same applies to many of the recent applications of rational inattention, including Luo (2008), Matejka (2015a,b), and Matejka et al. (2015); but not to those that let rational inattention be, in effect, a specific micro-foundation of strategic uncertainty, such as Maćkowiak and Wiederholt (2009, 2015). See Section 8.5 for further discussion.

but do not accommodate the kind of strategic uncertainty, or the friction in coordination, that we are after in this chapter.[q]

By contrast, when Aumann (1974, 1987) introduced the concept of correlated equilibrium and formalized its connection to Bayesian rationality in games, he ruled out fundamental uncertainty and used incomplete information to model *exclusively* uncertainty about the actions of others. A similar point applies to Rubinstein's (1989) "email game": an imperfection in communication was introduced in order to inhibit coordination, not to add uncertainty in payoffs.

The preceding observations help clarify the theme of this chapter. However, as noted in the Introduction, a sharp separation between fundamental and strategic uncertainty may not always be possible, especially once we impose equilibrium.

For instance, in the models we study in Sections 7 and 8, the equilibrium is a unique irrespective of the information structure and the action of each agent is pinned down by her hierarchy of beliefs about the fundamentals. It follows that uncertainty about the equilibrium value of $\mathbf{K}$ can obtain in any particular state if and only if the agents lack common knowledge of fundamentals in that state, which in turns means that at least some of them face fundamental uncertainty in some state of nature.

What is more, suppose that all agents share the same fundamental, namely $\theta_i = \overline{\theta}$ for all i, and that the information of each agent i consists of a single private signal $x_i = \overline{\theta} + \epsilon_i$, where ϵ_i is idiosyncratic noise that is drawn from a commonly known distribution and that washes out at the aggregate level. Under these assumptions, $\overline{\theta}$ becomes a sufficient statistic for $\mathbf{\Omega}$, the profile of information in the cross section, and thereby also for the equilibrium profile of actions. One we impose equilibrium, no *apparent* difference therefore remains between the two types of uncertainty: predicting $\mathbf{K}$ is the same as predicting $\overline{\theta}$.

To better appreciate the role of strategic uncertainty, it is therefore useful to do one, or both, of the following: (i) consider sufficient rich information structures that make sure that the uncertainty agents face about $\mathbf{K}$ is not spanned by their uncertainty about fundamentals; (ii) momentarily drop equilibrium and, instead, contemplate the kind of higher-order reasoning that may justify certain forecasts of $\mathbf{K}$ on the basis of the common knowledge of the environment and of the rationality of the agents. Both of these paths will be explored throughout this chapter.

[q] The same statement applies to the baseline model in Lorenzoni (2009): the key mechanism in that paper rests on the uncertainty consumers face about future fundamentals (TFP) and the deviation of monetary policy from replicating flexible-price allocations. The paper contains also an extension that features dispersed information. This extension is used to justify higher fundamental uncertainty and to raise the quantitative potential of the aforementioned mechanism. It also makes a contribution to the literature we are concerned with, because of the elegant ways in which the author deals with the geography of information and the dynamics of learning. Yet, the key mechanism does not rest on the dispersion of information, and is therefore outside the scope of our chapter.

Remark 12 No matter the information structure, once equilibrium is imposed, the distinction between the two forms of uncertainty becomes blurred in the eyes of the agents inside the model: both the fundamentals and the endogenous outcomes are treated by each agent as exogenous stochastic processes. In the eyes of the outside observer, however, the two forms of uncertainty can be distinguished by the different marks they leave on the observables of the model. Dissecting this difference is an integral part of this chapter.[r]

3.2 Informational Friction

As noted in the Introduction, the present chapter views incomplete information as a device for introducing frictions in coordination, rather than as a device for introducing uncertainty about payoff-relevant fundamentals such as technology and preferences. In this context, we now proceed to clarify two distinct forms of informational friction and to point out which one is more relevant for our purposes.

Definition 9 Information is said to be *perfect* if $(\theta_i, \mathbf{S})$ is known to all agents in all states $\mathbf{S}$. If the converse is true, information is said to be *imperfect*.

Definition 10 Information is said to be *complete* if $\mathbf{\Omega}$ is known to all agents (or at least to all but a zero-measure set of agents) in all states of nature. If the converse is true, information is said to be *incomplete*.

This terminology differs somewhat from the one typically found in the related applied literature, where terms "informational friction" and "imperfect," "dispersed," or "heterogeneous" information are often used to refer to either of the two concepts defined above. The problem is that, although many papers depart from the clearly defined benchmark of perfect information, it is not always clear whether their key departure is (i) the introduction of some type of fundamental uncertainty or (ii) the relaxation of common knowledge and the associated strategic uncertainty. With the above two definitions, we seek to clarify the difference between two important types of informational frictions.[s]

Note, in particular, that perfect information imposes common knowledge of both $\mathbf{\Omega}$ and $\mathbf{\Theta}$, whereas complete information imposes only common knowledge of $\mathbf{\Omega}$.[t] For example, suppose that $\mathbf{\Omega}$ is commonly known, but $\mathbf{\Theta}$ is not. Then, information is *imperfect*

[r] This remark is consistent with the Bayesian approach to games (eg, Harsanyi, 1967-1968; Aumann, 1987; Aumann and Heifetz, 2002); but it also reinforces why we treat incomplete information and higher-order beliefs as instruments for generating first-order uncertainty about equilibrium outcomes, rather than objects of independent interest.

[s] Note that our notion of incomplete information is the same as the one found in game theory, but our notion of imperfect information is distinct. Although we regret this discrepancy, we believe that the definitions we adopt here help clarify both the language used in macroeconomics and the mechanisms featured in different applications.

[t] To be precise, our definition of complete information requires only *mutual* knowledge of $\mathbf{\Omega}$. But since $\mathbf{\Omega}$ is the entire profile of Harsanyi, mutual knowledge of $\mathbf{\Omega}$ implies common knowledge of $\mathbf{\Omega}$.

in the sense that the agents are uncertain about one another's fundamentals, and yet information is *complete* in the sense that the agents face no uncertainty about one another's information sets. Finally, note that our definition of complete information rules out private information about aggregate shocks, but allows for private information about idiosyncratic shocks; that is, complete information imposes that all agents share the same beliefs about both $\mathbf{\Theta}$ and more generally about $\mathbf{S}$, but allows each agent to know more than others about the idiosyncratic component of her fundamental.[u]

3.3 Imperfect Coordination Equals Incomplete Information

So far, the notions of imperfect coordination, strategic uncertainty, and incomplete information have been defined in a manner that allows them to be disconnected from one another. However, once equilibrium is imposed, these notions become essentially the same.

Proposition 1 (Information and Coordination) *The following are true only when information is incomplete:*

(a) *higher-order beliefs of fundamentals diverge from first-order beliefs;*

(b) *agents face uncertainty about the actions of others in any given equilibrium;*

(c) *coordination is imperfect in any given equilibrium.*

We prove this by showing that complete information rules out (a), (b), and (c). Recall first that first-order beliefs $\mathbf{B}$ are pinned down by $\mathbf{\Omega}$. If information is complete, meaning that $\mathbf{\Omega}$ is commonly known, then so is $\mathbf{B}$. The common-prior assumption then imposes that higher-order beliefs collapse to first-order beliefs. Next, note that, in any given equilibrium, the endogenous outcome $\mathbf{K}$ is given by a known function $\mathbf{K}^*$ of the realized $\mathbf{\Omega}$. If the latter is commonly known, then so is $\mathbf{K}$.[v] That is, complete information rules out strategic uncertainty in equilibrium. Finally, note that the equilibrium beliefs of $\mathbf{K}$ are pinned down by the beliefs of $\mathbf{\Omega}$, which themselves collapse to a Dirac on the true $\mathbf{\Omega}$

[u] The last point underscores that the heterogeneity, or dispersion, of information that is relevant for our purposes is the one that regards aggregate shocks, as opposed to the one that is at the center of the Mirrlees literature (aka New Public Finance): that literature introduces private information about idiosyncratic fundamentals in order to accommodate certain incentive problems, but maintains the assumption that all aggregate shocks (including the cross-sectional distribution of information sets) are commonly known, thus ruling out all the effects that are of interest to us in this chapter.

[v] If there are multiple equilibria, the function $\mathbf{K}^*$ varies across equilibria, but it is commonly known to the agents in any given equilibrium. This is true even in the case of sunspot equilibria; this case is nested by allowing ω_i to contain a publicly observed sunspot. Finally, if we consider correlated equilibria with imperfect correlation, the function $\mathbf{K}^*$ is still commonly known in any equilibrium, but now different agents have different beliefs about the realized $\mathbf{K}$, simply because they have private information about the underlying correlation devices and therefore also about $\mathbf{\Omega}$.

when information is complete. It follows that all agents share the same belief about $\mathbf{K}$, which means the friction in coordination vanishes.

These arguments prove that incomplete information is necessary for obtaining (a), (b), and (c) in Proposition 1. Sufficiency is also true, provided we consider nontrivial forms of incomplete information; this will become clear as we proceed.[w] Furthermore, one can strengthen part (b) to read as follows: a agent can face uncertainty about $\mathbf{K}$ even if she knows $\boldsymbol{\Theta}$, or even if she knows both $\boldsymbol{\Theta}$ and $\{\mathbf{B}^h\}_{h=1}^H$ up to any finite H. This illustrates the richness of the uncertainty that agent can face about one another's equilibrium actions and the associated economic outcomes when, and only when, information is incomplete.

Apart from clarifying how complete information rules out the friction we are after in this chapter, the above result also highlights the following point.

Fact Workhorse macroeconomic models typically rule out imperfect coordination, not only because they assume complete information, but also because they impose a strong solution concept.

This is due to the fact that the Rational-Expectations Equilibrium concept *itself* rules out uncertainty about the *strategies* of other agents, regardless of the information structure and the number of equilibria. It follows then that the agents can face uncertainty about the actions of others in equilibrium if and only if they do not share the same information, which is how incomplete information is defined.

Yet, this does not mean that incomplete information must be taken literally: as hinted in the Introduction, one can view incomplete information also as a substitute for relaxing the equilibrium concept.

Let us elaborate on what we have in mind. Fix a standard macroeconomic model, such as the textbook RBC model or its New-Keynesian sibling. These models deliver certain predictions about the nature of the business cycle, namely about the comovement of key economic outcomes such as employment, consumption, and investment with one another, as well as with the underlying shocks to TFP or other fundamentals. These predictions rely, not only on features of the micro-foundations such as the Frisch elasticity of labor supply and the degree of price stickiness, but also on the combination of a strong solution concept with strong informational assumptions—a combination that imposes flawless coordination in beliefs.

Macroeconomists have long debated about the elasticity of labor supply, the degree of nominal rigidity, and numerous other aspects of the micro-foundations of macroeconomic models, but have paid relatively little attention to the bite that the aforementioned combination has on the predictions of their models, and thereby on their interpretation of the data. By contrast, it is precisely this combination which we are uncomfortable with and whose bite we would like to relax. Incomplete information

[w] See also Weinstein and Yildiz (2007a).

is a vehicle for doing so and, in this sense, it is also as a substitute for relaxing the solution concept.[x]

Summing up, we have now formalized the following point, which was anticipated in the Introduction.

Fact Incomplete information is a modeling device that expands the sources of uncertainty the economic agents face about endogenous economic outcomes and that allows the expectations of these outcomes to deviate from the expectations of fundamentals.

We conclude this section by noting that, for our purposes, the uncertainty agents may face about their *own* fundamentals is of little consequence on its own right. To formalize this point, let

$$\hat{U}(k_i, \mathbf{K}, b_i) \equiv \int U(k_i, \mathbf{K}, \theta_i) db(\theta_i, \Theta),$$

where b_i is the agent's first-order belief.[y] The following is then immediate.

Proposition 2 *The equilibrium set of the original model coincides with the equilibrium set of a variant model in which the agents are always perfectly informed about their own fundamentals. In this variant games, payoffs are given by $\hat{U}(k, \mathbf{K}, \hat{\theta}_i)$, individual fundamentals by $\hat{\theta}_i \equiv b_i$, and aggregate fundamentals by $\hat{\Theta} \equiv \mathbf{B}$.*

In this sense, we can always recast any given model as one in which the agents are perfectly informed about their own fundamentals. Of course, this does not mean that uncertainty about fundamentals is irrelevant. There is a long tradition in macroeconomics that studies different kinds of fundamental uncertainty within complete-information, and often representative-agent, models. Examples include a voluminous literature on the equity premium, as well as the recent literatures on news and uncertainty shocks. The sole role of the above result is to underscore, once more, the type of uncertainty that we are interested in. For our purposes, it is not relevant per se whether each agent knows her *own* fundamental; what is relevant is whether agents have common knowledge of *aggregate* fundamentals. To isolate the latter channel, we will therefore occasionally consider examples in which every agent *i* knows her own θ_i perfectly, but not necessarily $\mathbf{\Theta}$. By contrast, the leading example in the literature shuts down heterogeneity ($\theta_i = \theta$ for all *i*)

[x] In this regard, the approach we take in this chapter can be viewed as complementary to, albeit distinct from, the approaches taken by Evans and Honkapohja (1999, 2009), Guesnerie (1992, 2008), Fuster et al. (2010), Kurz (1994, 2012), and Sargent (2008). See also the review in Woodford (2013). What is common between our approach and these alternative approaches is the desire to relax, or enrich, the stochastic structure of beliefs.

[y] Note that the definition of $\hat{U}$ treats $\mathbf{K}$ as a deterministic variable: the possibility that $\mathbf{K}$ correlates with $\mathbf{\Theta}$ in equilibrium is not relevant here. By the same token, $\hat{U}$ depends on b_i only through the marginal of b_i over θ_i; that is, it depends on the agent's belief about her *own* fundamental, but is otherwise invariant to her belief about the *aggregate* fundamentals.

and confounds the lack of common knowledge of aggregate fundamentals with uncertainty about own fundamentals.

3.4 Complete-Information Benchmark and Coordination Motives

Section 3.3 clarified the sense in which incomplete information introduces a friction in coordination. To elaborate on the observable implications of this friction and the applied value of the approach taken in this chapter, we will have to get more concrete. This is what we do in Sections 4–8. As a prelude to this, however, it is useful to work out the complete-information benchmark of our abstract framework. We do so in this section with the help of additional, simplifying, assumptions on the action space D_k and the payoff function U. We then use this benchmark to identify a feature that distinguishes the type of models studied in Sections 4–6 ("global games" and their applications) from those studied in Sections 7 and 8 ("beauty contests" and their applications), namely a feature that regards the strength of coordination motives.

With the function $\hat{U}$ defined as before, we impose the following:

Assumption 1 D_k is a closed interval in $\mathbb{R}$ and $\hat{U}$ is strictly concave and twice differentiable in k.

This simplifies the analysis by letting the action be uni-dimensional and by guaranteeing both the existence and the uniqueness of a solution to the individual's decision problem, that is, to the choice of k_i that is optimal for given beliefs of fundamentals and of the actions of others. We express this solution as follows:

$$k_i = \gamma(\mathbf{K}, b_i) \equiv \arg\max_k \hat{U}(k, \mathbf{K}, b_i).$$

Note here that agent i faces no uncertainty about $\mathbf{K}$, not only because information is complete, but also because she knows that others play a given equilibrium. That is, as anticipated, strategic uncertainty has been ruled out by the combination of a strong informational assumption with a strong solution concept.

With this in mind, we have the following result.

Proposition 3 (Complete Info) *Suppose information is complete. There exists a function Γ, which depends only on the function $\hat{\mathrm{U}}$, such that the following is true: in any equilibrium and any state of nature, $\boldsymbol{K}$ solves*

$$\mathbf{K} = \Gamma(\mathbf{K}, \mathbf{B}). \tag{8}$$

This result follows trivially from letting $\Gamma(\mathbf{K}, \mathbf{B})$ be the distribution of k that obtains when $k = \gamma(\mathbf{K}, b)$ and by noting that the function γ is essentially exogenous, in the sense that it is pinned down by the function $\hat{U}$. When the equilibrium is unique, this result means that equilibrium actions are pinned down *exclusively* by first-order beliefs of fundamentals. When, instead, there are multiple equilibria, the theory makes room for sunspots, that is, for random selections across the multiple solutions of condition (8). For any such selection, however, the equilibrium value of $\mathbf{K}$ remains tied to $\mathbf{B}$ through condition (8). Furthermore, whether the equilibrium is unique or not, the heterogeneity in the cross-section of actions is pinned down by the heterogeneity in beliefs of own fundamentals.[z] Finally, if information is perfect (which, recall, is a stronger requirement than the requirement of complete information), nothing essential changes: we merely need to replace b_i with θ_i and $\mathbf{B}$ with $\mathbf{\Theta}$.

By contrast, these restrictions are relaxed when information is incomplete: equilibrium outcomes, whether at the aggregate level or in the cross section, may depend on higher-order beliefs, which themselves may differ from first-order beliefs. What exactly this implies for the observables of any given model will be explored in the subsequent sections of this chapter. For now, we continue with the complete-information benchmark.

To further sharpen the analysis, we next impose the following:

Assumption 2 $\hat{U}_k$ depends on $\mathbf{K}$ only through the mean action K and is differentiable in it.

We can then rewrite the best response of agent i as $k_i = g(K, b_i)$, where g is a function that is differentiable in K and that is implicitly defined by the solution to $\hat{U}_k(g(K,b),K,b)=0$. Aggregating gives

$$K = G(K,\mathbf{B}),$$

where the function $G: D_k \times D_{\mathbf{B}} \to D_k$ is defined by $G(K,\mathbf{B}) \equiv \int g(K,b)d\mathbf{B}(b)$. This is essentially the same as Proposition 3 above, except for the fact that we now have a fixed-point relation in the mean action K alone as opposed to the entire distribution $\mathbf{K}$.

Since g is differentiable in K, G is also differentiable in K (as well as continuous in it). This, together with the compactness of D_k, guarantees the existence of equilibrium. To characterize the determinacy of the equilibrium, we introduce the following notions.

Definition 11 The economy exhibits *strategic substitutability* if and only if $G_K(K, \mathbf{B}) < 0$ for all $(K, \mathbf{B})$; and *strategic complementarity* if and only if $G_K(K, \mathbf{B}) > 0$ for all $(K, \mathbf{B})$.

Definition 12 Strategic complementarity is of the *weak* form if $G_K(K, \mathbf{B}) \in (0, 1)$ for all $(K, \mathbf{B})$, and of the *strong* form if $G_K(K, \mathbf{B}) > 1$ for some $(K, \mathbf{B})$ such that $K = G(K, \mathbf{B})$ and K is in the interior of D_k.

We have the following result.

Proposition 4 (Equilibrium Determinacy) *Suppose information is complete.*

(i) *An equilibrium always exists.*

(ii) *The equilibrium is unique if the economy exhibits either strategic substitutability or weak strategic complementarity.*

(iii) *There exist multiple equilibria if the economy exhibits strong strategic complementarity.*

[z] To see this, note that $k_i = \gamma(\mathbf{K}, b_i)$ can differ across two agents only if these agents have a different b_i; and because all agents have the same beliefs about $\mathbf{\Theta}$, different b_i means different beliefs about θ_i.

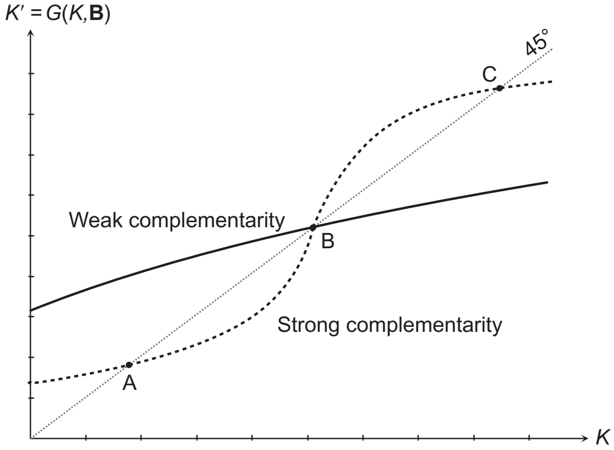


Fig. 1 Best responses with weak and strong complementarity.

We illustrate these possibilities in Fig. 1 for two examples. In both examples, we fix a particular realization of $\mathbf{B}$ and draw $G(K, \mathbf{B})$ against K. The case of weak complementarity corresponds to the solid curve, which intersects only once with the 45-degree line; point B then gives the unique equilibrium value of K, for a given value of $\mathbf{\Theta}$. The case of substitutability is similar and is thus omitted. The case of strong complementarity, on the other hand, corresponds to the dotted curve, which intersects three times with the 45-degree line; points A and C then give the two stable equilibrium outcomes of the economy, whereas point B represents an unstable equilibrium outcome.[aa]

The above categorization anticipates the two classes of models we study in the sequel: Section 4 introduces a class of models that feature strong complementarity and that have been used to study self-fulfilling crises; Section 7 turns attention to a class of models that feature weak complementarity and that are commonly used to study business cycles and asset prices. These two classes of models behave quite differently, not only under complete information, but also under incomplete information. This is because the impact of higher-order uncertainty depends on how essential coordination is in the first place. As a result, higher-order uncertainty has a more pronounced effect in the former class than in the latter one.[ab]

We conclude this section by noting how complete information ties "animal spirits" to equilibrium indeterminacy. To formalize this point, we define the notion of animal spirits as follows.

[aa] The notion of stability used here is that of iterated best responses.

[ab] In particular, the beauty contests studied in Section 7 pass Weinstein and Yildiz's (2007b) criterion of "global stability under uncertainty," which means that beliefs of sufficiently high order have a vanishing effect. This is not the case with either the static global games of Section 4, or the related dynamic games of Section 6.

Definition 13 The economy features *animals spirits* if there are two states ($\mathbf{S}_1$, $\mathbf{S}_2$), with respective beliefs of fundamentals ($\mathbf{B}_1$, $\mathbf{B}_2$) and respective outcomes ($\mathbf{K_1}$, $\mathbf{K}_2$), such that $\mathbf{B}_1 = \mathbf{B}_2$ and $\mathbf{K}_1 \neq \mathbf{K}_2$.

This definition is consistent with those found in Cass and Shell (1983) and the extant macroeconomic literature on sunspot fluctuations: it simply says that equilibrium actions can vary without any variation in the agent's beliefs of the fundamentals. For instance, if we consider the textbook RBC model, animal spirits would require that employment, investment, and consumer spending vary without any commensurate variation in the firms' and the consumers' beliefs of preferences and technologies.[ac]

Note now that the action of any given agent is always pinned down by her belief of her fundamental and her belief of the actions of others. It follows that two states can be associated with different equilibrium outcomes even when they contain the same beliefs about the fundamentals only if the two states are associated with different equilibrium beliefs about the endogenous outcomes themselves. In this sense, the notion of animal spirits is inherently tied to the notion of self-fulfilling beliefs. Finally, the following is true.

Corollary 1 *Suppose information is complete. The economy can feature animal spirits if and only if it admits multiple equilibria.*

This result highlights more generally the tendency in the profession to associate the notion of animal spirits with models that admit multiple equilibria, and to treat this notion as inconsistent with the class of unique-equilibrium DSGE models often used in the study of business cycles and monetary policy. In this context, a key point of our analysis will be to show how incomplete information blurs the distinction between multiple- and unique-equilibrium models in two complementary ways: first, by inducing equilibrium uniqueness in models that admit multiple equilibria under complete information; and second, by allowing animal spirits to obtain despite a unique equilibrium.[ad]

Remark 13 Recall that our definition of complete information requires that agents share the same information, not only with regard to the underlying fundamentals, but also with regard to *everything* else. This allows for public sunspots but rules out *imperfect* correlation devices as in Aumann's (1974; 1987) work on correlated equilibrium. If, instead, one imposes common knowledge of fundamentals but allows private information about payoff-irrelevant variables, then one can devise examples in which (i) there is a unique correlated equilibrium and (ii) this equilibrium features animal spirits because agents respond to their private information about the underlying correlation devices in this equilibrium. To see this, consider a three-player variant of the "matching pennies" game

[ac] Our notion of animal spirits should therefore not be confused with the less-common notion adopted in Lorenzoni (2009) and Barsky and Sims (2012): in these papers, "animal spirits" are defined as the noise in a public signal of future TFP, that is, as a particular shock to beliefs of fundamentals.

[ad] An alternative approach, which will not be considered in this paper, is to formalize "animal spirits" as a deviation from individual rationality. See, eg, Akerlof and Shiller (2010).

in which there is no (Nash) equilibrium in pure strategies, but there is a correlated equilibrium in which the two players coordinate on playing a jointly mixed strategy against a mixed strategy of the third player. Although this example is too contrived to be of any practical relevance, it clarifies the following point: in general, incomplete information can help sustain "animal spirits," not only by introducing higher-order uncertainty about the fundamentals, but also by sustaining strategic uncertainty even in settings that maintain common knowledge of fundamentals.

Consider now an overlapping-generations macroeconomic model in which the fundamentals are common knowledge and remain constant over time, but a different sunspot is realized in every period. Even if the sunspot is publicly observed in the period that is realized, the model is akin to a game with imperfect correlation devices, because the generation of agents that acts in any given period does not see the sunspot upon which future generations can condition their choices. This example underscores that the concept of sunspot equilibria in *dynamic* macroeconomic models is closely related to the concept of correlated equilibria, which itself allows for incomplete information. As a consequence, the sunspot volatility that has been documented in prior work does not *strictly* require equilibrium multiplicity. For a more detailed discussion of these issues, we refer the reader to Aumann et al. (1988) and Peck and Shell (1991); see also Jackson and Peck (1991) and Solomon (2003) for two applications of these ideas.

4. GLOBAL GAMES: THEORY

In this section, we restrict attention to a special case of our framework that imposes strong complementarity. We view this case as representative of a variety of multiple-equilibrium models that were developed in order to formalize the idea that macroeconomic outcomes are driven by coordination failures, self-fulfilling beliefs, and animal spirits.[ae] Of course, our framework is too stylized to capture either the rich micro-foundations or the intricate dynamics of some of these models. Nonetheless, the case studied contains two key features from the related literature: the role of coordination and the existence of multiple self-fulfilling equilibria under complete information.

The main theme of this section is to show how the introduction of incomplete information in such settings can induce a unique equilibrium, possibly one in which outcomes are pinned down merely by the underlying fundamentals. In addition, we establish the existence of a certain type of discontinuity in equilibrium outcomes as we move from the complete-information benchmark to a perturbation with arbitrarily small noise in the

[ae] See, inter alia, Azariadis (1981), Benhabib and Farmer (1994, 1999), Cass and Shell (1983), Cooper and John (1988), Cooper (1999), Diamond and Dybvig (1983), Farmer (1996), Farmer and Woodford (1997), Guesnerie (1992), Howitt and McAfee (1992), Murphy et al. (1989), Matsuyama (1991, 1995), Obstfeld (1986, 1996), Shell (1977, 1987), and Woodford (1986, 1991).

agent's information of the fundamentals. Combined, these results provide a sharp illustration of the potential sensitivity of the predictions of the theory the kind of friction we study in this chapter. They also highlight that the driving force is the uncertainty faced by agents about one another's actions, rather than their uncertainty about the underlying fundamentals. With this in mind, we then proceed to describe the distinct role that private and public information play in shaping the level of strategic uncertainty.

The analysis in this section builds heavily on the influential contributions of Carlsson and Van Damme (1993a,b) and Morris and Shin (1998, 2001, 2003). Carlsson and Van Damme's work highlighted the possible fragility of multiple equilibria by showing how small perturbations of the payoff and information structures can guarantee a unique rationalizable outcome within arbitrary two-by-two games (two players, two actions). This outcome was also shown to coincide with the one selected by the risk-dominance criterion proposed by Harsanyi and Selten (1988). Morris and Shin's work extended this type of result to models that were better suited for macroeconomics and finance, thus paving the way to a large applied literature; it also expanded the theoretical foundations.

Remark 14 The closely related contributions of Chamley (1999), Frankel and Pauzner (2000), and Burdzy et al. (2001) were made at roughly the same time as those of Morris and Shin. An important precedent to all these works is Rubinstein (1989), which first highlighted the fragility of coordination to perturbations of common knowledge.

Remark 15 The term "global game" was introduced by Carlsson and Van Damme (1993b) within the context of two-by-two games that had the following key features: first, information was incomplete; second, types could be ordered in such a way that one action is dominant for sufficiently low types and another action is dominant for sufficiently high types. In the subsequent literature, the number of players and actions is often larger, and strategic complementarity is often imposed, yet appropriate versions of the aforementioned key features are maintained. For the purposes of our paper, we restrict attention to a particular class of games that are often referred to in the literature as "games of regime change."

4.1 Setup

Throughout this section, we impose that actions are binary, that there is no heterogeneity in fundamentals, and that payoffs take a particularly simple form.

Assumption 3 (Global Game) $D_k = \{0, 1\}$ and $D_\theta = \mathbb{R}$. Furthermore, θ_i is identical across i and is henceforth denoted simply by θ. Finally,

$$u_i = U(k_i, K, \theta) = \begin{cases} k_i(b-c) & \text{if } K \geq \theta \\ -k_i c & \text{if } K < \theta \end{cases}$$

where K is the average action (equivalently, the fraction of the population that chooses $k = 1$) and where b and c are known positive scalars, and b > c.

The key feature of this specification is that it represents environments with strong strategic complementarity. As long as the agent faces no uncertainty about either K or θ, her best response is given by $k = G(K, \theta)$, where

$$G(K,\theta) \equiv \arg \max_{k \in \{0,1\}} U(k,K,\theta) = \begin{cases} 1 & \text{if } K \geq \theta \\ 0 & \text{if } K < \theta \end{cases}$$

It is therefore as if the best response of an agent has infinite slope in a neighborhood of $K = \theta$.

Remark 16 The definition of strong complementarity that we introduced in Section 3.4 requires that G is continuous and that it has slope higher than one in a point where it crosses the 45^o line. Here, G is discontinuous, so the earlier definition does not apply anymore, but the essence remains the same. Also, note the following slight abuse of notation: before, G was a function of $\mathbf{B}$, the entire cross-sectional distribution of first-order beliefs; now, it is a function of θ, because the latter happens to be sufficient statistic for $\mathbf{B}$ under the specification we consider in this section.

4.2 Interpretation

To ease the transition to applications, and following Angeletos et al. (2007), we interpret the aforementioned payoff specification as a "game of regime change."

There are two possible regimes, the status quo and an alternative. Each agent can choose between an action that is favorable to the alternative regime and an action that is favorable to the status quo. We henceforth refer to these actions as, respectively, "attack" and "not attack." We denote the regime outcome with $R \in \{0, 1\}$, where $R = 0$ represents the survival of the status quo and $R = 1$ represents its collapse. We similarly denote the action of an agent with $k_i \in \{0, 1\}$, where $k_i = 0$ represents "not attack" and $k_i = 1$ represents "attack." Next, we normalize the payoff from not attacking to zero and let the payoff from attacking be $b-c > 0$ if the status quo is abandoned and $-c < 0$ - otherwise. This means that it is individually optimal to attack the status quo if and only if the latter is expected to collapse. Finally, we assume the status quo is abandoned ($R = 1$) if and only if

$$K \geq \theta,$$

which means that θ identifies the minimal size of attack that is needed for the status quo to be overturned. It then follows that the payoff of the agent can be expressed as in Assumption 3.

One may wish to interpret the above game as a model of revolutions and political change.[af] Perhaps more interestingly for macroeconomics, the above game can be used to capture the role of coordination in the context of financial crises.

[af] See, eg, Atkeson (2000) and Edmond (2013) for related interpretations.

Consider, in particular, the literature on self-fulfilling currency attacks, such as Obstfeld (1996). In this context, the players can be interpreted as speculators choosing whether to attack a currency peg or not; θ represents the resources (eg, reserves) that are available to the central bank for defending the peg, or more generally, the ability and the willingness of the policy maker to withstand a speculative attack; c represents the interest-rate differential between domestic and foreign assets, or other costs suffered by a speculator when short-selling the domestic currency; b represents the gains of an attacking speculator in the event of devaluation; and a "regime change" occurs when a sufficiently large mass of speculators attacks the currency, forcing the central bank to abandon the peg.

Similarly, in models of self-fulfilling bank runs, such as Diamond and Dybvig (1983), θ may represent the liquidity of a bank or of the financial system at large. A "regime change" occurs once a sufficiently large number of depositors decides to withdraw their deposits, forcing the bank to suspend its payments.

In models of self-fulfilling debt crises, such as Calvo (1988), θ may represent the long-run profitability of a firm or the growth potential of a country, and "regime change" occurs when the firm's/country's creditors decline to roll over its outstanding short-term debt because they fear that other creditors will refuse to roll over, forcing the firm/country into default.

Finally, regime change can also mean the "big push" in Murphy et al. (1989), that is, a situation where the incentives of an agent to switch from one technology to another, or to leave the village for the city, or to enter a particular industry, depend crucially on how many other agents are doing the same.

Remark 17 The aforementioned papers assume complete information. The more recent literature that incorporates incomplete information in such applications is discussed in the next section. Also, the applications we have in mind are intrinsically dynamic. One may thus question whether it is appropriate to model them as one-shot games. We will revisit this issue at the end of the next section.

4.3 Complete Information and Multiple Equilibria

Suppose for the moment that information is complete. Without any serious loss, suppose further that information is perfect, meaning that the true realization of θ is known to all agents in all states of Nature.

Let $\theta_L \equiv 0$ and $\theta_H \equiv 1$. For $\theta \leq \theta_L$, the fundamentals are so weak that the regime is doomed with certainty and the unique equilibrium has every agent attacking. For $\theta > \theta_H$, the fundamentals are so strong that the regime can survive an attack of any size and the unique equilibrium has every agent not attacking. For intermediate values, $\theta \in (\theta_L, \theta_H]$, the regime is sound but vulnerable to a sufficiently large attack and there are multiple equilibria sustained by self-fulfilling expectations. In one equilibrium, individuals expect everyone else to attack, so they then find it individually optimal to attack, the status quo is abandoned and expectations are vindicated. In another, individuals expect no one else to attack, they thus find it individually optimal not to attack, the status quo is spared and expectations are again fulfilled.

Proposition 5 (Multiple Equilibria) *Suppose information is complete (θ is common knowledge). When $\theta \leq \theta_L$, the unique equilibrium outcome is $K = R = 1$. When $\theta > \theta_H$, the unique equilibrium outcome is $K = R = 0$. And when $\theta \in (\theta_L, \theta_H]$, both $K = R = 0$ and $K = R = 1$ are equilibrium outcomes.*

The interval $(\theta_L, \theta_H]$ thus represents the set of "critical fundamentals" for which multiple equilibria are possible under complete information. Each equilibrium is sustained by different self-fulfilling expectations about what other agents do.

Under the assumed payoff function, the two equilibria can be readily ranked: the equilibrium with $K = R = 1$ is Pareto superior to that with $K = R = 0$. One can thus associate the latter equilibrium with a "coordination failure." More generally, which equilibrium represents a coordination failure varies depending on the context of interest. In the context of self-fulfilling currency attacks, for example, a coordination failure is the no-attack equilibrium from the perspective of foreign speculators, whereas it is the attack equilibrium from the perspective of domestic agents. Regardless of which perspective is adopted, however, the notion of coordination failure is associated with a particular selection across multiple equilibria.

In Sections 4.4 and 4.5, we will show how allowing information to be incomplete can remove the indeterminacy of the equilibrium. In fact, we will show that this can be true even if the noise in the available information is arbitrarily small. While this implies that a realistic perturbation of standard models can potentially destroy the *traditional* view of coordination failures and animal spirits, we will argue that these notions can be meaningfully resurrected even in models with a unique equilibrium.

4.4 Incomplete Information and Unique Equilibrium

Following Morris and Shin (1998, 2001), we restrict the stochastic structure as follows. Nature draws θ from a uniform distribution over the entire real line.[ag] Conditioning on θ, Nature draws a private signal for each agent i. This signal is given by

$$x_i = \theta + \sigma\epsilon_i,$$

[ag] This specification of the common prior about θ is only for simplicity. For the results that follow, it suffices to consider any smooth prior over an interval that strictly contains the critical region and to let σ, the noise in the signal, be small enough. What the assumption of a uniform (or uninformative) prior does is to guarantee that uniqueness obtains for any σ, not just for σ small enough. The role of precise priors, or the related role of public information, is discussed later.

where ϵ_i is an idiosyncratic noise term and $\sigma > 0$ parameterizes the level of noise. The noise is independent of θ and is drawn from a smooth distribution over the real line, with a strictly increasing c.d.f. denoted by Φ. These facts are common knowledge, but the realizations of θ and the signals are not; instead, the information set, ω_i, of agent i is simply the signal x_i. Without any loss, we will assume that an agent attacks whenever she is indifferent between attacking and not attacking.[ah]

Note that the scalar σ parameterizes how informed each agent is about the realization of θ. When σ is *exactly* zero, the model reduces to the complete-information benchmark. When, instead, σ is positive but small enough, every agent is nearly perfectly informed, and we have a seemingly tiny perturbation in the exogenous primitives of the environment. One may have expected that such a tiny change in the assumptions of the model would imply a tiny difference in its predictions. This turns out not to be the case: the predictions of the theory are discontinuous at $\sigma = 0$.

Proposition 6 (Morris-Shin) *There is a unique equilibrium for all $\sigma > 0$. In this equilibrium, the size of the attack K is monotonically decreasing in θ, and regime change occurs ($R = 1$) if and only if $\theta < \theta^*$, where $\theta^* = 1 - c/b \in (0, 1)$.*

This result highlights the sharp discontinuity mentioned above: a seemingly tiny perturbation in a model's assumptions regarding the agents' information implies a huge difference in the model's predictions. This discontinuity seems troubling as it calls into question the insights and policy lessons obtained in the large multiple-equilibria literature of the 80s and 90s.

We prove this result in Section 4.5. We then proceed to elaborate on its robustness, its theoretical underpinnings, and its implications for applied work.

4.5 Proof of Equilibrium Uniqueness

The proof is based on the procedure of iterated deletion of dominated strategies. For any $\hat{x} \in [-\infty, +\infty]$, let $K_{\hat{x}}(\theta)$ denote the size of aggregate attack when every (or almost every) agent attacks if and only if $x \leq \hat{x}$. Next, define the function

$$V(x,\hat{x}) = \mathbb{E}[\ \ U(1, K_{\hat{x}}(\theta), \theta) - U(0, K_{\hat{x}}(\theta), \theta)\ \ |\ \ x\ \ ].$$

This is the difference in utility between attacking and not attacking for an agent who has private information x and expects the other agents to attack if and only if their signals fall below $\hat{x}$.

Let us determine V. First, note that, when other agents follow a threshold strategy with threshold $\hat{x}$, the resulting size of the attack is given by the following:

[ah] This assumption resolves an indeterminacy that obtains in a zero-probability event, namely when the private signal x takes the particular value x^* characterized in the proof of Proposition 6 below.

$$K_{\hat{x}}(\theta) = Prob(x \leq \hat{x}|\theta) = Prob(\theta + \sigma\epsilon \leq \hat{x}|\theta) = Prob\left(\epsilon \leq \frac{1}{\sigma}(\hat{x} - \theta)\right) = \Phi\left(\frac{1}{\sigma}(\hat{x} - \theta)\right).$$

By implication, the regime collapses ($R = 1$) if and only if $\theta < \hat{\theta}$, where $\hat{\theta} = \hat{\theta}(\hat{x})$ is the unique solution to $K_{\hat{x}}(\hat{\theta}) = \hat{\theta}$, or equivalently the inverse of the following:

$$\hat{x} = \hat{\theta} + \sigma\Phi^{-1}(\hat{\theta}).$$

It follows that the agent with signal x attaches the following probability to the event of regime change:

$$Prob(R = 1|x) = Prob(\theta \leq \hat{\theta}|x) = Prob(\theta - x \leq \hat{\theta} - x) = Prob\left(\epsilon \geq \frac{1}{\sigma}(x - \hat{\theta})\right)$$
$$= 1 - \Phi\left(\frac{1}{\sigma}(x - \hat{\theta})\right).$$

We can express the payoff V as follows:

$$V(x, \hat{x}) = b - b\Phi\left(\frac{1}{\sigma}[x - \quad \hat{\theta}(\hat{x})]\right) - c,$$

where $\hat{\theta} = \hat{\theta}(\hat{x})$ is the aforementioned unique solution to $K_{\hat{x}}(\hat{\theta}) = \hat{\theta}$.

Before we proceed, we wish to emphasize that the specific functional form of V is not essential. As it will become clear below, the result is driven by the monotonicity and continuity properties of the function V and of the related function h that we introduce below.

With this point in mind, note $\hat{\theta}$ is increasing in $\hat{x}$. It follows that $V(x, \hat{x})$ is increasing in $\hat{x}$: The more aggressive the other agents are, the higher the expected payoff from attacking. Moreover, $V(x, \hat{x})$ is decreasing in x: The higher the value of the private signal, the lower the expected payoff from attacking.

Next, note that $V(x, \hat{x})$ is continuous in x and satisfies $V(x, \hat{x}) \to b - c > 0$ as $x \to -\infty$ and $V(x, \hat{x}) \to -c < 0$ as $x \to +\infty$. We can thus define a function h such that $x = h(\hat{x})$ is the unique solution to $V(x, \hat{x}) = 0$. Because $V(x, \hat{x})$ is continuous in both arguments, decreasing in x, and increasing in $\hat{x}$, $h(\hat{x})$ is continuous and increasing in $\hat{x}$.

The function h defined above summarizes best responses within the set of monotone strategies: assuming that agents $j \neq i$ attack if and only if $x_j \leq \hat{x}$, agent i finds it optimal to attack if and only if $x_i \leq h(\hat{x})$. By the same token, the fixed points of h identify the set of monotone equilibria: if there is an equilibrium in which an agent attacks if and only if $x < x^*$, then x^* solves $x^* = h(x^*)$.

Finally, note that $x^* = h(x^*)$ if and only if $V(x^*, x^*) = 0$. Using the definition of V, and letting $\theta^* = \hat{\theta}(x^*)$, we have that $V(x^*, x^*) = 0$ if and only if $\theta^* = 1 - c/b$. It follows that there exists a *unique* threshold x^* such that $x^* = h(x^*)$. This proves that there exists a unique equilibrium in monotone strategies.

We next prove that there is no other equilibrium. In fact, we prove a stronger result: the only strategy that survives iterated deletion of dominated strategies is the monotone equilibrium strategy identified by the aforementioned threshold.

Construct a sequence $\{\underline{x}_j\}_{j=0}^{\infty}$ by $\underline{x}_0 = -\infty$ and $\underline{x}_j = h(\underline{x}_{j-1})$ for all $j \geq 1$. In particular, letting $\underline{\theta}_{j-1}$ be the solution to

$$\underline{x}_{j-1} = \underline{\theta}_{j-1} + \sigma\Phi^{-1}(\underline{\theta}_{j-1}), \tag{9}$$

we have

$$V(x, \underline{x}_{j-1}) = b - b\Phi\left(\frac{1}{\sigma}(x - \underline{\theta}_{j-1})\right) - c$$

and thus

$$\underline{x}_j = \underline{\theta}_{j-1} + \sigma\Phi^{-1}\left(1 - \frac{c}{b}\right). \tag{10}$$

Hence, $\underline{x}_0 = -\infty$, $\underline{\theta}_0 = 0$, $\underline{x}_1 = \frac{1}{\sqrt{\alpha_x}}\Phi^{-1}\left(\frac{b-c}{b}\right)$, and so on. Clearly, the sequence $\{\underline{x}_j\}_{j=0}^{\infty}$ is increasing and bounded above by x^*. Hence, the sequence $\{\underline{x}_j\}_{j=0}^{\infty}$ converges to some $\underline{x}$. By continuity of h, the limit $\underline{x}$ must be a fixed point of h. But we already proved that h has a unique fixed point. Hence, $\underline{x} = x^*$.

Next, construct a sequence $\{\bar{x}_j\}_{j=0}^{\infty}$ by $\bar{x}_0 = +\infty$ and $\bar{x}_j = h(\bar{x}_{j-1})$ for all $j \geq 1$. Note that this sequence is decreasing and bounded below by x^*. Hence, the sequence $\{\bar{x}_j\}_{j=0}^{\infty}$ converges to some $\bar{x}$. By continuity of h, $\bar{x}$ must be a fixed point of h. But we already proved that h has a unique fixed point. Hence, $\bar{x} = x^*$.

What is the meaning of these sequences?

Consider $\underline{x}_1$. If nobody else attacks, the agent finds it optimal to attack if and only if $x \leq \underline{x}_1$. By complementarity, if some people attack, the agent finds it optimal to attack *at least* for $x \leq \underline{x}_1$. That is, for $x \leq \underline{x}_1$, attacking is *dominant*. Next, consider $\underline{x}_2$. When other agents attack if and only if it is dominant for them to do so, that is, if and only if $x \leq \underline{x}_1$, then it is optimal to attack if and only if $x \leq \underline{x}_2$. By complementarity, if other agents attack at least for $x \leq \underline{x}_1$, then it is optimal to attack *at least* for $x \leq \underline{x}_2$. That is, attacking becomes dominant for $x \leq \underline{x}_2$ after the second round of deletion of dominated strategies.

More generally, for any $j \geq 1$, we have that attacking becomes dominant $x \leq \underline{x}_j$ after the j round of deletion of dominated strategies. Hence, $\{\underline{x}_j\}_{j=0}^{\infty}$ represents iterated deletion of dominated strategies "from below." Symmetrically, $\{\bar{x}_j\}_{j=0}^{\infty}$ represents iterated deletion of dominated strategies "from above."

To recap, the only strategies that survive j rounds of iterated deletion of dominated strategies are functions k such that $k(x) = 1$ for all $x \leq \underline{x}_j$ and $k(x) = 0$ for $x > \bar{x}_j$. The value of $k(x)$ is still "free" at the jth round only for $x \in (\underline{x}_j, \bar{x}_j)$. But we already proved that both $\underline{x}_j$ and $\bar{x}_j$ converge to x^* as $j \to \infty$. Hence, in the limit, the only strategy that survives is the function k such that $k(x) = 1$ for $x \leq x^*$ and $k(x) = 0$ for $x > x^*$.

We have thus proved that there exists a unique rationalizable strategy, and hence a fortiori also a unique equilibrium. In this equilibrium, an agent attacks if and only if $x \leq x^*$, and the status quo collapses if and only if $\theta \leq \theta^*$, where the pair (x^*, θ^*) is the unique fixed point to the iteration defined in (9) and (10). From this last fact, it is then immediate that $\theta^* = \frac{b-c}{b}$.

Remark 18 The property that the best-response function h is increasing and admits a unique fixed point means, in effect, that the incomplete-information game under consideration features weak complementary. But recall that the complete-information counterpart features strong complementarity. It follows that the introduction of incomplete information has effectively transformed the game from one of strong complementarity to one of weak complementarity, which in turn helps explain the uniqueness result. See Vives (2005), Van Zandt and Vives (2007), and Mathevet (2010) for a further discussion of this point, and for complementary analyses of how the uniqueness result can be understood via either super-modular methods (Vives, 2005; Van Zandt and Vives, 2007) or a contraction mapping argument (Mathevet, 2010).

4.6 The Role of Public Information

In the preceding analysis, we assumed that the common prior about θ was uninformative (ie, an improper uniform over entire real line). If the prior is not uniform, uniqueness obtains for σ small enough, but may not obtain otherwise. A similar property applies if we introduce a public signal: to the extent that there is sufficient public information, multiplicity survives the introduction of dispersed private information.[ai]

To illustrate this point consider the following modification of the information structure. The private signal is now given by

$$x_i = \theta + \sigma_\epsilon \epsilon_i,$$

where ϵ_i is drawn from a standardized Normal distribution (with mean 0 and variance 1). Similarly, the public signal is given by

$$z = \theta + \sigma_\zeta \zeta,$$

where ζ is an aggregate noise term that is also drawn from a standardized Normal distribution. The scalars σ_ϵ and σ_ζ parameterize the level of noise in the two signals.

One can establish the following:

[ai] In terms of strategic uncertainty, there is no difference between a public signal and the common prior: when studying the hierarchy of beliefs and/or solving for the equilibrium strategies, one can always re-cast a public signal as part of the common prior. The distinction between the prior and a public signal is therefore useful only for applied purposes.

Proposition 7 (Public Info) *The equilibrium is unique if and only if* $\sigma_\epsilon \leq \sqrt{2\pi}\sigma_\zeta^2$.

The distinct nature of the two types of information can be further illustrated by comparing the limit as $\sigma_\epsilon \to 0$ (for fixed $\sigma_\zeta > 0$) to that as $\sigma_\zeta \to 0$ (for fixed $\sigma_\epsilon > 0$). In either limit, the beliefs of every agent about θ converge in probability to the true θ. Both limits therefore approach "perfect knowledge" with regard to fundamentals, and are indistinguishable from the perspective of how informed the agents are about the fundamentals. Yet, the two limits look very different from the perspective of how well the agents can predict one another's actions. Because of this, they also make very different predictions about economic behavior.

Proposition 8 (Limits)

(i) *In the limit as* $\sigma_\epsilon \to 0$ *for given* $\sigma_\zeta > 0$, *the probability of regime change converges to 1 for all* $\theta < \theta^*$ *and to zero for all* $\theta > \theta^*$, *where* $\theta^* = 1 - c/b$.

(ii) *Pick an arbitrary compact subset of the critical region,* $A \subset (\theta_L, \theta_H]$. *In the limit as* $\sigma_\zeta \to 0$ *for given* $\sigma_\epsilon > 0$, *there is an equilibrium in which the probability of regime change converges to 1 for all* $\theta \in A$, *as well as an equilibrium in which the probability of regime change converges to 0 for all* $\theta \in A$.

Part (i) recasts Proposition 7 as a particular limit in which private information is infinitely precise *relative* to public information. Part (ii), by contrast, recasts the complete-information benchmark as the limit in which it is the relative precision of public information that is infinite. In combination, these two results therefore underscore the distinct role played by the two types of information in shaping the equilibrium beliefs that agents form about one another's actions and in determining their ability to coordinate.

Remark 19 Recall the function h from the proof of Proposition 6 in Section 4.5; that function described best responses within the set of monotone strategies. In that proof, we saw that h admitted a unique fixed point. We then commented that this property meant that the introduction of private information had transformed the game from one of strong complementarity to one of weak complementary. The role of public information is the exact opposite: when public information is sufficiently precise relative to private information, the function h admits multiple fixed points. That is, sufficiently precise public information brings back strong complementarity.[aj]

4.7 Intuition and Take-Home Lessons

A deeper understanding of the precise logic behind the preceding results requires a review of important advances in game theory, an endeavor which is beyond the scope of this paper. We refer the interested reader to Rubinstein (1989) for an early seminal contribution that highlighted the fragility of coordination to perturbations of common

[aj] These points can be inferred from the proof of Proposition 7 in Appendix, noting that the function G in that proof is a transformation of the function h from the space of strategies (as indexed by the threshold x^*) to the space of regime outcomes (as indexed by θ^*).

knowledge; to Monderer and Samet (1989, 1996) and Kajii and Morris (1997a,b) for what it means to have "approximate" common knowledge and the related robustness of equilibria to incomplete information; to Morris et al. (1995) for the contagious effects of higher-order uncertainty; to Morris and Shin (1997, 2003), Frankel et al. (2003), and Mathevet and Steiner (2013) for extensions of the global-games uniqueness result to richer settings than the simple one we have studied in this section; to Mathevet (2010) for a variant proof based on a contraction mapping argument; to Vives (2005) and Van Zandt and Vives (2007) for related techniques from supermodular games; to Weinstein and Yildiz (2007a) for a powerful result that we discuss briefly in Section 4.8; and to Morris et al. (2016) for the importance of "rank beliefs" (the probability the players assign to their signal being higher than that of their opponents) and for the common belief foundations of global games. For our purposes, it suffices to emphasize the following key intuitions.

First, note that equilibrium imposes that agents know one another's *strategies*, that is, they know the mappings from their information sets (or Harsanyi types) to their actions. If we assume that all agents share the same information, then this imposes that all agents face no uncertainty about their actions. The absence of this kind of strategic uncertainty is conducive to multiple equilibria: it is "easy" to coordinate on one of many equilibrium actions when the agents are confident that other agents will do the same. But once information is incomplete, the agents may face uncertainty about one another's actions, and this type of uncertainty can hinder coordination. It follows that the determinacy of the equilibrium hinges on the level of strategic uncertainty: the higher the level of strategic uncertainty, the harder to sustain multiple equilibria.

Next, note that the level of strategic uncertainty is not necessarily tied to the level of fundamental uncertainty: the uncertainty an agent faces about the beliefs and actions of other agents has to do more with the heterogeneity of the information and the associated higher-order uncertainty, and less with the overall level of noise in the observation of the fundamentals. In fact, when private information becomes more precise, the uncertainty that an agent i faces about the fundamentals necessarily decreases, yet it is possible that her uncertainty about beliefs and actions of any other agent j increases. This is because an increase in the precision of private information implies that the beliefs and actions of agent j become more anchored to her own private information, which is itself unknown to agent i. This anchoring effect in turn explains why private information can exacerbate higher-order uncertainty and thereby hinder coordination.

This intuition can be formalized as follows. First, note that, when information is complete, the equilibrium belief of any agent about K is a direct measure of the realized value of K. That is, agents are perfectly informed about the size of the attack in equilibrium, irrespective of the equilibrium selected. Next, note that, in the diametrically opposite scenario where an agent is completely agnostic about the size of the attack, her belief about K is uniform over the $[0, 1]$ interval. Finally, consider what happens under incomplete information. Let σ_ϵ be small enough so that the equilibrium is unique and consider the "marginal" agent, that is, the type who is indifferent between attacking and not attacking in equilibrium. Morris and Shin (2003) show that the following is true.

Proposition 9 *In the limit as private information becomes infinitely precise (namely, $\sigma_\epsilon \to 0$ for given $\sigma_\zeta > 0$), the marginal agent's belief about* K *converges to a uniform distribution over* [0, 1]. *That is, the marginal agent learns perfectly the fundamental in the limit, and yet she remains completely uninformed about the actions of others.*

The sharpness of this last result hinges on focusing on the beliefs of the marginal agent, or of agents whose distance from the marginal agent vanishes as $\sigma_\epsilon \to 0$, as opposed to agents sufficiently far from the threshold x^*. Nevertheless, the combination of this result with the preceding observations helps explain why private information contributes to equilibrium uniqueness, whereas public information is conducive to multiplicity.

To summarize, in settings with coordination motives, information plays a dual role: it shapes the beliefs of each agent, not only about the exogenous payoff-relevant fundamentals, but also about the endogenous actions of the others. In the preceding result, it is the second channel that matters most. Private and public information are similar vis-a-vis the first function, but are distinct vis-a-vis the second function.

We close this section by discussing the nature of private and public signals.

For certain applied purposes, one may seek a rather literal interpretation of these signals. For instance, the private signal x_i may correspond to the proprietary information of a hedge fund, the "local" knowledge that a bank may have about the entire financial network, or the information a firm or consumer extracts about the aggregate economy from its own transactions. Similarly, the public signal z may be a proxy for financial news in the media, announcement or choices made by policy makers, or market signals such as prices. We discuss some of these interpretations, and the additional lessons they can lead to, in Section 5.

A literal interpretation of the signals is therefore useful to the extent that the researcher can envision empirical counterparts to them, and potentially even measure them. It is also useful if one wishes to study questions relating either to the collection and aggregation of information or to the welfare effects of the signals disseminated by markets, the media, and policy makers—issues that we address latter on.

That said, the mapping from the theory to the real world may not be as easy as suggested by the sharp dichotomy between private and public information employed above. For instance, suppose that the noise in the private signals happens to be the sum of two components, one idiosyncratic and one aggregate: $x_i = \theta + u + \epsilon_i$, where u is the aggregate noise and ϵ_i is the idiosyncratic noise. Suppose further that we introduce a public signal of the aggregate noise u, as opposed to a public signal of the fundamental θ. Then, this public signal will only increase the reliance of individual decisions on the private signals, which in turn may contribute to higher strategic uncertainty.

The above example may appear to be esoteric, but once one starts thinking about applications, the distinction between "fundamentals" and "noise" can get fussy. For instance, in the context of asset markets, should one think of shocks to discount factors as "noise" or as "fundamentals"? This question that can be meaningfully addressed only within the context of a fully-specified micro-founded model.

Finally, a literal interpretation of the signals is too narrow for some of the issues we have in mind. In the real world, and especially in the context of macroeconomic phenomena, it seems hard to measure, or even comprehend, all the ways in which economic agents collect and exchange information about either their idiosyncratic circumstances or the overall state of the economy. More generally, it is unclear, at least to us, why the most useful explanations of real-world phenomena are those based on models that rule out any friction either in how agents form beliefs about the relevant economic outcomes or in how they coordinate their behavior—which is what workhorse macroeconomic models typically do. From this perspective, the signal structures we employ throughout this chapter are ultimately modeling devices that permit the researcher to "open up the black box" of how agents form expectations about endogenous economic outcomes and how they coordinate their behavior. In our view, this basic point is central to understanding the precise applied value, and the testable implications, of incorporating incomplete information in macroeconomics.[ak]

4.8 Extensions and Additional Lessons

The preceding analysis has focused on a rather narrow class of binary-action games that can be described as games of regime change. Morris and Shin (2003) extend the global-games uniqueness result to a broader class of binary-action games, allowing for a more flexible form of payoff interdependence. Frankel et al. (2003) provide a generalization of the global-games uniqueness result to settings where the action belongs to an arbitrary, finite set, while maintaining strategic complementarity (appropriately defined in terms of supermodularity). Guimaraes and Morris (2007) consider an application with a continuous action (a portfolio choice). Goldstein and Pauzner (2005) consider an example that relaxes the assumption of strategic complementarity, in exchange for a weaker single-crossing condition[al] and the assumption that both the prior about the fundamentals and the distribution of the idiosyncratic noise are uniform.

[ak] Bergemann and Morris (2013) corroborate this point in the related class of games studied in Section 7, by showing that the equilibrium outcomes that can be sustained by any Gaussian information structure can be replicated by those that are sustained under the assumed private and public signals. For two alternative approaches within the global-games literature, see Cornand and Heinemann (2009) and Izmalkov and Yildiz (2010). The first paper studies a model with multiple group-specific signals; the second uses a heterogeneous-prior specification of the belief hierarchy.

[al] This single-crossing condition is the following: $u(K, x)$ crosses zero once, going from negative to positive, as K varies from 0 to 1, where $u(K, x)$ is the net payoff from attacking when the agent receives signal x and knows K.

Morris and Shin (2003) also establish the result in Proposition 9, namely that the marginal agent is agnostic about the size of the attack in the limit as private information becomes infinitely precise—a property to which they refer as "Laplacian belief." Whether this property is realistic or not is open for debate. Morris and Shin (2003) argue that this is a "plausible" restriction the theorist can impose in order to sharpen the predictions he can deliver. By contrast, Atkeson (2000), Angeletos and Werning (2006), Angeletos et al. (2006, 2007), and others have argued that, for many applications of interest, the endogeneity of the available information may lead to situations where this property is violated: market signals such as prices, and past economic outcomes, may endogenously limit the uncertainty that agents, including the marginal agent, face about one another's actions. More on this in the next section.

The aforementioned works, as well as the applications considered in the next section, limit attention to specific payoff and information structures. Weinstein and Yildiz (2007a) aim for greater generality and obtain a revealing result: for any given game and any given rationalizable action *a* of any given type in the Universal Type Space, there is a nearby game in which the given action *a* becomes the unique rationalizable action for a perturbed version of the original type.[am] A corollary of this is that equilibrium multiplicity can be viewed as a knife-edge situation, which can always be eliminated by considering an arbitrarily small perturbation of the payoff and information structures.

At first sight, Weinstein and Yildiz's result may appear to deliver a deathblow to multiple-equilibria models and their applications: multiplicity is degenerate! However, the actual meaning of the result is different. The result applies to *every* rationalizable action of the original game. It follows that, whenever we have a model with multiple equilibria, we can find small perturbations—indeed, open sets of such perturbations—that select *any* of the original equilibrium outcomes as the unique rationalizable outcome.

For applied purposes, this means the following. In models that admit multiple equilibria, the inability of theorist to tell which equilibrium is selected should be interpreted as follows: the theorist is unable to reach sufficiently sharp predictions on the basis of the assumptions he has made in his model (that is, on the basis of his prior knowledge about the environment). While Weinstein and Yildiz (2007a) result suggests that this type of multiplicity is fragile, the inability to make sufficiently sharp prediction about the equilibrium remains: but it now concerns lack of knowledge about the details of the information structure rather than lack of knowledge about which equilibrium is played.[an]

What lesson can be drawn? The global-games methodology is not a panacea for getting rid of multiple equilibria, or for giving policy makers the satisfaction of sharp policy

[am] What "nearby" and "perturbed" mean can be delicate; see Weinstein and Yildiz (2007a) for details.

[an] To put it differently, the common thread is that the variation in the observable outcomes of the model is not panned by either the underlying payoff relevant fundamentals or the agent's beliefs about them. In the original, multiple-equilibrium, model, the residual variation is attributed to a pure sunspot. In the perturbed, unique-equilibrium, model, the residual variation is attributed to higher-order uncertainty and rich "noise" in the information structure.

advice. Instead, the applied value of the global-games uniqueness result rests on elucidating the mechanics of coordination and on highlighting the importance of information, and communication, for the questions of interest. We try to give concrete examples of this value later in Section 5.

5. GLOBAL GAMES: APPLICATIONS

In this section, we review recent applications of global games to macroeconomics and finance. Many of these applications can be classified into two broad categories. The first one treats the information structure as exogenous, makes assumptions that guarantee equilibrium uniqueness, and uses this to shed new light on phenomena such as bank runs, financial crises, and business cycles and on related policy questions. The second one endogenizes the information structure in manners that seem empirically relevant, studies how this can bring back multiple equilibria, and sheds further light on applied questions.

Sections 5.1–5.6 review the first set of applications. Sections 5.7–5.9 turn to the second set. Section 5.10 mentions additional research that does not necessarily fit in the previous two categories.

5.1 Currency Crises and Bank Runs

In the paper that popularized the global-games approach, Morris and Shin (1998) introduced incomplete information in the currency-attack model of Obstfeld (1996). Although multiple equilibria exist when the fundamentals (such as reserves) are common knowledge, a unique equilibrium is obtained by adding a small amount of idiosyncratic noise in the speculators' information about the fundamentals. The size of the attack and the devaluation outcome no longer depend on sunspots but may exhibit a strong nonlinearity (or near-discontinuity) with respect to the fundamentals. A large attack can thus be triggered by small changes in the underlying fundamentals, helping reconcile the uniqueness of the equilibrium with the fact that speculative attacks are abrupt and often happen without a significant change in measured fundamentals. Moreover, policy analysis can now be conducted without the difficulties associated with multiple equilibria and arbitrary, ad hoc, selections. The paper shows how a marginal increase in a "tax" on capital outflows or in domestic interest rates can curtail a speculative attack, a policy conclusion that would not be have been obtained in the presence of multiple equilibria.[ao]

Turning to the context of bank runs, Goldstein and Pauzner (2005) "globalize" Diamond and Dybvig (1983) and proceed to study the implications for the design of the optimal deposit contract. To understand the contribution of this paper, note first that the characterization of optimal contract in Diamond and Dybvig (1983) is based on the assumption that a self-fulfilling run never takes place. This entails a particular equilibrium selection, which is possible in the original, complete-information, setting of Diamond

[ao] The original paper had a mistake in the characterization of the comparative statics of the equilibrium with respect to the aforementioned tax. See Heinemman (2000) for the correction.

and Dybvig, but not once incomplete information is introduced. Instead, the unique equilibrium now makes self-fulfilling runs *inevitable* for sufficiently low bank fundamentals. Goldstein and Pauzner's key contribution is therefore to study how this inevitability reshapes the design of the optimal design contract. The paper shows that the optimal contract under incomplete information penalizes early withdrawals relative to the complete-information benchmark. This entails a welfare cost in terms of providing less insurance to "impatient" consumers; but this cost is now justified by the reduction in the probability and the size of a bank run.

The broader lesson emanating from this paper, as well as from Morris and Shin (1998), is the following: just as the optimal contract characterized in Diamond and Dybvig is dependent on the complete-information assumption, many of the policy recommendations found in the literature are dependent on shutting down frictions in coordination and are often driven by arbitrary equilibrium selections. By contrast, the global-games methodology offers a way to study policy without these caveats.

We finally refer the reader to Goldstein and Pauzner (2004) and Goldstein (2005) for global-games applications that study the role of contagion effects and the phenomenon of "twin crises" (ie, the coincidence of currency crises and bank runs); to Goldstein et al. (2011) for how strategic complementarity among speculators can be the by-product of the signal-extraction problem faced by the central bank when the latter is uncertain about the underlying fundamentals; and to Kurlat (2015) for the optimal stopping problem that such a central bank has to solve in the midst of a speculative attack.

5.2 Big Players and Lenders of Last Resort

Corsetti et al. (2004) sheds new light on the role of large players in speculative markets. The model is similar to the one in Morris and Shin (1998), except for one key difference: a large speculator (Soros) coexists with a continuum of small investors.

When a small speculator chooses whether to attack, she takes the probability of devaluation as exogenous to her own choice. When, instead, the large player chooses whether to attack, she takes into account that she has a nontrivial impact on the probability of devaluation. Other things equal, this effect makes the large speculator more aggressive than any small speculator.[ap] Perhaps more interestingly, the presence of a large player facilitates more coordination among the small players, even if her actions are not observable: because it is known that the large player is more aggressive, each small player finds it optimal to be more aggressive for any realization of her own signal, each small player expects the same from other small players, and so on. What is more, as the small players get more aggressive, the large player finds it optimal to become even more aggressive, and so on. It follows that the introduction of a large player in market can have a disproportionately strong effect on equilibrium outcomes, even if her actions are not

[ap] In effect, the large player is like a pool of small players that have managed to act in a coordinated fashion, thus overcoming the friction faced by other small players.

observable. Allowing the small players to observe the actions of the large player further amplifies this property.

Davila (2012) studies the role of a different type of large players in a different context: the role of large banks in the context of the subprime crisis. The starting point of the analysis is Farhi and Tirole (2012), who identified the following collective moral hazard problem: because the government is more likely ex post to bailout an individual bank when the entire banking system is in trouble, each bank is ex ante more willing to expose itself to an aggregate risk when it expects other banks to do the same. In Farhi and Tirole (2012), this collective moral hazard problem is modeled as a complete information coordination game that ultimately features multiple equilibria: one with low "systemic risk" and another with high.[aq] By contrast, Davila (2012) uses a global-game adaptation to obtain a unique equilibrium. The presence of large banks is shown to intensify the severity of the collective moral hazard problem, increase economy-wide leverage, and make the crisis occur for a large set of fundamentals. The intriguing policy implication is then that systemic risk can be reduced by breaking up the large banks into multiple small ones.

Rochet and Vives (2004) considers a bank-run model in which investors may refuse to renew their credit on the interbank market during the liquidity crisis. In their "globalized" model the equilibrium is unique and the probability of a crisis is linked to the fundamentals. In this unique equilibrium, there is an intermediate range for bank's fundamentals in which in the absence of the government intervention the bank is solvent but may still fail if too large a proportion of investors withdraw their money. In other words, there exists a potential for coordination failure. The authors proceed to study the interaction between ex ante regulation of solvency and liquidity ratios and ex post provision of emergency liquidity assistance. They show that liquidity and solvency regulation can solve the coordination problem, but typically the cost is too high. These prudential measures must be complemented with emergency discount-window loans.

Corsetti et al. (2006) build a similar model as the one in Rochet and Vives (2004), featuring liquidity crises caused by the interaction of bad fundamentals and self-fulfilling runs. The authors focus on the "catalytic" effects of the liquidity provision from the official lender. Drawing on the results of Corsetti et al. (2004) about the role of large players, the paper models the official creditor (IMF) as a large player in the world economy. Even if IMF by its own does not have enough resources to close the large financing gaps generated by speculative runs, the range of economic fundamentals over which liquidity crises do not happen is enlarged through the "catalytic" effect. As a result, liquidity provision during crisis is justified even if it exacerbates debtor's moral hazard problem. A similar point is made in an independent, contemporaneous contribution by Morris and Shin (2006).

[aq] Note that this rests on the government lacking commitment: if the government could commit ex ante not to bail out the banking system ex post, it could alleviate the problem. See Ennis and Keister (2009) for a related point on how the ex post optimal policy response to a bank run distorts ex ante incentives, raising the occurrence of self-fulfilling runs.

The policy lessons that emerge from these papers rely on the uniqueness of the equilibrium. They therefore could not have been obtained in the complete-information versions of the relevant models. However, these lessons also hinge on abstracting from the possibility that policy actions, when made intentionally, convey valuable information about the willingness and ability of the policy maker to defend the status quo (a currency peg, a particular bank, or the entire banking system). As we discuss later on, this possibility is explored in Angeletos et al. (2006) and is shown to lead to a potentially very different lessons.

5.3 Debt runs, Default Premia, and Credit Ratings

Morris and Shin (2004a) globalize the coordination problem among creditors of a distressed borrower in an otherwise conventional debt-run model a la Calvo (1988). There are three periods. The borrower—who is interpreted as a firm but could also be a country—owns a long-term investment project and owes a certain level of short-term debt. The return of the project materializes in period 3. The debt expires in period 2 and must be rolled over to period 3, or else the borrower "dies" and the project is destroyed. There is a large pool of lenders, each of whom is too small relative to the size of the borrower. It follows that the borrower can survive if and only if enough lenders coordinate on rolling over their debts in period 2. By the same token, a coordination failure is possible: even when the borrower's fundamentals are sound, a creditor's fear of other creditors' premature foreclosure may lead to pre-emptive action, thus undermining the project.

Incomplete information guarantees that this coordination failure materializes when the profitability of the project is positive but not high enough: there exists a unique equilibrium in which default takes place in period 2, not only when the project is unprofitable, but also when profitability is positive but not high enough to preclude a self-fulfilling debt run. In period 1, market price debt anticipating that outcomes in period 2 will be determined in the aforementioned fashion. It follows that thanks to the uniqueness of the equilibrium default outcome, there is now also a unique equilibrium price—a property that is not shared by the complete-information version of the model. Furthermore, the equilibrium default risk can be now decomposed in two components: one that measures insolvency (the probability that the project is unprofitable), and another that measures roll-over risk (the probability that a coordination failure materializes).

In a follow-up paper, Zabai (2014) endogenizes the behavior of borrowers and shows how optimal borrowing internalizes the effect that the size of debt has on the probability of coordination failure and thereby on default premia. Together, these papers therefore answer two important questions: how the risk of coordination failure is priced; and how policy makers can influence it.

In another related paper, Holden et al. (2014) study the role of credit-rating agencies. By influencing the behavior of creditors, the ratings published by such agencies affect the probability of default, which in turn affects the ratings of other agencies. Through this feedback effect, the presence of credit-rating agencies can exacerbate volatility and reduce welfare.

Finally, He and Xiong (2012) study the dynamics of debt runs in a model that obtains uniqueness with the help of the techniques found in Frankel and Pauzner (2000) and Burdzy et al. (2001). More specifically, the paper characterizes how the occurrence and the dynamics of debt runs depend on the level fundamental volatility, the presence of credit lines, and the debt maturity structure. Interestingly, they find that commonly used measures that are meant to ease runs, such as temporarily keeping the firm alive and increasing debt maturity, can actually backfire and exacerbate runs—yet another example of how policy conclusions drawn from complete-information models can go wrong.

5.4 "Liquidity Black Holes" and "Market Freezes"

Morris and Shin (2004b) study how the incompleteness of information can help explain sudden drops in asset prices and liquidity in otherwise smoothly functioning markets. A set of sophisticated, risk-neutral traders (who can be thought as "hedge funds") interact with a representative unsophisticated, risk-averse, trader. This means that the value of the asset, and hence also its price, is higher when it is held by the former rather than by the latter. Importantly, the sophisticated traders have short horizons and face privately known loss limits. When the price of the asset is sufficiently above the loss limits of the sophisticated traders, their trades are strategic substitutes: the less the other traders buy of the asset, the cheaper (and more attractive) it becomes for an individual trader to buy. However, once the price of the asset falls sufficiently close to, or below, the loss limits, sales of the risky asset become strategic complements: the more the others sell, the more one has to sell "in distress." This opens the door to a "liquidity black hole" analogous to the run equilibrium in Diamond and Dybvig (1983). A global-game specification is then used to select a unique equilibrium and to obtain a unique trigger point in the fundamentals, below which the liquidity black hole comes into existence. The paper proceeds to show how this helps generate a sharp V-shaped pattern in prices around the time of the liquidity black hole and connect to historical experiences.

Bebchuk and Goldstein (2011) study a debt crisis model in which a credit-market freeze—a situation in which banks abstain from lending to good firms—may arise because of the banks' self-fulfilling expectations that other banks will not lend. The authors "globalize" the model to study the effectiveness of alternative government interventions, including interest rate cuts, infusions of capital into banks, direct lending to operating firms, and provisions of government capital or guarantees to encourage privately managed lending. See also Liu (2016) for an application that explains the joint occurrence of systemic credit runs and interbank market freezes.

5.5 Safe Assets

He et al. (2016) study the pricing of sovereign debt when debt valuations depend on both insolvency and roll-over risk. Their model features two countries that issue sovereign bonds on the international capital markets. An investor's valuation of a sovereign bond depends not only on the country's fundamental, but also on the number of other investors who purchase that bond. For a country's bonds to be safe, the number of investors who invest in the bond must exceed a threshold. As a result, investor actions feature a strategic complementarity similar to the one found in Section 4. What is new is that the model also features strategic substitutability when the number of investors who invest in the bonds exceeds the threshold required to roll over debts: above the threshold, more demand for the bond drives up the bond price, leading to lower returns. Recall that a similar feature was present in Goldstein and Pauzner (2005).

He et al. (2016) "globalize" the environment by removing common knowledge of the relative fundamental of the two countries. Under certain conditions, they prove the existence of a unique, monotone equilibrium. In this equilibrium, a sovereign's debt is more likely to be safe if its fundamentals are strong relative to the other country, but not necessarily strong on an absolute basis. Investors coordinate to invest in the country with relatively better fundamentals, and thus relative valuation determines which country's bonds have less rollover risk and, in this sense, more safety. This prediction may help explain why the valuation of US debt may increases despite a deterioration in its fiscal conditions: the fact that other countries are in even worse shape means that US roll-over risk has gotten smaller.

An additional insight is that a higher level of debt may, perhaps paradoxically, carry a higher price. This happens when the global demand for assets is strong. When this is true, investors are attracted to the country with the highest level of debt, because this helps to satisfy their high demand for assets. But this means that the country with the higher level of debt faces lower roll-over risk, which in turn means that the bond of this country is safer and commands a higher price. By contrast, when the global demand for safe assets is low, investors coordinate on the country with the smallest debt size.

Finally, the authors use their results to evaluate the recent proposal for "Eurobonds," that is, of a common bond for a club of countries. When a significant amount of such bonds is issued, all countries within the club can benefit by reducing the overall roll-over risk. It follows that Eurobonds can be beneficial for, say, both Germany and Greece. For the same reason, however, they can also upset the dominance of US debt as the international benchmark for safety.

5.6 Business Cycles and the Great Recession

A few papers sought to use the global-games methodology to shed light on business-cycle phenomena. The first attempt was Chamley (1999), which studies the dynamics of regime switches in a relatively abstract investment game. A related attempt was made by Frankel and Burdzy (2005), using the technique reviewed in Section 6.1.

More recently, Schaal and Taschereau-Dumouchel (2015) develop a global-games adaptation of the RBC framework and use it to study the Great Recession. We find this paper to be particularly interesting because it succeeds to merge the more abstract global-games literature with the more canonical business-cycle paradigm. What is more, it takes up the challenge of confronting the theory with the data. Below, we briefly explain the key ingredients of the paper.

Firms have the option to pay a fixed cost in order to increase their productivities and reduce their variable costs. This option introduces a nonconvexity in technology. The combination of this nonconvexity with monopolistic competition and aggregate demand externalities introduces a coordination problem that is akin to the one we have formalized in our binary-action regime-change setting: a firm is willing to pay the aforementioned fixed cost if and only if it expects aggregate demand to be sufficiently high, which in turn is true if and only if enough other firms also pay the aforementioned fixed cost.

Under complete information, the model admits multiple equilibria. Schaal and Taschereau-Dumouchel consider a perturbation that maintains common knowledge of past outcomes (including the capital stock) but removes common knowledge of the current TFP level. This permits them to obtain, not only a unique equilibrium, but also tractable dynamics. Although the equilibrium is now unique, there are two locally stable steady states, leading to very different dynamics from those in the standard RBC model. In particular, a large transitory shock may push the economy into a quasi-permanent recession, helping explain the slow recovery and other salient features of the Great Recession. Importantly, these outcomes are the product of a coordination failure. But since this coordination failure is not the symptom of arbitrary equilibrium selection, a meaningful policy analysis is also possible. In particular, the framework justifies certain types of fiscal stimuli when the economy is transitioning between steady states.

A complementary policy point is made in Guimaraes et al. (2014). Similarly to the above paper, this paper studies a micro-founded model with a nonconvexity in production; but it abstracts from capital and productivity shocks. This precludes the type of quantitative assessment that Schaal and Taschereau-Dumouchel (2015) are after; but it also facilitates a sharper analysis of the equilibrium effects of fiscal policy. The key contribution of the paper is then to formalize and characterize the expectational/coordination channel of fiscal policy—a channel that is customarily invoked in real-world policy debates but does not have a sufficiently meaningful counterpart in standard macroeconomic models.

5.7 Prices and Other Endogenous Public Signals

The preceding applications have treated the information structure as exogenous and have used it primarily as a device to select a unique equilibrium. In fact, the literature on applied global games has often focused attention on the limit in which public information is minimal and strategic uncertainty is maximal, in the sense defined earlier on. During times of crises, however, it is unlikely that agents are in the dark. On the contrary, they are closely monitoring the activity of others.

More broadly, agents may have access to various public signals of the underlying fundamentals and/or the actions of others. Think of the price of a peso forward in the context of currency crises, or of media reports about long ATM queues in the context of bank runs. Alternatively, think of speculators observing the actions of a policy maker who is anxious to preempt a speculative attack. How does this kind of information affect the determinacy of the equilibrium and the observables of the model? In the remainder of this section, we briefly review a few papers that have sought to answer this question.

We start with Angeletos and Werning (2006). This paper augments the basic coordination game introduced in Section 4.1 with two kinds of endogenous public signals. The first public signal is a price of an asset whose dividend is correlated either with the fundamental, θ, or the size of attack, K. The second signal is a direct signal of the size of the attack. One can think of a country's stock market or the forward foreign exchange market as examples of the first type, and the ATM queues or the size of protests as examples of the second. In what follows, we focus on the second type, because it can easily be accommodated by our framework; the first type leads to similar lessons.

The payoff structure is the same as in Section 4.1, and so is the specification of private information. In particular, agent i's private signal is once again given by $x_i = \theta + \sigma_\epsilon \epsilon_i$, where the noise $\epsilon_i \sim N(0,1)$ is i.i.d. across agents and $\sigma_\epsilon > 0$. The only change is in the specification of the public signal. The latter is now given by a noisy signal of the size of the attack:

$$z = S(K, \zeta),$$

where S is a monotone function, K is the size of attack, and $\zeta \sim N(0,1)$ is noise independent of θ and $\{\epsilon_i\}$.

Equilibrium can now be defined as follows.

Definition 14 An equilibrium consists of an endogenous signal, $z = Z(\theta, \zeta)$, an individual attack strategy, $k(x, z)$, and an aggregate attack, $K(\theta, \zeta)$, such that

$$k(x,z) \in \arg \max_{k \in \{0,1\}} \mathbb{E}[\, U(k, K(\theta,z), \theta) \,|\, x, z \,] \quad \forall (x,z) \tag{11}$$

$$K(\theta,z) = \mathbb{E}[\, k(x,z) \,|\, \theta, z \,] \quad \forall (\theta,z) \tag{12}$$

$$z = S(K(\theta,z), \nu) \quad \forall (\theta, \zeta, z) \tag{13}$$

Condition (11) requires individual choices to be optimal given all available information, including the one contained in the realized signal z. Equation (12) aggregates. Equation (13) imposes that the signal is generated by the joint equilibrium behavior of the agents. The only novelty in the above definition relative to Definition 3 is the fixed-point relation between the way that agents react to their available information and the way some of that information (namely the signal z) is generated in the first place. This fixed-point relation is standard in noisy rational-expectations models, including those used in the context of financial markets (eg, Grossman and Stiglitz, 1980; Hassan and Mertens, 2014a).

Note that, in equilibrium, K is correlated with θ. It follows that, *in equilibrium*, z becomes also a signal of θ. This fact permits to map the equilibrium analysis of the present model to that of Section 4. What is different from before is that the precision of the information contained in z about θ is now endogenous to the actions of the agents.

To preserve Normality of the endogenous information structure, Angeletos and Werning (2006) impose the following functional form for the S function:

$$S(K,\zeta) = \Phi^{-1}(K) + \sigma_\zeta \zeta,$$

where Φ^{-1} is the inverse of the c.d.f. of the standardized Normal distribution and σ_ζ is a scalar parameterizing the *exogenous* noise in the signal. This functional form, which was first proposed by Dasgupta (2007), guarantees the existence of equilibria in which the observation of z is equivalent to the observation of a Gaussian signal of the following form:

$$\tilde{z} = \theta + \tilde{\sigma}_\zeta \zeta,$$

for some scalar $\tilde{\sigma}_\zeta$, which is itself determined as part of the equilibrium. Taking $\tilde{\sigma}_\zeta$ as given, the equilibrium strategies and the regime outcome can be characterized in exactly the same fashion as in the case where information was exogenous (modulo replacing the scalar σ_ζ that appears in the analysis of Section 4 with the scalar $\tilde{\sigma}_\zeta$ introduced herein). What remains then is to characterize the equilibrium value of $\tilde{\sigma}_\zeta$.

To complete this task, we must solve the fixed-point relation between the strategies and the information. Intuitively, the sensitivity of $k(x, z)$ to x determines the sensitivity of $K(\theta, z)$ to θ, which in turn determines the precision of the signal $\tilde{z}$ with respect to θ, which in turn determines the sensitivity of $k(x, z)$ to x, and so on. Solving this fixed-point problem ultimately yields the following relation between the endogenous $\tilde{\sigma}_\zeta$ and the exogenous parameters of the model:

$$\tilde{\sigma}_\zeta = \sigma_\epsilon \sigma_\zeta.$$

In other words, the precision of the endogenous public signal is proportional to the precision of the exogenous private signal, and inversely proportional to the exogenous noise in the observation of the size of attack.

The intuition is simple. When private signals are more precise, individual decisions are more sensitive to private information. As a result, the equilibrium aggregate attack reacts relatively more to the underlying fundamental θ than to the common noise ζ, thus also conveying more precise information about θ.

This elementary observation has important implications for the determinacy of equilibria and the volatility of outcomes. As the precision of private information increases, the *endogenous* increase in the precision of the available public information permits agents to better forecast one another's actions and thereby makes it easier to coordinate. Consequently, uniqueness need not obtain as a perturbation away from the perfect-information benchmark. Indeed, in the present model, multiplicity obtains when noise is small.

Proposition 10 *In the model described above, an equilibrium always exists. There are multiple equilibria if* either *of the exogenous noises is small, namely if* $\sigma_\zeta^2 \sigma_\epsilon < 1/\sqrt{2\pi}$.

This result does not upset the theoretical insights developed earlier: it remains true that uniqueness obtains if and only if strategic uncertainty is sufficiently high, which in turn is true if and only if private information overwhelms public information. It nevertheless weakens the empirical relevance of the global-games uniqueness result: in many applications of interest, there may be good reasons to expect that the information structure may be such that multiplicity obtains or, more generally, that outcomes remain highly sensitive to shocks that are largely or totally unrelated to the underlying fundamentals.

Angeletos and Werning (2006) reinforce this message by showing the following result: even if we restrict attention to the region of the parameter space in which the equilibrium is unique (namely, the region of σ_ζ and σ_ϵ such that $\sigma_\zeta^2 \sigma_\epsilon > 1/\sqrt{2\pi}$), a local reduction in either σ_ϵ or σ_ζ leads to an increase in the sensitivity of the equilibrium regime outcome to ζ relative to θ. In short:

Corollary 2 *Less exogenous noise can contribute to more nonfundamental volatility, not only by opening the door to multiple equilibria and sunspot volatility, but also by raising the sensitivity of the equilibrium outcome to higher-order uncertainty even when the equilibrium outcome is unique.*

This result, a variant of which we will encounter in Section 7, contrasts with results in single-agent decision-theoretic settings: in such settings, less noise, whether in the form of more information or in the form of lower cognitive friction, contributes to fewer "mistakes" and less nonfundamental volatility in outcomes. A similar property holds in noisy rational-expectations models of financial markets as those studied in Grossman and Stiglitz (1980) and Hassan and Mertens (2014b): in these settings, less exogenous noise, either in the form of a lower volatility in the demand of noise traders or in the form of more precise information in the hands of sophisticated traders, typically leads to more informative asset prices and less nonfundamental volatility in investment. What differentiates these settings from the models we study in this chapter is the absence of strategic complementarity. Once such complementarity is present, the "paradoxical" result reviewed above emerges, whether we consider models that are prone to multiple equilibria (as those studied presently) or models that are prone to equilibrium uniqueness (as those studied in Section 7).

We conclude by mentioning a few additional papers that also concern the informational role of markets in global-games settings: Hellwig et al. (2006), Tarashev (2007), and Ozdenoren and Yuan (2008). Each of these papers focuses on different institutional details and offers additional insights. They nevertheless share the same core message with Angeletos and Werning (2006): there are good reasons to think that the price mechanism, or other endogenous signals of the actions of others, may "tame" the level of strategic uncertainty relative to what is assumed in Morris and Shin (1998, 2001), thus contributing to nonfundamental volatility. Somewhat complementary is also the paper by Iachan and Nenov (2015), which studies how the quality of private information can contribute to instability, even if one holds constant the precision of public information.

We conclude with a basic but important point, which is neatly illustrated in the work of Hellwig et al. (2006). In many applications of interest, strategic complementarity emerges because, and potentially only because, of the effect that the actions of others have on prices. It follows that strategic uncertainty is relevant in such applications only insofar as agents face uncertainty about the relevant prices. Hellwig et al. (2006) explore the implications of this basic observation within the context of self-fulfilling currency crises. More broadly, the take home message is that the price mechanism plays a dual role: it is the *source* of strategic uncertainty, in the sense that it induces agents to care about the actions of others; but it is also the *regulator* of strategic uncertainty, in the sense that the observability of prices controls the magnitude of the relevant strategic uncertainty.

5.8 Policy Interventions

All the applications reviewed earlier on contain important policy insights. They nevertheless abstract from the possibility that active policy interventions may reveal valuable information about the state of the economy and/or the objectives and intentions of the policy maker. Importantly, because such policy interventions are highly visible, they could affect the level of strategic uncertainty, opening the door to effects that were assumed away in the aforementioned applications.

Motivated by this elementary observation, Angeletos et al. (2006) and Angeletos and Pavan (2013) study global-game settings that add a policy maker, who is informed about the underlying fundamentals and can take an action that may influence the size of attack. Think of a central bank trying to preempt a speculative attack by raising domestic interest rates or imposing a tax on capital outflows; or think of a dictator deploying the police or the army to quell a protest. Such actions are likely to reveal that the status quo is neither too strong nor too weak: if the status quo were sufficiently strong, the policy maker would not bother to intervene, for there would be little risk of an attack; and if the status quo were sufficiently weak, it would be pointless to pay the cost of such interventions, for the status quo would be doomed in any case.

Angeletos et al. (2006) thus show that such policy interventions may induce high common belief that the fundamental θ is within the critical region, even if they do not reveal precise information about the exact value of θ. They then proceed to show how this can open the door to multiple equilibria, not only in the regime outcome, but also in the policy action itself. In a certain sense, the policy maker therefore may find himself trapped in a situation in which he has to conform to market expectations that are beyond his control, as opposed to being able to shape economic outcomes with seemingly powerful tools such as taxes and interest rates.

While this result raises an important caveat to policy applications of global-games results, it does not necessarily eliminate the predictive power of the global-games approach. Angeletos and Pavan (2013) show that, even though the signaling role of policy interventions brings back multiple equilibria, the resulting multiplicity is “much smaller" than the one that is present in the complete-information benchmark. We refer the reader to that paper for the explanation of what this "much smaller" means. The bottom line, however, is that a number of concrete policy predictions emerge that are robust across all equilibria and that could not have been reached without the introduction of incomplete information.

Edmond (2013) studies a different kind of policy interventions: interventions that can manipulate the information that is available to the agents, without however directly signaling information to them. The particular application concerns a dictator manipulating the media or the internet in an attempt to preempt a revolution. The action of the dictator is not directly observable. Instead, it is confounded with the signal the agents observe about the underlying fundamentals. This helps preserve equilibrium uniqueness. Yet, the nature of the optimal manipulation hinges on the effect it has on strategic uncertainty, as opposed to merely the information the agents extract about the fundamentals.

Combined, these papers indicate more generally how policy can interact with frictions in coordination, in manners that are absent from workhorse macroeconomic models. Exploring this possibility within a business-cycle context seems an interesting direction for future research.

5.9 Dynamics

The framework studied in Section 4 is static, and so are the applications we have studied so far. The preceding analysis has thus abstracted from the possibility that agents take multiple shots against the status quo and that their beliefs about their abilities to induce regime change vary over time. Yet, these two possibilities are important from both an applied and a theoretical perspective. First, crises are intrinsically dynamic phenomena. In the context of currency crises, for example, speculators can attack a currency again and again until they force devaluation; and their expectations about the ability of the central bank to defend the currency in the present may naturally depend on whether the bank has successfully defended it in the past. Second, learning in a dynamic setting may critically affect the level of strategic uncertainty (ie, uncertainty about one another's actions) and thereby the dynamics of coordination and the determinacy of equilibria.

Motivated by these considerations, Angeletos et al. (2007) study a repeated version of the regime-change game considered in the previous section. Whenever the regime survives an attack, it leads to common knowledge (or at least a strongly correlated belief) that the regime is "not too weak," or else it would not have survived the attack. This kind of endogenous shift in beliefs opens again the door to multiple equilibria, but it also leads to a number of distinct predictions. First, fundamentals may predict the eventual regime outcome but not the timing or the number of attacks. Second, equilibrium dynamics tend to alternate between phases of *tranquility*, where no attack is possible, and phases of *distress*, where a large attack can occur. Finally, attacks take the form of relatively infrequent and acute changes in the level of economic activity.

Costain (2007) provide a complementary result, showing how the public observation of past actions can lead to herding as well as to multiplicity. By contrast, Heidhues and Melissas (2006), Dasgupta (2007), Dasgupta et al. (2012), and Kováč and Steiner (2013) study settings where learning is private, helping preserve the uniqueness of the equilibrium; these papers then proceed to study questions that have to do with the structure of the dynamics, such as the role of irreversible actions, the option value of waiting-to-see, and the synchronization of actions.

Other dynamic global-games applications that maintain the uniqueness of the equilibrium include Chamley (1999), Giannitsarou and Toxvaerd (2006), Guimaraes (2006), He and Xiong (2012), Huang (2014), Mathevet and Steiner (2013), Steiner (2008), and Toxvaerd (2008). Finally, Chassang (2010) studies a dynamic exit game in which multiplicity survives but is much "smaller" than the one that obtains under complete information, thus helping deliver sharper predictions—a message echoed in Angeletos et al. (2007) and Angeletos and Pavan (2013).

Finally, He and Manela (2016) consider a variant of Abreu and Brunnermeier (2003) that derives interesting dynamics from the interaction of the same two features as in Angeletos et al. (2007): the public signal generated by the fact that the regime has survived past attacks, and the arrival of new private information over time. An important difference, however, is the arrival of private information is endogenous: the paper studies the incentives of acquiring information and show how this endogenously leads to the possibility of a new run after an unsuccessful one.

5.10 Other Related Research

We conclude our review of the global-games literature by briefly mentioning three additional lines of research, one theoretical and two empirical.

The first line endogenizes the acquisition of information in global games. Important contributions include Hellwig and Veldkamp (2009) and Yang (2015). The first paper restricts the agents' information choice in a fixed set of private and public Gaussian signals. The second paper considers an attention-allocation problem as in Sims (2003): the agents are free to choose any signal they wish, including a non-Gaussian signal, at a cost that is proportional to the entropy reduction attained by the chosen signal structure. Both papers reach a similar conclusion: because strategic complementarity raises the incentive to observe the *same* information, there is yet another reason why multiplicity may survive. See, however, Denti (2016) and Morris and Yang (2016) for important qualifications to this conclusion and for even more flexible specifications of the information-acquisition technology.

The second line focuses on testing the empirical implications of global games. Prati and Sbracia (2002) use data on consensus forecasts for six Asian countries to measure the cross-sectional heterogeneity of beliefs in the context of speculative attack; they then proceed to document that both the level effect of this heterogeneity and its interaction with measures of fundamentals is consistent with the predictions of global games. Bannier (2006) and Tillmann (2004) provide complementary reduced-form evidence on the role of informational disparities and belief heterogeneity. Daníelsson and Pe naranda (2011) offer a structural estimation of a global game within the context of carry trades in the yen-dollar market. Chen et al. (2010) use mutual fund data to provide evidence that strategic complementarities among investors contribute to financial-market fragility, and interpret this evidence under the lenses of a global game. Nagar and Yu (2014) provide evidence on the coordinating role of public information, interpreting accounting data as a form of public signal. Finally, Nimark and Pitschner (2015) provide complementary evidence on the endogenous correlation of the information disseminated by newspapers.

The third line is also interested in testing the predictions of global games, but does so within the context of laboratory experiments. See Cabrales et al. (2007), Cornand (2006), Duffy and Ochs (2012), Heinemann et al. (2004, 2009), and Shurchkov (2013).

6. COORDINATION AND SYNCHRONIZATION

So far we have focused on settings where incomplete information impedes coordination within any given period. We now shift attention to another aspect: the agents' ability to synchronize their choices.

More specifically, we review two important contributions. The first one, which is by Frankel and Pauzner (2000) and Burdzy et al. (2001), studies the role of adding a Calvo-like friction in a dynamic game of regime change. By preventing synchronous choice, this friction is shown to help select a unique equilibrium. Importantly, this is true even when the friction is vanishingly small. The second contribution, which is by Abreu and Brunnermeier (2003), is also concerned with asynchronous choice. However, the primitive friction is now the asynchronous awareness of a change in the environment, as opposed to a Calvo-like friction in the agents' ability to act. Such asynchronous awareness is shown to cause a significant delay in the response of the economy, even if all agents become aware pretty fast. We discuss how these contributions shed light on the subtle relation between synchronization and coordination; how they are connected to the global-games literature as well as to one another; and how they illustrate, once again, importance of strategic uncertainty and the potentially fragility of standard macroeconomic models.

6.1 The Calvo Friction, Asynchronous Choice, and Coordination

In this section we discuss the contribution of Frankel and Pauzner (2000) and Burdzy et al. (2001). Unlike the "canonical" example in the global-games literature, these papers study dynamic games of regime change in which the fundamental is perfectly observed in every period. They then proceed to show how equilibrium uniqueness can be induced by the combination of persistent shocks to the fundamentals together with a Calvo-like friction in the ability of the agents to reset their actions.

At first glance, this result appears to be distinct from the one in the global-games literature, and possibly of higher relevance. A closer look, however, reveals a tight connection. We elaborate by reviewing Frankel and Pauzner (2000). Burdzy et al. (2001) contains an extension that may be useful for applications but is not needed for our purposes.

The model in Frankel and Pauzner (2000) is a stochastic version of the two-sector model in Matsuyama (1991), which itself a dynamic variant of the "big-push" model of Murphy et al. (1989). Time is continuous, indexed by $t \in [0, \infty)$, and agents are infinitely lived. At each point of time, each agent can be in one of two idiosyncratic states, state 0 or state 1. Let $k_{i,t} \in \{0, 1\}$ denote the state at which agent i is at time t. The interpretation is analogous to the interpretation of the actions in our static framework. In the aforementioned papers, for example, $k_{it} = 0$ corresponds to living in the village and working in agriculture, whereas $k_{it} = 1$ corresponds to living in the city and working in manufacturing. Furthermore, there is again strategic complementarity: it is more profitable to "attack," that is, to live in the city and work in manufacturing, when enough other agents do the same. What is different is that each agent can not instantaneously switch between "attacking" and "not attacking," which in turn explains why k_{it} is a state variable rather than a control variable.

Let $K_t \in [0, 1]$ denote the fraction of the population that rests at state 1 (city/manufacture) at time t. The life-time utility of agent i is given by

$$\int_0^\infty e^{-\rho t} U(k_{it}, K_t, \theta_t)\, dt,$$

where $\rho > 0$ is the discount rate, θ_t is an exogenous fundamental that affects the relative return to being in the village/agriculture, and $U(k_{it}, K_t, \theta_t)$ is the flow utility. The latter is given as in Assumption 3: the flow payoff from $k_{it} = 0$ is normalized to zero, whereas the flow payoff from $k_{it} = 1$ is given by $b - c > 0$ whenever $K_t \geq \theta_t$ and by $-c < 0$ otherwise.[ar]

We introduce stochasticity in the underlying fundamental by assuming that θ_t follows a Brownian motion with zero drift and volatility parameter $\sigma > 0$: $d\theta_t = \sigma dv_t$. For future reference, we note that the case of no aggregate shocks can be approximated by taking the limit as $\sigma \to 0$. We also let $\theta_L \equiv 0$ and $\theta_H \equiv 1$; these points identify the

[ar] Frankel and Pauzner (2000) allow the net payoff from $k_{it} = 1$ to be a more general function $A(\theta_t, K_t)$ that is increasing in K_t and decreasing in θ_t. The restriction we adopt is only for expositional simplicity. Also, we have changed the notation to make it consistent with the one we use in the rest of the chapter. In particular, note that θ_t equals $-z_t$ in their notation: our fundamental is the opposite of theirs. This explains why the functions κ_L, κ_L, and κ^* below have the opposite monotonicity that the corresponding functions in the original paper.

boundaries of dominance regions in the static benchmark (the one-shot game we studied in Section 4).

Unlike the rest of this chapter, we rule out private information: the realized values of both the exogenous aggregate state θ_t and the endogenous aggregate state K_t are public information at t. Instead, the key friction is idiosyncratic inertia of the same type as in Calvo (1983): each agent's option to switch follows an idiosyncratic Poisson process with arrival rate $1/\lambda$. The scalar λ therefore parameterizes the level of inertia in individual adjustment: the higher λ is, the more likely that an agent is "stuck" in her current location for a while. For future reference, note that instantaneous adjustment is approximated by the limit $\lambda \to 0$.

For comparison purposes, let us momentarily consider the knife-edge case in which $\sigma = 0$ and $\lambda = 0$. By this we do not mean the double limit of the aforementioned model as $\sigma \to 0$ and $\lambda \to 0$. Rather, we mean a variant that *literally* shuts down both the shocks to fundamentals and the Calvo-like friction. The following is then trivially true.[as]

Proposition 11 *Suppose $\lambda = \sigma = 0$ and let $\theta \in (\theta_L, \theta_H]$. For any t, and regardless of the history of past play, there is an equilibrium in which all agents "attack" (locate in the city/work in manufacture), as well as an equilibrium in which all such agents choose the opposite.*

This is essentially a multiperiod version of the multiplicity result we encountered in Section 4.3. The only novelty, in terms of observable implications, is that the multiplicity can take the form of rich fluctuations over time: there can be equilibria in which K_t is constant at either 0 or 1 forever, as well as equilibria in which K_t jumps up and down at arbitrary points of time. In short, sunspot volatility manifests in the time series of aggregate economic activity.

Consider next the case in which $\sigma > 0$ but $\lambda = 0$. That is, allow for shocks to fundamentals, but continue to rule out idiosyncratic inertia. Clearly, the multiplicity result survives. The only minor difference is that the possibility of sunspot volatility may now vary over time, depending on whether the current value of the fundamental happens to lie within the critical region $(\theta_L, \theta_H]$.

Consider next the diametrically opposite case in which $\lambda > 0$ but $\sigma = 0$. That is, introduce the Calvo friction, but rule out shocks to fundamentals. Now, K_t becomes a state variable, which moves slowly over time. This expands the regions over which it is dominant for an agent to choose either location. In particular, if K_t is close enough to 1, then it is dominant for an agent to choose city/manufacture in a neighborhood of θ_t above 0, simply because that agent knows that many other agents will be "stuck" in the same location for a while; and symmetrically, when K_t is close enough to 0, it is dominant to choose village/agriculture in a neighborhood of θ_t below 1. Nonetheless, at least insofar as λ is not too high, multiplicity survives.

[as] The solution concept is Perfect Bayesian Equilibrium.

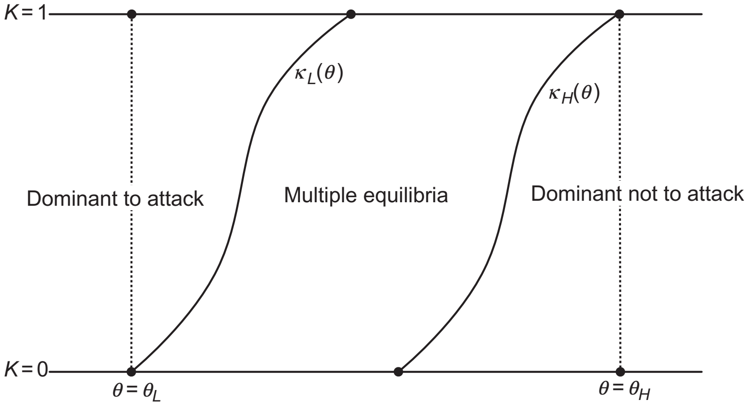


Fig. 2 Multiple equilibria with $\lambda > 0$ and $\sigma = 0$.

This case is illustrated in Fig. 2. One can readily show that there exist increasing mappings $\kappa_L : [\theta_L, \theta_H] \to [0,1]$ and $\kappa_H : [\theta_L, \theta_H] \to [0,1]$, illustrated by the solid lines in the figure, such that the following properties are true. Consider the agents who have the option to act at time t. If $K_t < \kappa_L(\theta_t)$, it is dominant to attack (ie, choose village/agriculture). If $K_t > \kappa_H(\theta_t)$, it is dominant not to attack (ie, choose city/manufacture). Finally, if $\kappa_L(\theta_t) < K_t < \kappa_H(\theta_t)$, there is an equilibrium in which all current agents choose village/agriculture, as well as an equilibrium in which all current agents choose city/manufacture. These equilibria are sustained by the expectation that all *future* agents will do the same as the current agents. In short, the mappings κ_L and κ_H identify the boundaries of, respectively, the lower and the upper dominance regions. For any $\lambda > 0$, we have that $0 < \kappa_L(\theta)$ and $\kappa_H(\theta) < 1$ for all $\theta \in (\theta_L, \theta_H)$. This fact reflects the expansion of the dominance regions mentioned above. But as λ converges to 0, these dominance regions converge uniformly to the dominance regions of the aforementioned variant model in which $\lambda = 0$. The latter are identified by the dotted vertical lines in the figure.

Let us now consider the case of interest, which is the case in which both λ and σ are strictly positive (although potentially arbitrarily small). The following is then true.

Proposition 12 (Frankel–Pauzner) *Suppose $\lambda > 0$ and $\sigma > 0$. There exists a unique, and increasing, mapping $\kappa^* : \mathbb{R} \to [0,1]$ such that the following is true in equilibrium: all the agents who have the option to act at time* t *choose not to attack (ie, locate in the city/work in manufacture) whenever the current aggregate state satisfies $K_t > \kappa^*(\theta_t)$, whereas they make the opposite choice whenever $K_t < \kappa^*(\theta_t)$.*

This result is illustrated in Fig. 3. The two dashed lines give the boundaries of the dominance regions; the solid line gives the mapping κ^*, which of course lies strictly in between the two dominance regions. Different realizations of the Brownian motion may induce the aggregate state (θ_t, K_t) to travel anywhere in the space $\mathbb{R} \times [0,1]$. However, the path of K_t is uniquely pinned down by the path of θ_t. In particular, K_t has a

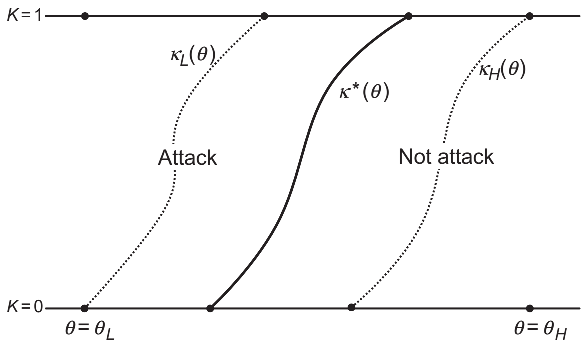


Fig. 3 Unique equilibrium with $\lambda > 0$ and $\sigma > 0$.

positive drift whenever (θ_t, K_t) is on the left of κ^*, and a negative drift on the right of κ^*. In the limit as shocks vanish, K_t therefore converges either to 0 or to 1, depending on initial conditions.

The above result applies even when λ and σ are vanishingly small. This means that the multiplicity result in Proposition 11 is fragile: a small perturbation selects a unique equilibrium.

Furthermore, if we take the limit as the level of idiosyncratic inertia becomes arbitrarily small, we obtain the following sharp characterization of the unique equilibrium.

Proposition 13 *Let* $\theta^* \equiv 1 - \frac{c}{b}$*, fix any* $\sigma > 0$*, and let* $\lambda \to 0$*. The following properties hold in this limit:*

(i) *The mapping* κ^* *becomes vertical at* θ^**:* $\kappa^*(\theta) \to 0 \;\; \forall \;\; \theta < \theta^*$ *and* $\kappa^*(\theta) \to 1 \;\; \forall \;\; \theta > \theta^*$*.*

(ii) *The distribution of* K_t *conditional on* θ_t *converges to a Dirac at* 1 *for all* $\theta_t < \theta^*$ *and to a Dirac at* 0 *for all* $\theta_t > \theta^*$*.*

This result is qualitatively the same as the Morris–Shin limit result in Proposition 8: in essence, every agent is "attacking" whenever the current fundamental is below θ^*, and nobody is attacking whenever the current fundamental is above θ^*, regardless of the history of either past fundamentals or past play. What is more, the threshold θ^* that shows up here is exactly the same as the one that shows up in Proposition 6.

At first glance, this coincidence is surprising, even mysterious. The perturbation considered here is of a different nature than the one considered before: instead of removing common knowledge of fundamentals and introducing private information and strategic uncertainty, we only had to assume (a bit of) idiosyncratic inertia at the individual level along with aggregate shocks to fundamentals. How could it be that two very different perturbations lead to essentially the same selection?

The answer is that there is a deep connection: the combination of idiosyncratic inertia and aggregate shocks transforms the dynamic model under consideration to a game that has a similar mathematical structure as the static global game we have studied in the preceding sections. This fact is most evident if one compares the proof of the uniqueness result in Frankel and Pauzner (2000) with the proof of the uniqueness result that Frankel et al. (2003) provide for a class of static global games with multiple actions: the proofs in the two papers are essentially the same!

To economize space on space and effort, we do not dwell into the formal details. Instead, we provide a sketch of the proof of the uniqueness result in Proposition 12 and relate it to the proof of Proposition 6.

Let us start with the following preliminary observations. At any given point of time, both the exogenous fundamental θ_t and the endogenous state variable K_t are publicly known to all agents. Yet, the mass of agents that have the option to act at any point of time is zero. We can therefore reinterpret the model as one in which a single agent acts at each moment. Because this agent will be "stuck" at her currently chosen action for a while, she must forecast the actions of the agents that will act in the near future. But these agents must themselves forecast the actions of other agents that will move further away in the future. It follows that we can understand the behavior of an agent as a function of her hierarchy of beliefs about the actions of other agents that will have the option to act in either the near or the far future.

These forward-looking higher-order beliefs would not have been relevant if λ were identically zero: it is essential that choices are asynchronous (due to the Calvo-like friction). This, however, does not necessarily mean that the role of higher-order belief vanishes as λ converges to zero from above. As long as both λ and σ are positive, the agents moving at any given point face higher-order uncertainty; and because we are in an environment with strong complementarities, higher-order uncertainty can leave its mark on equilibrium outcomes even if both λ and σ become vanishingly small.

To elaborate, let us fist make the following elementary observation. No matter how small σ is, as long as σ is strictly positive, the fundamentals can drift away over time to either very low or very high levels. As a result, the aggregate state variable (θ_t, K_t) can eventually enter either one of the two dominance regions, even if it is currently far away from both of them.

Consider a moment in time such that the aggregate state is outside the dominance regions but arbitrarily close to one of them, say, the upper one. An agent who acts at that moment does not have a dominant action. However, this agent expects the aggregate state to drift into the upper dominance region with probability close to 1/2, at which point it becomes dominant for future agents to attack (ie, choose village/agriculture). Given this, it becomes optimal–indeed, iteratively dominant–for the current agent to attack as well.

Repeating this argument gives a process of iterated deletion of dominated strategies from "above," which is similar to the one we encountered in the proof of Proposition 6. The only difference is that the process is now over the space of (θ, K), whereas before it was over the space of the private signal x_i. This underscores that the current value of the aggregate state plays a similar role as private information plays in global games.

One can construct a symmetric process of iterated deletion from "below" and can then attempt to prove the uniqueness result by showing that both processes converge to the same point, namely to the strategy described by the mapping κ^*. As it turns out, it is more convenient to consider a modification of the one of the two processes; see Frankel and Pauzner (2000). The substance, however, remains the same: as with the global games, the key mechanism is the impact of higher-order uncertainty, manifested as a contagion effect from the dominance regions.

A natural question is then the following. In global games, uniqueness obtains when private information is sufficiently precise, whereas multiplicity survives if public information is sufficiently precise. Is there an analogue in the present context?

Frankel and Pauzner do not provide an answer to this question. We nevertheless conjecture that multiplicity survives if the fundamental is mean-reverting towards a point in the critical region and, importantly, the mean reversion is sufficiently strong relative to λ and σ. Our intuition is that strong mean reversion in the fundamental has a similar coordinating effect as public information: it helps the current and future agents reach a high common belief about the state.[at]

To recap, the results of Frankel and Pauzner (2000) underscore that, once combined with strategic complementarity and changing fundamentals, asynchronous choice in the form suggested by Calvo (1983) can be interpreted as a friction in coordination.

At some level, this idea is not entirely surprising. In the context of New-Keynesian models, the Calvo friction—and staggered pricing more generally—captures, not only price-stickiness at the individual level, but also the asynchronous price adjustment. It is this asynchronicity that one can reasonably interpret as a friction in coordination, even if one knows nothing about either the global-games literature or the related work of Frankel and Pauzner (2000) and Burdzy et al. (2001).

Yet, the findings of these papers indicate that the workings of the Calvo friction can be more subtle than those understood so far. They also raise the following questions. If the Calvo friction is meant to capture frictions in coordination, does it make sense to impose it *only* on the price-setting behavior of firms, or should we impose it on real choices as well? Also, if the Calvo friction is a proxy for a more primitive friction, what is the latter? Is it menu costs? Or is it an informational friction?

These questions add to the motivation for a later topic of our chapter, namely the study of business-cycle models in which nominal rigidity originates from incomplete information as opposed to Calvo-like asynchronicity. For now, we conclude with few remarks.

Remark 20 In the model described above, equilibrium outcomes are discontinuous in λ at $\lambda = 0$, reflecting the nonvanishing impact of beliefs of arbitrarily high order. This kind

[at] Burdzy et al. (2001) allow for mean-reversion, but take the limit as $\lambda \to 0$. This seems akin to allowing for public information but taking the limit as private information becomes infinitely precise.

of discontinuity, which is similar to the one encountered in global games, rests on the underlying strategic complementarity being of the strong form. If, instead, complementarity is of the weak form, as in the beauty contests and the business-cycle applications we study in Sections 7 and 8, the impact of beliefs of order h vanishes as $h \to \infty$ at a rate high enough that the aforementioned discontinuity does not emerge.[au]

Remark 21 The methods developed in Frankel and Pauzner (2000) and Burdzy et al. (2001) should appeal to macroeconomists, because equilibrium uniqueness is obtained with the help of two familiar modeling ingredients: aggregate shocks to fundamentals and a Calvo-like friction. It is then somewhat surprising that this approach has not attracted more attention in applied research. Exemptions include Frankel and Burdzy (2005) on business cycles, Guimaraes (2006) on currency attacks, and He and Xiong (2012) on debt runs.

Remark 22 The Calvo friction helps captures asynchronous choice in a brute way. A possible alternative is that asynchronous choice is the byproduct of an information friction. Dasgupta et al. (2012) offer an analysis that has such a flavor: they study a dynamic global game in which there is value to synchronize actions and show how lack of common knowledge can impede synchronization. An important earlier contribution in the same vein is Abreu and Brunnermeier (2003), which we review next: in that paper, asynchronous choice is the byproduct of asynchronous awareness.

6.2 Asynchronous Awareness and Coordination (with Application to Bubbles)

We now turn attention to Abreu and Brunnermeier (2003), an important contribution that highlights how asynchronous awareness breaks common knowledge and how this in turn may cause significant delay in the response of equilibrium outcomes to changes in the environment. The particular application considered in Abreu and Brunnermeier (2003) regards asset bubbles: asynchronous awareness is shown to delay the burst of a bubble. The lessons that emerge, however, extend well beyond the particular application.

The model is unlike the ones we have studied so far, because it admits a unique equilibrium regardless of whether information is complete or incomplete. It nevertheless shares a similar contagion mechanism working through higher-order beliefs. The paper's key contribution is to show how this kind of mechanism can help explain why a bubble may persist long after every trader has recognized that the price is unsustainable.

More specifically, the paper studies a continuous-time model in which a unit mass of sophisticated traders, each with relatively shallow pockets, decide whether to ride or attack an asset bubble. If enough of them attack the bubble, then the bubble will burst immediately. If, instead, enough of them choose to ride the bubble, then the bubble will

[au] See Weinstein and Yildiz (2007b) for a formalization of this kind of intuition, albeit within the context of a static game.

survive for a while—but not forever. In particular, the bubble will burst at some random date $t = T$, even if no trader ever attacks it.

Before we proceed, we should clarify what a bubble means. In Abreu and Brunnermeier (2003), a bubble is defined in terms of two possible price paths. The first path, which is interpreted as the "bubble path," features both a higher level and a higher growth rate than the second, the "fundamental path." Both of these paths are exogenous. What is endogenous is only the crash of the bubble, that is, the transition from the first to the second path. The contribution of the paper rests in the characterization of this transition.

This transition is characterized under two scenarios. In the first scenario, Nature publicly announces at $t = 0$ the aforementioned date T to all traders at once. In the second scenario, Nature gradually reveals this date to the traders over a small time interval of length δ; think of this as each trader becoming aware of, or "wakening up" to, the fact that there is a bubble, without however knowing right away how many other traders have also waken up. The two scenarios lead to different predictions, even when δ is arbitrarily small. This is because in the first scenario the aforementioned fact becomes common knowledge instantly, whereas in the second scenario asynchronous awareness implies that common knowledge of the aforementioned fact is not achieved even long time after every single agent has herself become aware of the fact.

Let us elaborate. To start, consider the first scenario, in which all agents wake up simultaneously. In this case, the bubble bursts immediately (at $t = 0$). The proof is as follows. Suppose that a trader believes that no other trader will ever attack. Then, this trader knows that the asset will continue to have abnormal returns for any $t < T$ and that its price will jump down exactly at T. It is then optimal for the trader to ride the bubble up to $t = T - \epsilon$ and sell the asset just at that moment.[av] But if all traders do the same, then the bubble will burst at $T - \epsilon$ rather than T, in which case it becomes optimal for the individual trader to sell the asset at $t = T - 2\epsilon$ rather than at $t = T - \epsilon$. Repeating this argument proves that the bubble bursts immediately, no matter how far in the future T happens to be.

Consider next the second scenario, in which agents wake up in an asynchronous manner. In this case, the bubble can persist for long time after *all* traders have waken up. To explain why, suppose that it takes half of the traders to attack in order for the bubble to burst before T. It follows that the bubble will survive at least till $t = \delta/2$, because not enough traders are awake prior to this point for a coordinated attack to trigger a crash. (Recall that δ denotes the length of time it takes for all traders to wake up; think of δ as small relative to T.)

[av] Although the model is in continuous time, it helps to recast it in discrete time, letting ϵ denote the length between any two consecutive points of time and thinking of ϵ as arbitrarily small.

Consider a trader who just woke up. At this moment, the trader does not know whether she is the first in line to wake up or the last one. As long as she is strictly below the middle of the line, she is better of riding the bubble, because it will take more time for half of the traders to be awake and therefore for a crash to be possible. But if this is true, it is optimal for the trader to ride the bubble even if she happens to be on or just above the middle of the line. Iterating, it is possible to prove that a trader who just woke up finds it optimal–indeed dominant—not to attack the bubble right away.

We can now use this fact to initiate a process of iterated deletion of dominated strategies from "below": given that it is dominant for a trader not to attack at the moment she wakes up, it becomes iteratively dominant for a trader to delay attacking for a certain time interval after she has woken up, which in turn makes it iteratively dominant not to attack for an even bigger time interval, and so on. The opposite process, the one from "above," gets initiated from the fact that it is dominant for a trader to attack once she reaches date $t = T - \epsilon$, for arbitrarily small ϵ.

Abreu and Brunnermeier (2003) prove that the aforementioned two processes converge to the same point, which means that the equilibrium is unique and is associated a critical threshold τ^* such that the following is true: the trader keeps riding the bubble until τ^* units of time have passed since she woke up, and starts attacking the bubble thereafter. It follows that the bubble bursts at date $t = t^* \equiv \delta/2 + \tau^*$, which is how long it takes for exactly 1/2 of the traders to attack. Importantly, t^* can be much larger than δ, although, of course, it has to be smaller than T. It follows that the bubble can survive for long time after it has become known—but not common knowledge—to all traders that the bubble will burst.

To recap, the unraveling that triggers the immediate burst of the bubble in the first scenario hinges on the assumption that traders can instantaneously reach common knowledge of the fact that there is a bubble that can be burst by a coordinated run. Once this assumption is relaxed by allowing for asynchronous awareness of the aforementioned fact, there can be significant delay in the kind of coordinated action that is needed for the bubble to burst and the market price to get in line with the asset's fundamental valuation.

Remark 23 We will encounter similar delay effects within the class of "beauty contests" and the related business-cycle applications that we study in Sections 7 and 8. In that context, the relevant delay effects will manifest as rigidity, or inertia, in the response of certain outcomes to innovations in the underlying fundamentals.

Remark 24 As anticipated, the effects of asynchronous awareness that we reviewed here have a similar flavor as the effects of asynchronous choice—or the Calvo friction—that we discussed in Section 6.1. This supports our earlier claim, namely that there is a tight relation between asynchronous choice and informational frictions. Understanding the exact nature of this relation is an open question; applying the insights to business-cycle settings is another fruitful direction for future research.

7. BEAUTY CONTESTS: THEORY

In this section we turn attention to a class of games that feature weak complementarity and linear best responses. The incomplete-information versions of these games are often referred to as "beauty contests," due to a resemblance with Keynes' parable for financial markets. For applied purposes, one can think of this class of games as a stylization of unique-equilibrium macroeconomic models whose equilibrium conditions can be represented as, or at least be approximated by, a system of (log)linear equations. The lessons we develop in the present section are therefore relevant for canonical business-cycle models of either the RBC or the New-Keynesian type, as well as for a class of asset-pricing models. Such applications are discussed in the next section; here, we develop some basic insights, building in part on the analyses of Morris and Shin (2002b), Angeletos and Pavan (2007, 2009), and Bergemann and Morris (2013).

Remark 25 The settings studied in this and the next section differ from those studied in Sections 4–6 in two related respects. First, they admit a unique equilibrium—in fact, a unique rationalizable outcome—irrespective of the structure of information. Second, they let higher-order beliefs matter in a more modest manner than in global games: beliefs of order h have a vanishingly small effect as $h \to \infty$.

7.1 Setup

In this section we use the framework introduced in Section 2 and impose the following restrictions.

Assumption 4 (Beauty Contest) $D_k = D_\theta = \mathbb{R}$ and U is quadratic in (k, K, θ). Furthermore, U satisfies the following restrictions: $U_{kk} < 0$, $0 \leq -U_{kK}/U_{kk} < 1$, and $U_{k\theta} > 0$.

The first restriction, $U_{kk} < 0$, imposes concavity at the individual level, ensuring that best responses are well defined. The second restriction, $0 \leq -U_{kK}/U_{kk} < 1$, is equivalent to imposing that the slope of best responses with respect to aggregate activity is positive but less than one. This means that strategic complementarity is of the weak form and guarantees that the equilibrium is unique. The last restriction, $U_{k\theta} > 0$, is innocuous: it means that, other things equal, a higher fundamental causes the agent to take a higher action. (If the converse had been true, we could have simply redefined θ as $-\theta$).

Remark 26 Although in this section we restrict attention to the case of weak strategic complementarity (namely $0 \leq -U_{kK}/U_{kk} < 1$), many of the key lessons, including those about the role of incomplete information in sustaining inertia and animal spirits, extend also to the case of strategic substitutability (namely $-U_{kK}/U_{kk} < 0$).

7.2 Complete-Information Benchmark

As in the case of global games, we start by studying the benchmark of complete information, which in turn nests the case of perfect information.

Proposition 14 (Complete Info) *There exist coefficients* $(\kappa_0, \kappa_1, \kappa_2)$*, pinned down by the payoff function U, such that the following is true: whenever information is complete, the equilibrium action is given by*

$$k_i = \kappa\left(\mathbb{E}_i\theta_i, \mathbb{E}_i\overline{\theta}\right) \equiv \kappa_0 + \kappa_1\mathbb{E}_i\theta_i + \kappa_2\mathbb{E}_i\overline{\theta}, \tag{14}$$

where $\overline{\theta} \equiv \int \theta d\Theta(\theta)$ *is the average fundamental and* $\mathbb{E}_i$ *is the rational expectation conditional on* ω_i.

The coefficients $(\kappa_0, \kappa_1, \kappa_2)$, which are characterized in the Appendix, depend on U and thereby on the micro-foundations of the particular application that lies behind U. Understanding what determines these coefficients and how they can be identified in the data is important within any individual application.[aw]

Recall that complete information means common knowledge of the distribution of information, $\mathbf{\Omega}$, but allows for the possibility that agents face uncertainty either about their own fundamental (θ_i) or about the aggregate fundamentals ($\mathbf{\Theta}$, or $\overline{\theta}$). This explains why the forecast of the fundamentals, $\mathbb{E}_i\theta_i$ and $\mathbb{E}_i\overline{\theta}$, show up in the above characterization. Trivially, the case of perfect information is then nested simply by letting $\mathbb{E}_i\theta_i = \theta_i$ and $\mathbb{E}_i\overline{\theta} = \overline{\theta}$. That is, as we move from perfect information to imperfect but complete information, all we have to do is to replace the actual fundamentals with the forecasts of them. This offers a sharp illustration of the more general point we made in Section 3: insofar as strategic uncertainty is ruled out, it makes little difference whether the agents face uncertainty about the fundamentals or know them perfectly.

7.3 Equilibrium with Incomplete Information

Suppose now that information is incomplete. In this case, equilibrium actions need not be pinned down by fundamentals (or forecasts of fundamentals), because the latter are generally not sufficient for forecasting the actions of others. This makes equilibrium outcomes vary away from, and around, their complete-information counterparts, in a manner that is formalized in the next two propositions.

Proposition 15 (Equilibrium and Coordination) *The equilibrium satisfies the following fixed-point relation*

[aw] For instance, in the context of the neoclassical economy introduced in Section 2, κ_1 measures the "micro" elasticity of the response of a farmer's output to an idiosyncratic shock to her own productivity, whereas $\kappa_1 + \kappa_2$ measures the "macro" elasticity of the response of aggregate output to an aggregate productivity shock. The former is representative of ZIP- or state-level elasticities that are estimated in Mian et al. (2013), Mian and Sufi (2014), and Nakamura and Steinsson (2014) on the basis of appropriate cross-sectional variation in the data; the latter is what macroeconomists are most often concerned with; the two differ because of general-equilibrium effects, which are captured here by the dependence of U (and of U_K in particular) on K.

$$k_i \;=\; \mathbb{E}_i\left[\kappa\left(\theta_i,\overline{\theta}\right)\right]+\alpha\cdot\mathbb{E}_i\left[K-\kappa\left(\overline{\theta},\overline{\theta}\right)\right], \tag{15}$$

where $\alpha \equiv -U_{kK}/U_{kk} \in [0, 1)$.

Proposition 16 (Equilibrium and Higher-order Beliefs) *Suppose that each agent knows her own fundamental. Then, the equilibrium action of agent* i *is given by*

$$k_i=\kappa_0+\kappa_1\theta_i+\kappa_2\mathbb{E}_i\left\{\sum_{h=0}^{\infty}(1-\alpha)\alpha^h\bar{\mathbb{E}}^h[\overline{\theta}]\right\}, \tag{16}$$

where $\bar{\mathbb{E}}^h[\overline{\theta}]$ *denotes the* h*th order average forecast of the mean fundamental.*[ax]

Proposition 15 highlights the dependence of equilibrium allocations on beliefs regarding aggregate activity (forecasts of the actions of others). To understand condition (15), recall that $\mathbb{E}_i\left[\kappa\left(\theta_i,\overline{\theta}\right)\right]=\kappa\left(\mathbb{E}_i\theta_i,\mathbb{E}_i\overline{\theta}\right)$ is the action agent *i* would have taken in equilibrium had information been complete. How much an agent deviates from this benchmark when information is incomplete depends on $\mathbb{E}_i\left[K-\kappa\left(\overline{\theta},\overline{\theta}\right)\right]$, which is her forecast of the deviation of the average action of the rest of the population from this benchmark, weighted by the coefficient α. In this sense, the coefficient α measures how much each *individual* cares about aligning her action with that of others, or equivalently the private motive to coordinate: it identifies the degree of strategic complementarity.

Proposition 16 then restates this result in terms of the hierarchy of beliefs (forecasts of the forecasts of others). For this part we have added the restriction that each agent knows her own fundamental. This restriction allows us to isolate the uncertainty that agents face about one another's actions from the uncertainty that each agent may face about her own preferences and technologies: we shut down the latter and concentrate on the former. Comparing the above result to Proposition 14, we then see that the key difference as we move from complete to incomplete information is that $\mathbb{E}_i\overline{\theta}$ has been replaced by $\mathbb{E}_i\left\{\sum_{n=0}^{\infty}(1-\alpha)\alpha^h\bar{\mathbb{E}}^h[\overline{\theta}]\right\}$, a weighted average of the entire hierarchy of beliefs about the underlying aggregate fundamental. This is because an agent's first-order belief of the aggregate shock is no longer sufficient for forecasting the equilibrium level of aggregate activity; the agent needs to forecast the forecasts of others. The coefficient α then determines the sensitivity of the equilibrium action to higher-order beliefs: the stronger the degree of complementarity, the stronger the impact of higher-order beliefs relative to first-order beliefs.

That said, note that the specifics of higher-order beliefs do not matter per se: if we change, say, the 3rd and the 7th order belief of an agent while keeping $\mathbb{E}_i\left\{\sum_{n=0}^{\infty}(1-\alpha)\alpha^h\bar{\mathbb{E}}^h[\overline{\theta}]\right\}$ constant, then her optimal action does not change. This is

[ax] The operator $\bar{\mathbb{E}}^h$ is defined recursively by letting, for any variable X, $\bar{\mathbb{E}}^0[X]=X$, $\bar{\mathbb{E}}^1[X]\equiv\bar{\mathbb{E}}[\bar{\mathbb{E}}^0[X]]\equiv\int\mathbb{E}_i[X]di$ and $\bar{\mathbb{E}}^h[X]\equiv\bar{\mathbb{E}}\left[\bar{\mathbb{E}}^{h-1}[X]\right]=\int\mathbb{E}_i\left[\bar{\mathbb{E}}^{h-1}[\overline{\theta}]\right]di \;\; \forall h\geq 2$.

because the agent's best response depends merely on her first-order belief of K, not the details of either the belief hierarchy about the fundamentals or the underlying information structure. This underscores, once again, that incomplete information and higher-order uncertainty are modeling devices that allow the researcher to accommodate interesting, and potentially testable, variation in the agents' expectations (first-order beliefs) of economic outcomes. We will discuss how this can be understood in a manner that is robust to the underlying micro-foundations.

7.4 Simplifying Assumptions

To simplify exposition, we make a few additional assumptions. First, we rescale the fundamental so that the following is true.

Assumption 5 $\kappa_0 = 0$ and $\kappa_1 + \kappa_2 = 1$.

Using the characterization of the coefficients κ_1 and κ_2 (equation (A.6) in the Appendix), one can then show that $\kappa_1 = 1 - \alpha$ and $\kappa_2 = \alpha$. This means that, under complete information, $1 - \alpha$ measures the *micro* elasticity of the individual's action to her own fundamental, α measures the general-equilibrium feedback in the case of aggregate shocks; and $1 = (1 - \alpha) + \alpha$ measures the *macro* elasticity of the aggregate action to an aggregate shock. It also implies that the agent's best response under incomplete information, or equation (15), reduces to the following:

$$k_i = (1-\alpha)\mathbb{E}_i[\theta] + \alpha \cdot \mathbb{E}_i[K]. \tag{17}$$

To simplify the exposition and also to stay close to the applied literature, we rule out payoff heterogeneity and assume a Gaussian specification for the underlying common fundamental and for the information the agents receive about it.

Assumption 6 There is a common fundamental: $\theta_i = \overline{\theta} = \theta$, for all i and all states of nature.

Assumption 7 The fundamental θ is drawn from a Normal distribution with mean 0 and variance $\sigma_\theta^2 > 0$.[ay] Each agent observes two signals, a private one and a public one. The private signal takes the form

$$x_i = \theta + u + \epsilon_i,$$

where u and ϵ_i are common and idiosyncratic noises that are independent of one another as well as of θ, and are drawn from Normal distributions with mean 0 and variances, respectively, $\sigma_u^2 \geq 0$ and $\sigma_\epsilon^2 \geq 0$. The public signal takes the form

$$z = \theta + \zeta,$$

[ay] Allowing for a nonzero prior mean for θ only adds a constant in the equilibrium action, without affecting any of the positive properties we document below.

where ζ is noise, independent of θ, u and all ϵ_i, and drawn for a Normal distribution with mean 0 and variance $\sigma_\zeta^2 \geq 0$.

Incomplete information requires $\sigma_\epsilon > 0$ and $\sigma_\zeta > 0$, that is, nontrivial private information about the aggregate fundamental, but does not necessarily require $\sigma_u > 0$. In fact, much of the related literature (eg, Morris and Shin, 2002b; Angeletos and Pavan, 2007) is nested here by restricting $\sigma_u = 0$, that is, by shutting down the correlated noise, u, in the private signal. The reason why we allow for such correlation is twofold. First, it helps mimic the correlation in beliefs that emerges as agents talk to their neighbors or participate in localized markets. Second, the noise u in combination with the noise ζ allow us to disentangle the aggregate variation in beliefs of aggregate activity from the variation in either the fundamental or the beliefs of it.

To formalize this point, consider the residuals of the projection of $\bar{\mathbb{E}}\theta$ and $\bar{\mathbb{E}}K$ on θ.[az] From Proposition 17 and Proposition 18 (see below), we have that the two residuals are linear transformations of u and ζ. When u is shut down, as is done in much of the existing literature, the two residuals become proportional to one another, implying that the noise in beliefs of endogenous economic outcomes is perfectly correlated, and thus indistinguishable from the noise in the beliefs of the fundamentals. When instead u is switched on, as we do here, the two residuals cease to be perfectly correlated, thus permitting us to disentangle the two types of noise. By the same token, the introduction of u helps accommodate a strong form of "animal spirits": as emphasized in Proposition 19 below, expectations of economic activity (and actual activity) can fluctuate holding constant *both* the true fundamentals and the agents' beliefs of them.

It is also worth noting that the assumed information structure is less restrictive than it seems. Let A be the set of joint distributions for $(\theta, \bar{\mathbb{E}}\theta, K, \bar{\mathbb{E}}K)$ that can be obtained in equilibrium under the assumed information structure, for some $(\sigma_u^2, \sigma_\epsilon^2, \sigma_\zeta^2)$. Next, let B be the set of joint distributions for $(\theta, \bar{\mathbb{E}}\theta, K, \bar{\mathbb{E}}K)$ that can be obtained in equilibrium under *arbitrary* Gaussian information structures. The latter nests cases in which the agents observe an arbitrary collection of Gaussian signals, not only about the underlying fundamental (θ), but also about the endogenous outcome (K) or even about one another's signals; it can also nest situations where the agents talk to one another, either directly, or through some market mechanism. Using a similar argument as in Bergemann and Morris (2013), one can show that $A = B$. One can thus think of the assumed signal structure merely as a convenient parameterization of the type of outcomes that can be obtained under arbitrary information structures, as opposed to seeking a literal interpretation of it.

Unless stated otherwise, we impose σ_ϵ, σ_ζ, $\sigma_u > 0$.

[az] To ease notation, we henceforth let $\bar{E}X$ be a short-cut for $\bar{\mathbb{E}}[X]$, for any variable X.

7.5 The Structure of Higher-Order Beliefs

As a stepping stone towards the characterization of the equilibrium, we next study how the hierarchy of beliefs varies with the underlying aggregate shocks, namely the fundamental θ and the noises ζ and u.

Proposition 17 (Higher-Order Beliefs) *For every $h \geq 1$, there exist positive scalars $(\omega_\theta^h, \omega_u^h, \omega_\zeta^h)$ such that the* h*th order average expectation of the fundamental is given by*

$$\bar{\mathbb{E}}^h[\theta] = \omega_\theta^h \theta + \omega_\zeta^h \zeta + \omega_u^h u,$$

for every realization of (θ, ζ, u). Furthermore, for every $h \geq 1$,

$$1 > \omega_\theta^h > \omega_\theta^{h+1} > 0 \quad \text{and} \quad 0 < \frac{Var\left(\bar{\mathbb{E}}^h[\theta]\middle|\theta\right)}{Var\left(\bar{\mathbb{E}}^h[\theta]\right)} < \frac{Var\left(\bar{\mathbb{E}}^{h+1}[\theta]\middle|\theta\right)}{Var\left(\bar{\mathbb{E}}^{h+1}[\theta]\right)} < 1.$$

This result means that beliefs of any given order h vary with the fundamental, θ, and the two sources of noise, u and ζ; but the higher the order h is, the smaller the absolute response of the hth order belief to the fundamental; and the higher the contribution of the noise relative to that of the fundamental in driving the volatility in hth order beliefs. The first property, namely the lower response to the fundamental, is the mirror image of the fact that higher-order beliefs are more anchored to the prior (which of course does not move with the realized fundamental) than lower-order beliefs. The second property, on the other hand, means that higher-order beliefs are more "noisy" than lower-order beliefs, in the sense that the R-square of the projection of the former on the fundamental is lower than that of the projection of the latter on the fundamental.

These properties have been established here only for a very specific information structure. However, using methods similar to those of Bergemann and Morris (2013), it is possible to show that these properties are shared by essentially any Gaussian information structure. By the same token, the positive implications we document next are not driven by any ad-hoc specifications of the information structure; rather, they are *generic* to the incompleteness of information.

7.6 Positive Implications

We now proceed to translate the preceding properties of higher-order beliefs into those of the observables of the model. To this goal, we must first take a stand on what the model's observables are. For the purposes of the present exercise, we envision that the researcher collects data on the exogenous fundamental θ, the endogenous aggregate outcome K, and possibly the average beliefs of these objects, namely $\bar{\mathbb{E}}\theta$ and $\bar{\mathbb{E}}K$. For instance, in the context of the neoclassical economy introduced in Section 2 and taken up again in the next section, K corresponds to aggregate output and θ to labor productivity. The researcher can observe these variables in standard macroeconomic data sets, and can extract information about agents' beliefs of these objects from surveys of expectations.

Definition 15 The model's observables are $(\theta, \bar{\mathbb{E}}\theta, K, \bar{\mathbb{E}}K)$. The model's predictions are the restrictions the model imposes on the joint distribution of $(\theta, \bar{\mathbb{E}}\theta, K, \bar{\mathbb{E}}K)$.

Note that the stochastic properties of θ and $\bar{\mathbb{E}}\theta$ are dictated directly by the assumptions the researcher makes about shocks and information. A model's essence is in the "cross-equation restrictions" it imposes on the stochastic properties of the endogenous objects K and $\bar{\mathbb{E}}K$. We now characterize these restrictions.

Proposition 18 (Positive Properties) *For any given information structure, there exist positive scalars* $(\phi_\theta, \phi_u, \phi_\zeta)$ *and* $(\psi_\theta, \psi_u, \psi_\zeta)$, *such that the aggregate outcome and the average forecast of it satisfy the following:*

$$K = \phi_\theta \theta + \phi_\zeta \zeta + \phi_u u \quad \textit{and} \quad \bar{\mathbb{E}}[K] = \psi_\theta \theta + \psi_\zeta \zeta + \psi_u u.$$

Furthermore,

$$1 > \phi_\theta > \psi_\theta > 0, \tag{18}$$

and

$$0 < \frac{Var\,(K|\theta)}{Var\,(K)} < \frac{Var\,\left(\bar{\mathbb{E}}[K]\middle|\theta\right)}{Var\,(\bar{\mathbb{E}}[K])} < 1. \tag{19}$$

This result represents the equilibrium counterpart of Proposition 17. To interpret it, note that ϕ_θ identifies the slope of the aggregate outcome to the fundamental, ψ_θ identifies the corresponding slope of the average forecast of the aggregate outcome, and finally the (normalized) slope of either object under complete information is 1. Condition (18) therefore means the average forecast under-reacts relatively to the actual outcome, and that both of them under-react relative to the complete-information benchmark. Condition (19), on the other hand, means that the average forecast of K is relatively more noisy than the actual K, in the sense that the relative contribution of the noises ζ and u to the volatility of $\bar{\mathbb{E}}K$ is larger than those to the volatility of K. In this sense, forecasts under-react to fundamentals but overreact to noise.

This result suggests a unifying explanation for two seemingly contradictory sets of facts: the *under-reaction* of forecasts documented in Coibion and Gorodnichenko (2012); and the *overreaction* of forecasts documented in Greenwood and Shleifer (2014) and Gennaioli et al. (2015). We expand on the relation to Coibion and Gorodnichenko (2012) in the next section; we leave the potential connection to Greenwood and Shleifer (2014) and Gennaioli et al. (2015) to future work.

We also note that the actual level of aggregate activity and the average belief of it are positively but only imperfectly correlated:

$$1 > Corr(K, \bar{\mathbb{E}}[K]) > 0.$$

By contrast, when information is complete, the above correlation becomes perfect.

Pushing this observation a bit further, we have the following result.

Proposition 19 (Animal Spirits) *The economy features "animal spirits," not only in the sense of Definition 13, but also in the sense that*

$$Var\left(K|\theta,\bar{\mathbb{E}}\theta\right)>0.$$

Furthermore, the following is necessarily true:

$$Var\left(\bar{\mathbb{E}}K\middle|\theta,\bar{\mathbb{E}}\theta\right)>0 \quad \text{and} \quad Cov\left(K,\bar{\mathbb{E}}K\middle|\theta,\bar{\mathbb{E}}\theta\right)>0.$$

Recall that Definition 13 identified animals spirits with situations in which outcomes vary without commensurate variation in the agents beliefs of their *own* fundamentals. In the present context, own and aggregate fundamentals coincide ($\theta_i=\overline{\theta}$ for all i and all states of nature). It follows that animal spirits obtain in the sense of Definition 13 if and only if $Var(K|\bar{\mathbb{E}}\theta)>0$.

The first part of Proposition 19 reinforces this possibility by showing that variation in outcomes can obtain holding constant, not only the beliefs of the fundamental, but also the realized fundamental. This helps disentangle "animal spirits" from the noise in beliefs of the fundamental: aggregate activity varies while holding constant both the true fundamental and the aggregate error in beliefs of the fundamental. It also underscores that the notion of "animal spirits" used here, and the related notion found in Angeletos and La'O (2013), is both conceptually and empirically distinct from that used in Lorenzoni (2009) and Barsky and Sims (2012): in these papers, "animal spirits" is defined as noise in beliefs of fundamentals. Instead, the notion used here is closely connected to the one used in the literature on multiple equilibria and sunspot fluctuations; as noted before, the only difference is that we engineer animal spirits with the help of rich higher-order beliefs as oppose to multiple equilibria and correlation devices.

The second part of Proposition 19 states that the volatility attributed to "animal spirits" is features positive comovement between actual and expected outcomes, underscoring once more its self-fulfilling nature.

We conclude that incomplete information allows the researcher to rationalize empirically relevant imperfections in the comovement of fundamentals, economic outcomes, and beliefs thereof—thereby accommodating a number of facts that may be prima-facie inconsistent with workhorse macroeconomic models. We elaborate on this point in the next section, within the context of specific applications.

7.7 Complementarity, Information, and Volatility

Before moving to the applications, we study how the degree of strategic complementarity α affects the different kinds of volatility in the model's observables. Clearly, the stochastic properties of θ and $\bar{\mathbb{E}}\theta$ are invariant to α; we thus need to consider only the effect of α on the stochastic properties of K and $\bar{\mathbb{E}}K$, which is what the next proposition does.

Proposition 20 (The Effect of Complementarity) *Stronger strategic complementarity (higher α) results in all of the following:*

(i) *Less covariation between the fundamental and either the aggregate outcome or the average belief of it:*

$$\frac{\partial Cov\,(K,\theta)}{\partial\alpha}<0 \quad and \quad \frac{\partial Cov\,(\bar{\mathbb{E}}K,\theta)}{\partial\alpha}<0$$

(ii) *Higher portion of volatility driven by noise:*

$$\frac{\partial}{\partial\alpha}\left(\frac{Var\,(K|\theta)}{Var\,(K)}\right)>0 \quad and \quad \frac{\partial}{\partial\alpha}\left(\frac{Var\,(\bar{\mathbb{E}}K|\theta)}{Var\,(\bar{\mathbb{E}}K)}\right)>0$$

(iii) *More room for animal spirits, not only in the sense of Definition 13, but also in the stronger sense defined in Proposition 19:*

$$\frac{\partial Var\,(K|\theta,\bar{\mathbb{E}}\theta)}{\partial\alpha}>0 \quad \frac{\partial Var\,(K|\theta,\bar{\mathbb{E}}\theta)}{\partial\alpha}>0 \quad and \quad \frac{\partial}{\partial\alpha}\left(Cov\left(K,\bar{\mathbb{E}}K\middle|\theta,\bar{\mathbb{E}}\theta\right)\right)>0$$

To appreciate this result, note that under the maintained normalization $\kappa_1+\kappa_2=1$, the value of α is completely immaterial for aggregate outcomes when information is complete. The above result therefore reveals how incomplete information interacts with strategic complementarity (or general-equilibrium effects) to shape the covariation of actual economic outcomes and their forecasts. To put it differently, this result hinges entirely on strategic uncertainty.

We conclude by noting that while incomplete information helps accommodate animal spirits and overreaction to certain forms of noise, it can do so only at the expense of dampening the overall volatility.

Proposition 21 (Volatility Bound) *The unconditional variance of* K *under incomplete information is lower than its perfect-information counterpart.*

This result follows from the basic fact that the variance of the forecast of any variable is always lower than the variance of the variable itself. Applying this fact recursively gives that the variance of higher-order forecasts is lower than the variance of lower-order forecasts, which in turn is lower than the variance of the fundamental. Because the equilibrium K is a weighted average of the hierarchy of forecasts, it then follows that the volatility of K is maximal when information is perfect.

That said, we wish to emphasize that this result concerns the *unconditional* variance of K, whereas researchers are often interested in certain *conditional* moments, such as those conditional on certain shocks or, in dynamic contexts, on certain past outcomes. When it comes to this kind of conditional moments, incomplete information may actually contribute to more volatility.[ba] Finally, although the volatility of the average fundamental

[ba] As a trivial example, recall that $Var(K|\bar{\mathbb{E}}\theta)>0$ and $Cov(K,\bar{\mathbb{E}}K)<1$*only* when information is incomplete: with complete information, $Var(K|\bar{\mathbb{E}}\theta)=0$ and $Cov(K,\bar{\mathbb{E}}K)=1$.

places an upper bound on the volatility of the aggregate outcome in the current setting, this bound can be relaxed in settings that allow for local interactions and/or idiosyncratic shocks; see Angeletos and La'O (2013) and Bergemann et al. (2015).

7.8 Dynamics and Learning

So far the analysis has been confined to a static framework in order to deliver key insights in a sharp and transparent manner. In applications, however, it is often central to allow for two different kinds of dynamics. The first has to do with learning over time (the dynamics of beliefs). The second has to do with intertemporal payoff interdependences, which make the decisions depend on past and/or future actions (backward- and/or forward-looking effects).

7.8.1 Slow Learning and Inertia

To illustrate the role of the dynamics that obtain from slow learning, we now consider a repeated variant of our static beauty contest game. In this variant, the best responses remain static, in the sense that they depend only on the contemporaneous actions of others. Yet, interesting dynamics obtain because the fundamental is persistent and there is slow private learning about it.

Example 1 The fundamental of each agent i follows a Gaussian random walk:

$$\theta_{it} = \theta_{it-1} + \nu_t + \xi_{it},$$

where $\nu_t \sim N(0, \sigma_\nu^2)$ is an aggregate innovation and $\xi_{it} \sim N(0, \sigma_\xi^2)$ is an idiosyncratic one. Furthermore, in every period t, player i observes a private signal of her own fundamental:

$$x_{it} = \theta_{it} + \epsilon_{it} \tag{20}$$

where $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ is an idiosyncratic noise.

The optimal action of agent i is still given by

$$k_{it} = \mathbb{E}_{it}[(1-\alpha)\theta_{it} + \alpha K_t].$$

What is novel relative to our static setting is only the dynamic evolution of beliefs implied by the combination of persistent fundamentals and noisy learning.

To see how these features matter for the observables of the model, let us first consider the special case in which there is no complementarity ($\alpha = 0$). In this case, equilibrium outcomes depend merely on first-order beliefs of fundamentals. Furthermore, thanks to the Gaussian specification, the dynamics of first-order beliefs can be shown to obey the following rule:

$$\mathbb{E}_{it}\theta_t = \mathbb{E}_{it-1}\theta_{it} + \lambda[x_{it} - \mathbb{E}_{it-1}x_{it}],$$

for some $\lambda \in (0, 1)$ that depends on the standard deviations of the innovations in the fundamental and the noise in the signal. This result is an application of the Kalman filter, with

λ being the Kalman gain.[bb] Using $\mathbb{E}_{it-1}x_{it} = \mathbb{E}_{it-1}\theta_{it}$ and rearranging gives $\mathbb{E}_{it}\theta_{it} = (1-\lambda)\mathbb{E}_{it-1}\theta_{it-1} + \lambda x_{it}$. Aggregating implies that

$$\bar{\mathbb{E}}_t\theta_t = (1-\lambda)\bar{\mathbb{E}}_{t-1}\theta_{t-1} + \lambda\theta_t.$$

Finally, when $\alpha = 0$, we have that $K_t = \bar{\mathbb{E}}_t\theta_t$. It follows that we can express the dynamics of K_t as follows:

$$K_t = (1-\lambda)K_{t-1} + \lambda\theta_t,$$

When information is perfect ($\sigma_\epsilon = 0$), λ becomes 1 and the expression above reduces to $K_t = \theta_t$, implying that K_t follows a random walk, just as the underlying fundamental. When instead information is imperfect ($\sigma_\epsilon > 0$), λ is strictly lower than 1, implying that K_t features sluggish response to the innovation in the fundamental. In particular, if we let IRF_j denote the cumulative effect of an aggregate innovation on the level of aggregate activity j periods after the innovation occurs,[bc] we have the following characterization:

$$IRF_0 = \lambda, \quad IRF_1 = \lambda[1 + (1-\lambda)], \quad IRF_2 = \lambda\left[1 + (1-\lambda) + (1-\lambda)^2\right], \quad \text{etc.}$$

Note then that IRF_j is an increasing sequence that starts from λ and converges to 1, with a higher speed of convergence when λ is higher. Combining this observation with the fact that λ is decreasing in σ_x, we infer that more noise (higher σ_x) induces more inertia in the response of the aggregate outcome K to any innovation in the underlying fundamental.

It is important to recognize that, when $\alpha = 0$, the information that matters is only the one that each agent has about her *own* fundamental, not the one about the aggregate fundamental. What explains the inertia documented above is therefore the lack of (first-order) knowledge of own fundamentals, not the lack of common knowledge of aggregate fundamentals. By the same token, this kind of inertia can be quantitatively important only insofar as the agents are sufficiently uninformed about their own fundamentals.

We can summarize all these points as follows.

Proposition 22 *Slow learning can induce inertia* by itself*, even if $\alpha = 0$ (ie, in the absence of strategic, or general-equilibrium, interdependence). However, the following properties hold whenever $\alpha = 0$:*

(i) *The inertia that is present at the aggregate level (in the response of aggregate outcomes to aggregate shocks) coincides with the inertia that is present at the individual level (in the response of idiosyncratic outcomes to idiosyncratic shocks).*

(ii) *Both types of inertia vanish as the agents learn their own fundamentals, regardless of the information they may or may not have about aggregate economic conditions.*

[bb] Strictly speaking, λ should be allowed to be a deterministic function of t, reflecting the dependence on the initial prior. But as time passes, the impact of the prior vanishes, and λ_t converges to a constant, which identifies the steady-state solution of the Kalman filter. Here, as in most of the applied literature, we focus on this solution.

[bc] Formally, for any $j \geq 0$, $IRF_j = \partial\mathbb{E}[K_{t+j}|K_{t-1}, v_t]/\partial v_t$.

By contrast, when $\alpha > 0$, inertia can result from lack of common knowledge of the aggregate fundamental and can be quantitatively important, even if each agents is arbitrarily well informed about both her own or the aggregate fundamentals. This is because higher-order beliefs are less sensitive to innovations in fundamentals than lower-order beliefs, not only contemporaneously (which is the effect already documented in the static setting), but also in terms of their dynamic adjustment.

We illustrate this property with the following modification of the preceding example.

Example 2 The fundamental of each agent i is given by $\theta_{it} = \overline{\theta}_t + \xi_{it}$, where $\xi_{it} \sim N\left(0, \sigma_{\xi}^2\right)$ is an idiosyncratic shock and $\overline{\theta}_t$ is the aggregate fundamental. The latter follows a random walk:

$$\overline{\theta}_t = \overline{\theta}_{t-1} + \nu_t,$$

where $\nu_t \sim N\left(0, \sigma_{\nu}^2\right)$ is an aggregate innovation. Finally, in every period t, agent i observes her own fundamental, θ_{it}, along with the following additional private signal about the aggregate fundamental:

$$x_{it} = \overline{\theta}_t + \epsilon_{it},$$

where $\epsilon_{it} \sim N\left(0, \sigma_{\epsilon}^2\right)$ is idiosyncratic noise.

We can then show the following.

Proposition 23 (Complementarity and Inertia) *When $\alpha = 0$, $K_t = \overline{\theta}_t$. When instead $\alpha \in (0, 1)$,*

$$K_t = \gamma_K K_{t-1} + \gamma_{\theta} \overline{\theta}_{t-1} + \gamma_{\nu} \nu_t,$$

for some scalars $\gamma_K, \gamma_{\theta}, \gamma_{\nu}$ that depend on $(\alpha, \sigma_{\epsilon}, \sigma_{\xi}, \sigma_{\nu})$.

This result follows from Angeletos and La'O (2010)'s adaptation of Woodford (2003). Woodford (2003) studies an environment where there is no payoff heterogeneity, thus combining two mechanisms: the inertia induced by mere lack of knowledge of own fundamentals, with the inertia induced by strategic complementarity and lack of common knowledge of the aggregate fundamentals. By contrast, Angeletos and La'O (2010) and the closely related example we study here isolate the second mechanism by imposing perfect information about own fundamentals. The dependence of K_t on K_{t-1} and $\overline{\theta}_{t-1}$ documented in the above result therefore reflects the endogenous persistence introduced by, *and only by*, the friction in coordination.[bd]

To gain further insight, we consider a numerical example. Fig. 4, which is borrowed from Woodford (2003), illustrates the dynamic response of the average first- and higher-order forecasts of $\overline{\theta}_t$ to a positive innovation in it. For any order h, the response of $\overline{\mathbb{E}}_t \overline{\theta}_t$,

[bd] The exact characterization of this dependence was elusive to both Woodford (2003) and Angeletos and La'O (2010), but becomes possible with the method developed in Huo and Takayama (2015a,b).

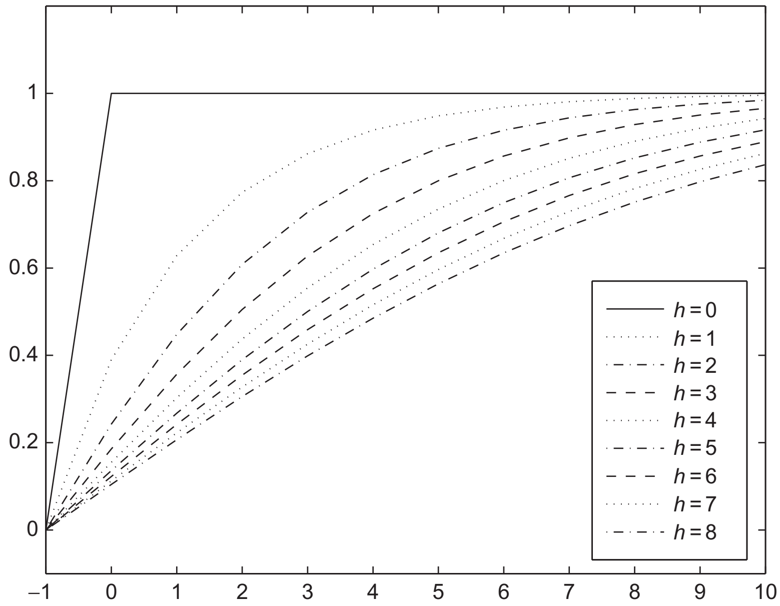


Fig. 4 Dynamics of first- and higher-order beliefs.

the *h*th order forecast, eventually converges to 1 as time passes, reflecting the accumulation of more and more information over time. Yet, in any finite period after the innovation has occurred, higher-order beliefs remain further away from 1 than lower-order beliefs, mirroring the property documented in Proposition 17 on the static framework. What is more, higher-order beliefs converge to 1 more slowly than lower-order beliefs. All in all, Fig. 4 therefore reveals that higher-order beliefs exhibit, not only lower sensitivity on impact, but also more sluggishness over time.

The translation of these properties to the observables of the model is straightforward. Importantly, if we fix the information structure (and hence also fix the dynamics of first- and higher-order beliefs) but increase the degree of strategic complementarity, we obtain both a smaller immediate response and a more sluggish adjustment of the equilibrium level of activity.

Last but not least, note that the mechanism we have documented above operates at the aggregate level but not at the idiosyncratic level (or in the cross section): because the agents are perfectly informed about their own fundamentals, they exhibit no inertia at all in their response to idiosyncratic shocks.

In the next section, we present two applications of this mechanism: one in the context of an RBC economy hit by technology shocks, another in the context of New-Keynesian economy hit by a monetary shock. Moving beyond these examples, the broader lesson is the following.

Fact Frictions in coordination, modeled as incomplete information coupled with strategic complementarity, can help generate significant inertia in the response of aggregate outcomes to aggregate shocks to fundamentals, even when the available information about the fundamentals is precise and even when there is no such inertia in the response of idiosyncratic outcomes to idiosyncratic shocks.

Remark 27 As anticipated in Section 6.2, the inertia we have documented here is of a similar nature as the delay effects documented in the asynchronous-awareness setting of Abreu and Brunnermeier (2003). In both cases, the key is that higher-order beliefs adjust more slowly that first-order beliefs–or, equivalently, that agents continue to lack confidence on whether others will adjust to a change in the environment long after they have themselves become confident that this change took place.

7.8.2 Intertemporal Payoff Interdependencies

We now turn attention to the role of intertemporal payoff interdependencies. In particular, we have in mind settings in which an agent's actions in any given period depend on her expectations of the actions of other agents, not only in the same period, but also in the future. Such forward-looking features are standard in dynamic macroeconomic models. Importantly, these features expand the room for higher-order beliefs.

Consider, for example, the textbook RBC model. In this model, the labor-supply and consumption-saving choices of a household in any given period depend, not only on the current wage and interest rate, but also on her expectations of wages and interest rates in all subsequent periods. Furthermore, the wage and the interest rate that clear the relevant markets in any given period depend on the joint behavior of all consumers and firms, which in turn depend on their own expectations of future wages and interest rates, and so on. It follows that the kind of high-order beliefs that are relevant can be vastly richer than the one we have allowed so far. In particular, even if the past and current fundamentals are common knowledge, one can sustain frictions in coordination by introducing private information about future fundamentals.

To illustrate these points, consider a dynamic game in which the best-response function of player i happens to take the following form:

$$k_{it} = \mathbb{E}_{it}[g(\theta_{it}, K_t; \theta_{it+1}, K_{t+1})], \tag{21}$$

where g is a linear function. The dependence of g on θ_{it} and K_t mirrors the one accommodated in our earlier static analysis; the novelty here is the inclusion of the next-period variables θ_{it+1} and K_{t+1}. This inclusion captures the forward-looking aspects of the environment.

To give an example, let us consider a monetary model that merges Woodford (2003) with Taylor (1979, 1980). In particular, firms face two frictions in their price-setting behavior: first, they have incomplete information about aggregate economic conditions (as in Woodford); and second, they can reset their prices only every two periods (as in

Taylor). In the absence of these two frictions, the optimal price in period t would have been given (after an appropriate log-linearization) by

$$p_{it}^{*} \equiv (1-\alpha)\theta_{it} + \alpha P_t$$

where θ_{it} summarizes exogenous demand and supply shocks hitting the firm in period t, P_t is the endogenous price level in period t, and α is the degree of strategic complementarity in pricing decisions. When only the first friction (incomplete information) is present, the optimal price set by firm i is

$$p_{it} = \mathbb{E}_{it} p_{it}^{*} = \mathbb{E}_{it}[(1-\alpha)\theta_{it} + \alpha P_t]. \tag{22}$$

When instead both of the aforementioned frictions are present, the optimal reset price is given by

$$\begin{aligned} p_{it} &= \mathbb{E}_{it}\left[\frac{1}{1+\beta} p_{it}^{*} + \frac{\beta}{1+\beta} p_{it+1}^{*}\right] \\ &= \mathbb{E}_{it}\left[\frac{1}{1+\beta}((1-\alpha)\theta_{it} + \alpha P_t) + \frac{\beta}{1+\beta}((1-\alpha)\theta_{it+1} + \alpha P_{t+1})\right] \end{aligned} \tag{23}$$

where β is the firm's discount factor. Clearly, while equation (22) is nested in our static framework, (23) is nested in the dynamic extension we have just introduced.[be]

Now we go back to the general set up, equation (21). We will show that strategic uncertainty can originate, not only from differential *current* information about the current and future fundamentals, but also from uncertainty of the information that others may receive in the *future*. To isolate the second channel, we shut down the former by restricting $g_2 = 0$. In other words, only future actions of other agents, not current actions of others, matter for an agent's best response. For simplicity, we also assume that there are no idiosyncratic shocks, so that $\theta_{it} = \theta_t$. Then, using the linearity of g and aggregating, we get

$$K_t = g_1 \bar{\mathbb{E}}_t \theta_t + g_3 \bar{\mathbb{E}}_t \theta_{t+1} + g_4 \bar{\mathbb{E}}_t K_{t+1}. \tag{24}$$

Next, let

$$\tilde{\theta}_t \equiv \theta_t + \frac{g_3}{g_1}\theta_{t+1},$$

[be] As already noted, the example considered here is a hybrid of Woodford (2003) and Taylor (1979, 1980). If one considers hybrids of Woodford (2003) and Calvo (1983), like those studied in Angeletos and La'O (2009) and Nimark (2011), the key difference is that p_{it} then depends on the expectation of $(\theta_{i\tau}, P_\tau)$, not only for $\tau \in \{0, 1\}$ as in the present example, but also for $\tau > 1$. This enriches the type of strategic uncertainty that is relevant, which can be important for quantitative purposes; but it only reinforces the message we wish to convey here.

Clearly, we can then rewrite the above as

$$K_t = g_1 \bar{\mathbb{E}}_t \widetilde{\theta}_t + g_4 \bar{\mathbb{E}}_t K_{t+1}. \tag{25}$$

With some abuse of notation, we henceforth redefine θ_t as $\widetilde{\theta}_t$. This indicates that the appropriate interpretation of what the "period-t fundamentals" mean in the theory is neither the exogenous payoff relevant variables that happen to get realized in period t nor the exogenous shocks that enter the period-t flow payoff of an agent; rather, it is the collection of all payoff-relevant variables that are relevant for period-t decisions, regardless of the time the values of these variables materialize or appear in the player's flow payoffs.

Notwithstanding these observations, let us now solve for the equilibrium K_t. To this goal, let $z_t^0 \equiv \bar{\mathbb{E}}_t \theta_t$ and, for all $j \geq 1$, let

$$z_t^j \equiv \bar{\mathbb{E}}_t z_{t+1}^{j-1} = \bar{\mathbb{E}}_t \{ \bar{\mathbb{E}}_{t+1} \{ \ldots \{ \bar{\mathbb{E}}_{t+j} \{ \theta_{t+j} \} \} \ldots \} \}.$$

Iterating (25) yields

$$K_t = g_1 \sum_{j=0}^{+\infty} (g_4)^j z_t^j. \tag{26}$$

It follows that K_t depends not only on today's forecasts of fundamentals (the forecasts captured in z_t^0) but also on today's forecasts of *tomorrow*'s forecasts of fundamentals (the forecasts captured in z_t^1), on today's forecasts of tomorrow's forecasts of forecasts two period ahead (z_t^2), and so on. Uncertainty about the information that others may receive in the *future* can thus serve as an independent source of higher-order uncertainty in the *present*. Nevertheless, there is no need to get lost in the wilderness of higher-order beliefs. If we put aside condition (26) and go back to (24), we can readily see that, conditioning on the equilibrium first-order beliefs of K_t and K_{t+1}, the actual outcome K_t is invariant to the underlying hierarchy of beliefs about the fundamentals. As with our static framework, it therefore remains true that the practical value of incomplete information is to enrich the testable restrictions on the observables of the model.

7.9 Endogenous Information Acquisition

We now discuss a strand of the literature that endogenizes the acquisition of information, or the allocation of attention, within beauty contests. We follow Myatt and Wallace (2012) and Pavan (2015) because their framework is essentially the same as ours. Other notable contributions include Hellwig and Veldkamp (2009), Vives (2016), Chahrour (2014), Colombo et al. (2014), Llosa and Venkateswaran (2015), Yang (2015), and Denti (2016).

This line of work is closely related to the one on rational inattention, such as Sims (2003) and Maćkowiak and Wiederholt (2009). It nevertheless departs from it in that it pays attention to the following fact: economic agents may wish to collect information that is useful for predicting and tracking the actions of others, even if that information is not particularly useful for predicting and tracking the underlying fundamentals.

We use the payoff structure introduced in Section 7, along with the simplifying assumptions introduced in Section 7.1 and 7.4. The only novelty is the introduction of an "information technology," by which we mean the following.

Each agent i has access to $N \in \mathbb{N}$ potential sources of information about θ, the underlying common fundamental, which is drawn from $N\left(0, \sigma_\theta^2\right)$. We let $\pi_\theta \equiv \sigma_\theta^{-2}$ denote the precision of the common prior about θ. The information contained in each source $n \in \{1, \ldots, N\}$ is given by

$$y_n = \theta + \epsilon_n,$$

where $\epsilon_n \sim N\left(0, \eta_n^{-1}\right)$ is i.i.d. Normally distributed noise, independent of θ. By paying attention $z^i = \left(z_n^i\right)_{n=1}^N \in \mathbb{R}_+^N$ to the available sources, agent $i \in [0,1]$ then receives private signals $x^i \equiv \left(x_n^i\right)_{n=1}^N \in \mathbb{R}^N$ as

$$x_n^i = y_n + \xi_n^i,$$

where $\xi_n^i \sim N\left(0, \left(t_n z_n^i\right)^{-1}\right)$ is i.i.d. Normally distributed noise, independent of θ and $\epsilon \equiv (\epsilon_n)_{n=1}^N$. Myatt and Wallace (2012) and Pavan (2015) interpret the parameter η_n as the *accuracy* of the source (the "sender's noise"), and the parameter t_n as the *transparency* of the source (how a marginal increase in the attention increases the precision of private signal that each agent ultimately obtains from "listening" to the source).

Agents incur a cost when paying attention to any given source. In particular, payoff are now given by

$$U(k_i, K, \theta) - C\left(z^i\right) \tag{27}$$

where U is as in Section 7 and C is an increasing and continuously differentiable function.

To simplify exposition, we normalize $U_{kk} = -1$ and restrict attention to the case in which attention is perfectly substitutable across different sources in the following sense.[bf]

Assumption 8 $C\left(z^i\right) = c\left(\sum_{n=1}^N z_n^i\right)$, where $c(\cdot)$ is strictly increasing, strictly convex, and differentiable.

One can then establish the following result regarding the equilibrium allocation of attention.

Proposition 24 *There exists a threshold* $\Delta > 0$ *such that, in the unique symmetric equilibrium,*

$$z_n = \frac{\eta_n max\left\{\left(\Delta - \frac{1}{\sqrt{t_n}}\right), 0\right\}}{(1-\alpha)\sqrt{t_n}}.$$

[bf] See Myatt and Wallace (2012) and Pavan (2015) for the more general case.

Proposition 24 follows from proposition 2 in Myatt and Wallace (2012), as well as from proposition 1 and corollary 1 in Pavan (2015). Transparency (t_n) determines whether a source of information will receive any attention and also how much attention it receives. Accuracy (η_n) instead only influences how much attention each information source receives, conditional on that source being used. The intuition why accuracy does not play a role in determining whether an information source receives positive attention is as follows. When z_n is small, the total amount of noise from information source n is dominated by the receiver noise, ξ_n^i. As a result, when thinking about which information source to pay attention, a player starts with the most transparent source.

Let us now define the *publicity* of an information source as the correlation of the noises in signals that two different agents receive from this source[bg]:

$$\rho_n \equiv corr\left(x_n^i, x_n^j | \theta\right) = \frac{z_n t_n}{z_n t_n + \eta_n}$$

We can now state the following result, which is due to Myatt and Wallace (2012) and regards the interaction of the coordination motive with the acquisition of information.

Proposition 25 *As the degree of strategic complementarity* α *rises, attention moves away from more private signals and towards more public signals: there is a* $\hat{\rho}$ *such that the attention paid to source n is locally increasing in* α *if and only if* $\rho_n > \hat{\rho}$.

Intuitively, public signals act as effective focal points for players' coordination. As the desire for coordination strengthens, agents pay more attention to such signals.

Pavan (2015) investigates the question of whether the equilibrium allocation of attention is efficient from the perspective of a social planner that is interested in maximizing ex ante welfare. This complements the earlier work by Angeletos and Pavan (2007), which studied the efficiency of the equilibrium use of information, taking the latter as exogenously given. We refer the reader to these works for an analysis of these normative questions; and to Angeletos and La'O (2012), Llosa and Venkateswaran (2015), and Paciello and Wiederholt (2014) for applications related to business cycles and macroeconomic policy.

We close this section by circling back to the connection between the present framework and the rational inattention literature. The applications of this literature that relate to monetary policy are briefly noted in Section 8.5; for a more extensive review, see Sims (2010). Here, we instead focus on two conceptual issues: the sense in which the framework we study can nest the works of Sims (2003) and Maćkowiak and

[bg] This is related to the notion of the *commonality* of information used in Angeletos and Pavan (2007). See that paper for how this notion helps understand both the positive and the normative properties of the class of games we have been studying.

Wiederholt (2009); and the sense in which it allows for an important departure from these works.

Let $x^i = (x^i_1, \ldots, x^i_N)$ and $y = (y_1, \ldots, y_N)$. Next, assume that instead of facing a cost for paying attention to different sources of information, the agents face a "capacity constraint" of the following form:

$$\Gamma\left(x^i, y\right) \leq \overline{\Gamma}, \tag{28}$$

for some function Γ. The agent's attention-allocation problem is then to choose z^i so as to maximize

$$\mathbb{E}\left[U(k_i, K, \theta) - \lambda\Gamma\left(x^i, y\right)\right]$$

where $\lambda \geq 0$ is the Lagrange multiplier associated with the capacity constraint. This is obviously nested in our present framework by letting $C(z^i) = \lambda\Gamma\left(x^i, y\right)$, no matter what the function Γ is.

To nest the information-acquisition problems studied by Sims (2003) and much of the related rational-inattention literature, we need to make two assumptions. The first is that Γ is the function that measures the mutual information between two random variables; in a Gaussian context like the one we are studying here, this means letting $\Gamma\left(x^i, y\right) = \frac{1}{2} log\left[det\left[V(x^i)\right] / det\left[V(x^i|y)\right]\right]$, where $\det[\cdot]$ denotes the determinant of a matrix, $V\,(x^i)$ denotes the covariance matrix of the random variable x^i, and $Var(x^i|y)$ denotes the covariance matrix of x^i conditional on y.

The second assumption is that there is a single and perfect source of information about the fundamental: $y = \theta$. This means that the attention-allocation problem reduces to one of trying to track *only* the fundamental, as opposed to various sources of information. This is not a serious limitation in a single-agent, decision-theoretic context. In a context with strategic, or general-equilibrium interactions, however, this restriction means there is no more a useful distinction between tracking the fundamental and tracking the actions of others: under this restriction, all the noise in the private signals of the agents becomes *purely* idiosyncratic, implying that the equilibrium value of **K** is pinned down by the fundamental θ. In equilibrium, predicting **K** is therefore the same as predicting θ.

It is this last property that is more relevant for our purposes. If none of the available sources of information is perfect, the underlying noise in the sources becomes correlated noise in the private signals received by different agents. As noted before, such correlated noise helps disentangle the expectations of endogenous outcomes from the expectations of exogenous fundamentals, for it accommodates aggregate shocks in the gap between first- and higher-order beliefs. The more flexible forms of information acquisition proposed by Myatt and Wallace (2012) and Pavan (2015) are therefore best suited for understanding how strategic, or general-equilibrium interaction may *endogenously* lead to more correlated noise.[bh]

On the other hand, an important feature of Maćkowiak and Wiederholt (2009) that is missing from both Myatt and Wallace (2012) and Pavan (2015) is the trade off in allocating attention to idiosyncratic vs aggregate shocks. Merging the two frameworks is an interesting venue for future research.[bi]

7.10 A Note on Solution Methods

Dynamic models with time-varying fundamentals and dispersed information can be hard to solve. Townsend (1983) first suggested that a finite state space solution for the equilibrium dynamics need not exist, reflecting an infinite-regress problem in forecasting the forecast of others.

To circumvent the problem, Townsend (1983) proposed the following short-cut: let the exogenous state of nature become common knowledge after T periods, where $T \geq 1$ is potentially large but finite. Then, for a class of linear models, one can guess and verify the existence of an equilibrium in which aggregate outcomes can be expressed as linear functions of the history of shocks over the last T periods, along with the relevant commonly-known state variables prior to that period.[bj]

It is worth noting here that Townsend (1983) was concerned with settings in which there were no strategic complementarity in actions and higher-order beliefs entered

[bh] We refer the reader also to Denti (2016) for an analysis that covers games with either a finite number or an infinity of players. See in particular the discussion in Section 9.4 of that paper regarding the significance of allowing for noise in the underlying sources of information if one wishes to obtain nonfundamental volatility in games with an infinity of players, like those studied in the present chapter.

[bi] Myatt and Wallace (2012) and Pavan (2015) restrict attention to settings where, as in the analysis above, there is a single fundamental that is common to all agents. Maćkowiak and Wiederholt (2009) study a setting in which there are two types of fundamentals: an aggregate monetary shock, and a purely idiosyncratic TFP shock. They study the optimal allocation of attention across these two types of fundamentals, under the restriction that each agent observes an pair of independent noisy private signals for the two fundamentals (as opposed to observing a joint signal of both fundamentals). Denti (2016) considers a more general structure that allows for arbitrary correlation among the fundamentals of different players, but does not focus on the aforementioned trade off and, unlike Myatt and Wallace (2012) and Pavan (2015), rules out the noise in the underlying sources of information.

[bj] For instance, if we consider Example 1 from Section 7.8, this would mean expressing K_t as a linear function of θ_{t-T-1} and of $(\nu_{t-T}, \ldots, \nu_{t-1}, \nu_t)$. More generally, one would have to include the values of any endogenous state variable (such as capital), as well as any aggregate shock to information (such as the noise in public signals), not just fundamentals.

decisions only though a signal-extraction problem.[bk] The method, however, readily extends to the class of environments we are interested in. For recent applications, see Bacchetta and van Wincoop (2006), Hellwig (2005); Hellwig and Venkateswaran (2009), and Lorenzoni (2009).

It is tempting to interpret Townsend's short cut as an approximate solution for the situation in which the past shocks never become common knowledge. However, the conditions under which this is true are not completely understood. On the one hand, Hansen and Sargent (1991) and Kasa (2000) verify that this is indeed the case for the leading example in Townsend (1983).[bl] On the other hand, Rondina and Walker (2014) present a version of Singleton's (1987) asset-pricing model in which the equilibrium is perfectly revealing when the lag T is finite (no matter how large), but not when T is infinite.

An alternative approximation method involves a truncation of the hierarchy of beliefs, rather than a truncation of the history of shocks. Nimark (2011) pursues this method. He also shows that the approximation error vanishes as the truncation gets larger for a class of dynamic linear models that exhibit a similar "stability" property as the static games studied in Weinstein and Yildiz (2007b), namely the property that the impact of higher-order beliefs on actions vanishes at an exponential rate as the order increases. Notwithstanding the theoretical appeal of this method, its computational efficiency and its comparison to Townsend's approach remain unclear.

Barring the above kinds of shortcuts and approximations, the literature has long struggled to obtain *exact* finite-state solutions for models with infinite horizons and perpetually dispersed information. A few papers have considered special examples in which it is possible to guess and verify the existence of such a solution using the Kalman filter; see Pearlman and Sargent (2005), Singleton (1987), and Woodford (2003). Another line of work has obtained analytic characterizations of the equilibrium by transforming the problem into the frequency domain; see Futia (1981), Kasa (2000), Kasa et al. (2007), Acharya (2013), and Rondina and Walker (2014). These papers have therefore been able to provide important insights into the dynamics of higher-order beliefs, often in settings

[bk] More specifically, Townsend (1983) used two examples. In the first example, there are two industries, each trying to extract a signal about the underlying fundamental from the price of the other industry. In the second, there is a continuum of industries, each trying to extract a signal of the underlying fundamental from aggregate outcomes. Townsend suggested that an infinite-regress problem exists in both examples, because an agent's signal extraction depends on the actions of others, which in turn depend on their signal extraction, and so on. However, Hansen and Sargent (1991) points out that the apparent infinite-regress problem vanishes in the first example: the solution turns out to be perfectly revealing due the limited dimensionality of the underlying uncertainty. The problem remains in the second example, which is the one that Townsend (1983) actually concentrates on, thanks to the fact that available signals are contaminated by measurement error.

[bl] That is, the example with a continuum of industries studied in section VIII of Townsend (1983).

with endogenous signals. Yet, the results of these papers appear to rest on special settings, which has limited their applicability and has not helped advance quantitative work.

A significant breakthrough was recently made by Huo and Takayama (2015a,b). The authors show that a finite-state space solution is possible for a large class of linear models insofar as the observed signals are exogenous and the underlying shocks (fundamentals and noises) follow ARMA processes. In particular, the solution itself has an ARMA structure, although typically of a higher order than those of the underlying shocks. The method may still be computationally intensive if there are multiple endogenous state variables and multiple shocks, but it is flexible and powerful.[bm]

A more radical approach is proposed in Angeletos et al. (2015). The key idea is to use a heterogeneous-prior specification that helps proxy the dynamics of higher-order beliefs that are induced by incomplete information, while bypassing all the computational difficulties. Each agent is assumed to believe that there is a bias in the signals of others. Variation in this perceived bias moves higher-order beliefs without moving first-order beliefs. The noise is taken to zero, shutting down noisy learning and forcing a low-dimensional representation of the stochastic process of the entire belief hierarchy. All in all, Angeletos et al. (2015) are able to augment a large class of DSGE models with rich yet highly tractable higher-order belief dynamics: the belief-augmented models can be solved and structurally estimated with the same ease and computational efficiency as standard, representative-agent models.[bn]

We review two related applications of the last two methods in Section 8.7 and use them to argue that the waves of optimism and pessimism that are rationalized by higher-order belief dynamics can offer a quantitatively potent explanation of business-cycle phenomena. Additional works that seek to push the quantitative frontier of the literature include Melosi (2014), who estimates a version of Woodford (2003); Mankiw and Reis (2007); Kiley (2007), and Reis (2009), who estimate DSGE models with sticky information; Maćkowiak and Wiederholt (2015), who study the quantitative performance of a DSGE model in which both firms and consumers are rationally inattentive; and David et al. (2014), who use firm-level data to gauge the cross-sectional misallocation of inputs caused by informational frictions.

[bm] The method also rules out signals of endogenous outcomes. This need not be a serious limitation for certain quantitative purposes, because the informational content of endogenous signals can be mimicked by appropriately chosen exogenous signals. Furthermore, an extension of the method can be used to obtain an approximate solution to models with endogenous signals; see Huo and Takayama (2015b) for details.

[bn] The potential downside is that this method gives the researcher the freedom to specify higher-order beliefs “at will” (ie, without the restrictions imposed by the combination of the common-prior assumption and specific information structures). How this freedom is used in practice is a delicate balancing act. Angeletos et al. (2015) propose that data on forecasts can offer much of the needed discipline, but do not pursue this idea in detail.

8. BEAUTY CONTESTS: APPLICATIONS

The analysis of the previous section has indicated that incomplete information has two key positive implications. First, it dampens the response of equilibrium actions to changes in the underlying fundamentals. Second, it accommodates forces akin to animal spirits along the unique equilibrium. In this section we explore what these properties mean within the context of specific applications; how they can help accommodate important facts that are inconsistent with certain workhorse macroeconomic models; and how they can inform policy. We also make a digression to discuss the connection between the mechanisms we study here and those in the complementary literature on sticky information (Mankiw and Reis, 2002; Reis, 2006) and rational inattention (Sims, 2003; Maćkowiak and Wiederholt, 2009).

8.1 Real Rigidity and Technology Shocks

In this section we discuss how frictions in coordination can be the source of real rigidity at the macro level[bo] and how this in turn can help reconcile the RBC paradigm with Gali (1999).

Gali (1999) argues that a key implication of the RBC model regarding the response of employment to technology shocks is grossly inconsistent with the data. More specifically, Gali uses a structural VAR method to estimate the impulse response functions of macroeconomic aggregates to an identified technology shock in US data: he run a VAR on key macroeconomic variables and identified the technology shock as the only shock that drives labor productivity in the long run.[bp] He then compares the impulse response functions obtained in the data with those predicted by the model.

Fig. 5, which is a replica of fig. 4 in Gali (1999), depicts the impulse responses of labor productivity, output, and employment obtained in the data. As it is evident in the first panel, the identified shock in the data triggers an immediate and persistent increase in productivity. This makes the identified shock in the data comparable to the theoretical technology shock in the RBC model. And yet, as seen in the last panel, the identified technology shock leads to a decline in hours/employment in the short run, which is exactly the opposite of what the RBC model predicts.

Gali (1999) proceeds to argue that this fact is consistent with the New-Keynesian framework. Suppose that prices are sticky. Suppose further that monetary policy fails

[bo] By "real rigidity" we refer to inertia in the response of real quantities, such as employment and output, to shocks in fundamentals, such as preferences and technology. See Angeletos et al. (2016b) for a complementary discussion of the sense in which incomplete information can be the source of either real or nominal rigidity (or of both) in the related applied literature.

[bp] Gali (1999) run a baseline bivariate VAR on hours and labor productivity; the same finding obtains in a five-variable VAR on hours, labor productivity, money supply, interest rates and price level. Fig. 5 below is based on the five-variable specification.

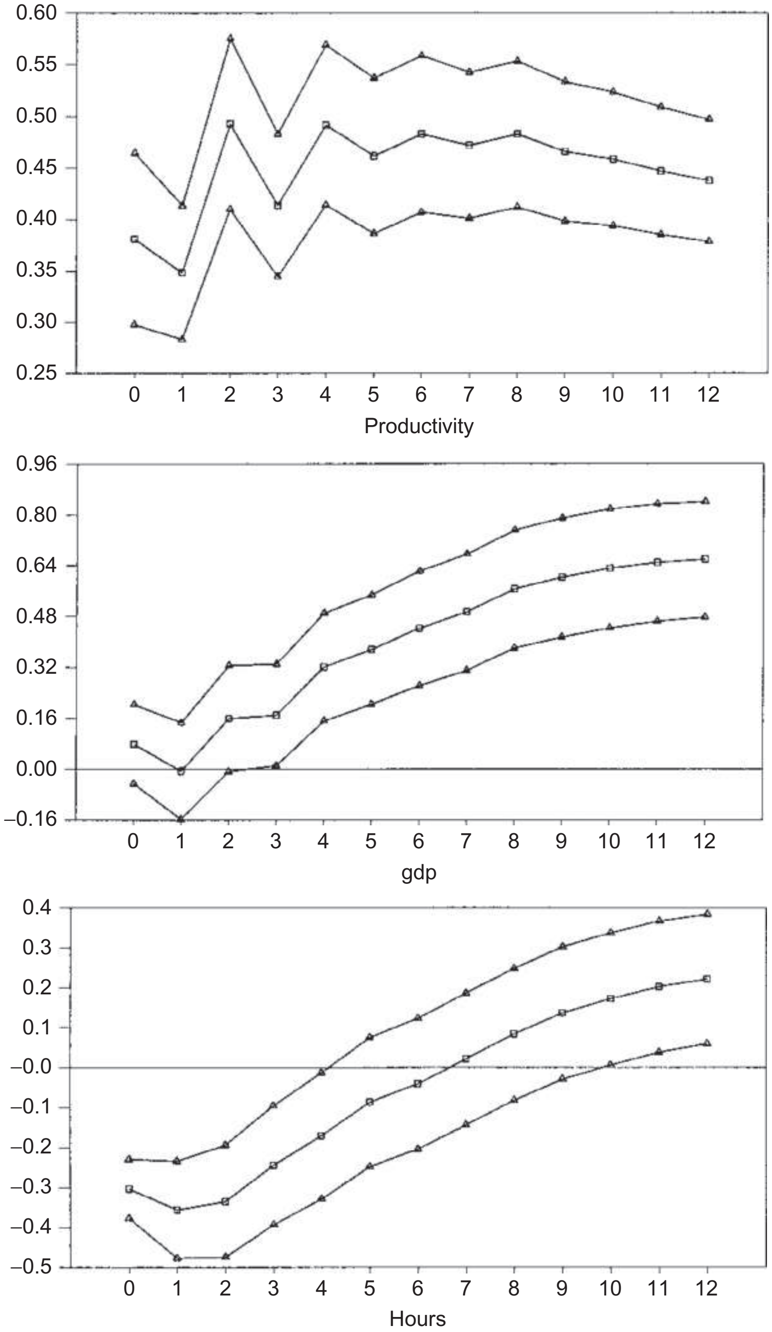


Fig. 5 Estimated impulse responses to a technology shock. Point estimates and ± 2 standard error confidence intervals; variables given in percentages. *Source: From Gali, J., 1999. Technology, employment, and the business cycle: do technology shocks explain aggregate fluctuations?. Am. Econ. Rev. 89 (1), 249-271.*

to "accommodate" the technology shock, that is, it responds to it by contracting aggregate demand *relative* to the level that would have obtained under flexible prices. Finally, suppose that this contractionary effect is strong enough. Then, aggregate output may increase less than productivity. By the same token, equilibrium employment may fall under sticky prices, even though it would have increased under flexible prices.

It remains debatable whether this is a satisfactory theoretical explanation, as well as whether the fact itself is robust to alternative empirical methodologies: see Gali and Rabanal (2005) for complementary evidence, Christiano et al. (2003) and McGrattan (2004) for criticisms. Here, we take the fact for granted and proceed to make two points: one regarding its structural interpretation and one regarding its policy implications.

The first is that, while the fact is inconsistent with the standard, complete-information, RBC model, it is not necessarily inconsistent with an incomplete-information extension of it. The reason is the dampening effect documented in Proposition 18: lack of common knowledge may cause significant inertia in the response of aggregate output, perhaps even a negative initial response in employment.

The second is that this dampening can be a reflection of the efficient use of the information that is dispersed in the economy.[bq] This contrasts the New-Keynesian explanation, which prescribes the observed fact to a failure of the monetary authority to replicate flexible-price allocations. To put it differently, the success of the New-Keynesian model in matching the fact in Gali (1999) is, in the context of that model, a manifestation of the failure of the policy maker to do the right thing. By contrast, augmenting the RBC model with a friction in coordination can help rationalize the fact without any policy failure.

These points are formalized in Angeletos and La'O (2010) by studying an incomplete-information version of the neoclassical economy we introduced in Section 2.2. This economy is essentially the same as the textbook RBC model, except for three modifications. First, capital is assumed away in order to simplify the analysis. Second, product differentiation is introduced, giving rise to strategic complementarity. Lastly, information is incomplete, giving rise to imperfect coordination.

More specifically, the economy is split into a large number of islands, indexed by $i \in [0, 1]$. Each island is populated by a competitive firm and a competitive worker, who produce a differentiated, island-specific, intermediate good. The goods produced by different islands enter the production of a single, final, consumption good. This introduces strategic complementarity across islands: the optimal production on each island depends on other islands' production. Incomplete information is introduced by assuming the following: when the firm and the worker of any given island decide on the local employment and output, they are uncertain about the productivity, employment and output of other islands. Despite the incompleteness of information, incomplete risk-sharing is assumed away by letting workers pool their income at the end of each period: everybody

[bq] By "efficient" we mean the notion of constrained efficiency discussed in Section 9.

belongs to the same "big family," whose income and consumption is fully diversified against the idiosyncratic information and choices of the various family-members. Finally, it is assumed that the local productivity of each island is perfectly observed by the firm and the worker of the island,[br] which guarantees that any other information is relevant only insofar as it helps predict the production levels of other islands.

To facilitate an exact mapping to the class of beauty-contest games we studied in the previous section, information is assumed to be Gaussian and preferences and technologies are assumed to take a power-form specification. Without further loss of generality, let us impose a linear technology: $y_{it} = a_{it} + n_{it}$, where y_{it} is the (log)level of the output produced in island i, a_{it} is the (log)level of the exogenous local productivity, and n_{it} is the (log) level of local employment. Angeletos and La'O (2010) show that the general equilibrium of the model reduces to the solution of the following fixed-point relation:

$$y_{it} = (1-\alpha)\chi a_{i,t} + \alpha \mathbb{E}_{it} Y_t \tag{29}$$

where Y_i is the (log)level of aggregate output, and $\chi > 0$ and $\alpha < 1$ are constants that depend on the underlying preference and technology parameters. With the output levels obtained from the solution of (29), local and aggregate employment are given by, respectively, $n_{it} = y_{it} - a_{it}$ and $N_t = Y_t - A_t$.

Clearly, the fixed-point relation in (29) is mathematically equivalent to the best-response condition encountered in the previous section. It follows that the insights developed there are readily applicable to the present context. We now elaborate what this means for the empirical properties of the RBC framework.

Assume that local productivity is given by the sum of an aggregate and an idiosyncratic component: $a_{it} = A_t + \xi_{it}$, where A_t is the aggregate technology shock and ξ_{it} is an idiosyncratic one. The latter is i.i.d. across both i and t (for simplicity), whereas the former follows a random walk:

$$A_t = A_{t-1} + u_t,$$

where u_t is the period-t innovation.

When information is complete (equivalently, coordination is perfect), we have that $\mathbb{E}_{it} Y_t = Y_t$ for all i. Substituting this into the above fixed-point relation and aggregating across i yields the following:

Proposition 26 *Under complete information, aggregate output and employment are given by, respectively,*

$$Y_t = \chi A_t \quad \textit{and} \quad N_t = (\chi - 1) A_t.$$

It follows that equilibrium outcomes are pinned down by the technology shock; that χ and $\chi - 1$ identify the general-equilibrium elasticities of, respectively, aggregate output

[br] This is the analogue of the assumption we made earlier on in Proposition 16.

and aggregate employment to technology shocks, as predicted by the complete-information RBC model; and that, holding χ constant, the scalar α is irrelevant for macroeconomic outcomes.

When instead information is incomplete (equivalently, coordination is imperfect), equilibrium outcomes depend, in effect, on the entire hierarchy of beliefs about A_t. The scalar α then starts playing a crucial role for the observables of the model, because it determines the relative importance of higher-order beliefs.

Proposition 27 *Under incomplete information, aggregate output is given by*

$$Y_t = \chi \sum_{h=0}^{\infty} (1-\alpha)\alpha^h \bar{\mathbb{E}}_t^h [A_t] \tag{30}$$

How α depends on the underlying primitives is discussed in detail in Angeletos and La'O (2010). There are two opposing mechanisms: a "demand-side" effect that contributes towards strategic complementarity ($\alpha > 0$) and a "supply-side" effect contributing towards strategic substitutability ($\alpha < 0$).

Let us explain. When an island expects aggregate output to go up, the demand for the local good, its relative price, and the local real wage are all expected to go up as well. This effect motivates the local worker to work more and the local firm to produce more. This is the "demand-side" effect, which induces local output to increase with expectations of aggregate output. The opposing "supply-side" effect originates from an income effect: when an island expects aggregate output to go up, income is also expected to go up, which tends to discourage labor supply and production. Whether α is positive or negative depends on which of the aforementioned two effects dominates. For the remainder of our analysis, we impose $\alpha > 0$.

If we vary α holding χ constant,[bs] we vary the incomplete-information outcomes without varying their complete-information counterparts. We can thus think of α as a "sufficient statistic" of the primitives of the environment that regulate the macroeconomic effects of strategic uncertainty, and thereby the observable aggregate implications of the incomplete-information model, holding constant the observable aggregate implications of the standard, complete-information macroeconomic model.[bt]

This last observation seems particularly useful if we adopt a "relaxed" interpretation of the model at hand as a representative of a broader class of RBC models in which

[bs] By varying the underlying preference and technology parameters one can match any pair of values (χ, α), which explains why one can indeed vary α holding χ constant.

[bt] Note, however, that the combination of α and χ also matters for the observable implications of both models in the cross section: in the economy under consideration, the product $(1 - \alpha)\chi$ is identified by the micro-elasticity of the response of local output to island-specific productivity shocks. This is indicative of how the combination of micro and macro data could help discipline the mechanisms we study in this section, in a manner that complements the approach reviewed in Section 8.6.

complementarity may originate, not only in the aforementioned demand-side effect, but also in "financial multipliers" or other kinds of market interactions that make the fate of any given firm and consumer sensitive to aggregate economic activity. For the purposes of the subsequent discussion, we invite the reader to entertain such a flexible interpretation of the strategic complementarity and the associated coefficient α.

Let us now go back to the original motivation of this section, namely the negative employment response to technology shocks. As already noted, any empirically plausible calibration of the standard, complete-information, RBC model makes the opposite prediction. In the context of the present framework (which assumes away capital), this prediction is equivalent to imposing $\chi > 1$, but leaves α unrestricted. But once information is incomplete, the predicted employment response crucially depends on α and can turn negative irrespective of the value of χ. That is, we can reconcile the RBC framework with Gali's findings by removing common knowledge of the technology shock and letting α be large enough.

Angeletos and La'O (2010) demonstrate the above possibility with numerical examples that allow learning to be slow and the incompleteness of information to persist indefinitely. Here, we illustrate the key idea with an example that forces the incompleteness of information to be transient, lasting for only one period.

Assumption 9 In each period t, the previous-period aggregate technology shock, A_{t-1}, becomes commonly known. Furthermore, for every i, any information that island i has about the current-period shock, A_t, is summarized in a private signal of the form

$$x_{it} = A_t + \epsilon_{it},$$

where $\epsilon_{it} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is i.i.d. across (i, t) and independent of A_t for all t.[bu]

The first part of this assumption rules out noisy dynamic learning and guarantees that the economy reduces, in effect, to a repeated version of the static beauty-contest game studied in Section 7: in each period, it is as if the islands play a beauty-contest game in which the fundamental $\theta = A_t$ and the common prior about it is a Normal distribution with mean $\mu = A_{t-1}$. The second part of the assumption abstracts from public signals and any other form of correlated noise, thus isolating the mechanism we care about in this section, namely the role of strategic uncertainty in dampening the response to fundamentals. The alternative function of accommodating forces akin to animal spirits is the subject of Section 8.7.

Following an approach similar to that used in the proof of Proposition 18, we can establish the following.

[bu] Because local productivity is itself a private (local) signal of aggregate productivity, x_{it} is meant to be a sufficient statistic for both the information contained in local productivity and any other private information that island i has about A_t.

Proposition 28 *Under Assumption 9, equilibrium output is given by*

$$Y_t = \chi A_{t-1} + \phi u_t,$$

for some positive scalars ϕ that satisfy $0 < \phi < \chi$. Furthermore, for any given $\chi > 0$ and $\sigma_\epsilon > 0$, ϕ converges to 0 from above as α converges to 1 from below.

This result illustrates two properties that go beyond the specific information structure assumed above. The first property is that the *long-run* effect of a technology shock under incomplete information remains the same as the one under complete information (which is χ). The second property is that the short-run effect (given by ϕ) is smaller and can even be arbitrarily close to zero if α is large enough. The first property holds because the shock becomes common knowledge in the long run (where "long run" means "next periods" in the above example and "asymptotically" in the example studied below). The second property holds because, as long as the fundamental is not common knowledge, the following is true: as h goes to ∞, the hth order belief of A_t converges to the common prior, no matter what the current innovation u_t is; letting α be high enough therefore guarantees that the equilibrium belief of Y_t does not move much with u_t, and that actual output does not move much either.

To derive the aggregate employment response, recall that $N_t = Y_t - A_t$. It follows that the effect of the technology shock on aggregate employment is given by $\phi - 1$, which is negative for high enough α.

Corollary 3 *Suppose $\chi > 1$ and take for granted Gali's finding that the short-run response of employment to technology shocks is negative. This finding rejects the frictionless version of the RBC model, but does not reject its incomplete-information extension insofar as α is large enough.*

The above result has so far been established only under a narrow interpretation of what the "short run" is: the response of employment turns positive after a one-period lag. However, this last property is an artifact of the simplifying assumption that the technology shock becomes common knowledge after a lag of only one period: if we allow the lack of common knowledge to persist for more periods, then we can accommodate a negative employment response for more periods. See Angeletos and La'O (2010) for details.

Angeletos and La'O (2010) further show that the equilibrium response of the economy to technology shocks is constrained efficient in the sense that there is no allocation that can improve upon the equilibrium without violating either resource feasibility or the informational constraints of the economy.[bv] It follows that, unlike the New-Keynesian interpretation, the Gali fact is not more a symptom of suboptimality (or constraints) in monetary policy.

To recap, insofar as there is sufficient lack of common knowledge and sufficient strategic complementary, the RBC framework can be reconciled with the Gali fact. This, however, pegs the question of how strong the relevant strategic complementary is. In

[bv] See Section 9 for a definition and a discussion of this kind of efficiency concept.

Angeletos and La'O (2010), the only source of strategic complementarity is that induced by the Dixit-Stiglitz preference specification. For conventional parameterizations, this implies a rather low value for α, and therefore also a rather low role for higher-order uncertainty. Financial frictions and feedback effects as those in Kiyotaki and Moore (1997) can "boost" the degree of complementarity that is present in the RBC framework, while at the same time deliver different normative conclusions. A promising direction for future research is therefore to introduce incomplete information in the models of the growing literature on financial frictions that has been spurred by the recent crisis.

We conclude by discussing a few complementary papers which have used informational frictions to generate real rigidity in different contexts, most notably in the canonical consumption-saving problem. Sims (2003, 2006), Luo (2008), and Tutino (2013) show how rational inattention can induce inertia in the response of consumption to income shocks. Luo et al. (2015) show how slow learning can help generate excess smoothness in durable and nondurable consumption, bringing the model closer to the data. Alvarez et al. (2011) study the interaction of transaction and observation costs for liquidity and consumption. The results in all these papers have a similar flavor to those derived here. But there is a key difference. These papers feature no strategic, or general-equilibrium, interaction. As a result, the rigidity they document is a rigidity at the *micro* level: it dampens the response of individual outcomes to individual fundamentals. It also requires enough noise in the observation of such fundamentals; that is, first-order beliefs of fundamentals must themselves be rigid. By contrast, the rigidity we have documented obtains at the *macro* level; it rests on lack of common knowledge (ie, on higher-order beliefs being rigid) as opposed to individual uncertainty; and it can thus be consistent with considerable flexibility at the micro level.

8.2 General-Equilibrium Dampening and Rigidity vs Overshooting

Angeletos and Lian (2016b) push the aforementioned insights further, showing (i) that incomplete information is equivalent to a certain relaxation of the solution concept and (ii) that it can dampen the general-equilibrium effects of macroeconomic models, while holding constant the underlying micro elasticities. This helps reduce the disconnect between recent empirical works such as Mian et al. (2013) and Mian and Sufi (2014), which identify the cross-sectional effects of regional shocks, and the key questions of interest, namely the macroeconomic effects of aggregate shocks.

The results in Angeletos and Lian (2016b) also clarify that the same mechanism—the dampening of general-equilibrium effects—could mean either rigidity or overshooting. If the general-equilibrium effect works in the same direction as the corresponding partial-equilibrium effect, meaning that the macro elasticity is higher than the micro one under complete information, then the introduction of incomplete information contributes to rigidity. If, instead, the general-equilibrium works in the opposite direction, meaning that the macro elasticity is smaller than the micro one under complete information, then the introduction of incomplete information contributes to overshooting.

The first scenario corresponds to settings in which the general-equilibrium effect is akin to strategic complementarity. The second scenario corresponds to settings in which the general-equilibrium effect is akin to strategic substitutability. The application by Angeletos and La'O (2010) that we discussed above and the monetary applications by Woodford (2003) and others that we review in the sequel are, in effect, examples of the first case. Angeletos and Lian (2016b) contain examples of the second case.

An interesting example of the second case is also Venkateswaran (2014). This paper considers an incomplete-information version of the Diamond–Mortesen–Pissarides model and shows that the incompleteness of information increases the volatility in aggregate unemployment, helping reconcile the model with the business-cycle data.

8.3 Nominal Rigidity and Monetary Shocks

In this section, we review Woodford (2003), an influential contribution that illustrated how incomplete information offers a potent substitute to more conventional formalizations of nominal rigidity.

Before reviewing Woodford's contribution, it is worth to recall some empirical evidence on the macroeconomic effects of monetary shocks. Christiano et al. (1999) and Christiano et al. (2005) use a structural VAR method to identify the impulse responses of macroeconomic variables to monetary shocks. Two of their key findings are the following: (i) Output, consumption, and investment respond in a hump-shaped fashion, peaking after about one to one-and-a-half years and returning to preshock levels after about three years; and (ii) inflation responds also in a hump-shaped fashion, peaking after about two years.

These findings are illustrated in Fig. 6, which is borrowed from fig. 1 of Christiano et al. (2005). The figure shows the estimated responses of output and inflation to an identified monetary shock.[bw] These responses provide support for the New-Keynesian framework insofar as they indicate that monetary shocks have large and persistent real effects. However, they also raise a challenge, which we turn to next.

The baseline version of New-Keynesian Philips Curve, which is at the core of the New-Keynesian model, takes the following form:

$$\pi_t = \kappa y_t + \beta \mathbb{E}_t \pi_{t+1},$$

[bw] The estimated responses of consumption and investment are omitted because we focus on a simple model without capital and consider only the response of output and inflation. Units on the horizontal axis are quarters; on the vertical, it is deviation from the unshocked path (annualized percentage points for inflation, percentages for output). Grey areas indicate 95% confidence intervals.

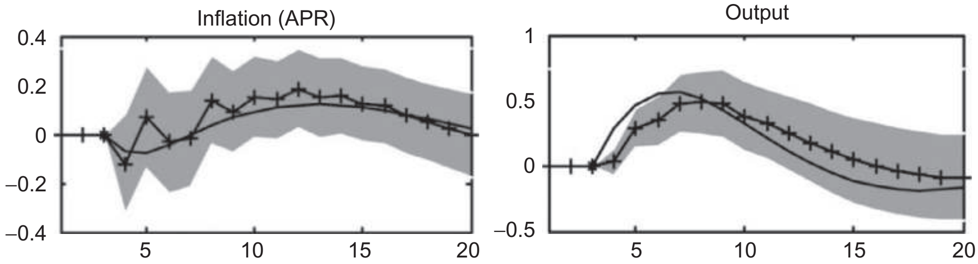


Fig. 6 Impulse responses of inflation and output to monetary shocks. Solid lines with plus signs are VAR-based impulse responses. Grey areas are 95% confidence intervals around the VAR-based estimates. Units on the horizontal axis are quarters. An asterisk indicates the period of the policy shock. The vertical axis units are deviations from the unshocked path. In the case of inflation, the unit is in annualized percentage points (APR). In the case of output, the unit is in percentage. Solid lines are the impulse responses generated by medium-scale new-Keynesian DSGE model (Christiano et al., 2005).

where π_t denotes inflation and y_t the output gap, κ is a function of the Calvo probability of resetting prices (along with preference and technology parameters), and β is the discount factor.[bx] Iterating the above condition gives current inflation as the best forecast of the present value of output gaps:

$$\pi_t = \kappa \mathbb{E}_t \left[\sum_{h=0}^{\infty} \beta^h y_{t+h} \right].$$

This implies that inflation must lead (or predict) output in response to monetary shocks: if the response of output to a monetary shock is hump-shaped as in the right panel of Fig. 6, then the peak in inflation has to come *before* the peak in output. But this is the opposite of what seen in the left panel of Fig. 6.

To overcome this failure, quantitative DSGE models have augmented the baseline New-Keynesian model with a number of ad hoc features, such as price indexation and certain kinds of adjustment costs, whose micro-foundations and immunity to the Lucas critique remain debatable. By contrast, Woodford (2003) shows that incomplete information, couple with strategic complementarity in price-setting decisions, can, not only substitute for the Calvo friction as a source of nominal rigidity, but also naturally produce the empirical pattern seen in Fig. 6.

Let us elaborate. Woodford (2003) considers a monetary economy similar to the one we sketched in Section 2.2. Unlike the New-Keynesian model, there is no Calvo

[bx] Strictly speaking, y_t measures the log deviation of the price-to-cost markup that obtains under sticky prices from the one that would have obtained if prices had been flexible. It is an open question what is the best empirical counterpart of this theoretical object. For simplicity, in the present discussion we interpret y_t as the output gap.

friction: firms are free to reset their prices period by period. Instead, nominal rigidity originates from the incompleteness of information about the underlying monetary shock. In particular, the (log) price set by the typical firm in period t is given by

$$p_{it} = \mathbb{E}_{it}[(1-\alpha)\theta_t + \alpha P_t], \tag{31}$$

where θ_t denotes (log) Nominal GDP, P_t denotes the (log) price level, and α is the degree of strategic complementarity in the firms' price-setting decisions. Nominal GDP is treated as an exogenous process. In line with the data, its growth rate is assumed to follow an AR(1) process:

$$\Delta\theta_t = \rho\Delta\theta_{t-1} + \nu_t,$$

where $\Delta\theta_t \equiv \theta_t - \theta_{t-1}$ is the growth of Nominal GDP, $\rho \in [0, 1)$ is the persistence, and ν_t is the innovation. Finally, the information received by the firm in any given period t is assumed to be a private signal x_{it} of the form

$$x_{it} = \theta_t + \epsilon_{it},$$

where $\epsilon_{it} \sim N\left(0, \sigma_\epsilon^2\right)$ is noise, i.i.d. across (i, t), and independent of θ_τ for all τ.

Woodford (2003) motivates the above information structure on the basis of rational inattention as in Sims (2003): the noise ϵ_{it} is interpreted as the byproduct of a cognitive friction. This helps bypass one of the critiques of the first generation of models that sought to attribute monetary nonneutrality to informational frictions (Lucas, 1972, 1973; Barro, 1976, 1977; Barro, 1978). That earlier literature was based on the assumption that there was no readily available information about the current monetary policy and the current aggregate price level. At face value, this assumption seems unrealistic. However, if agents face cognitive constraints in their ability to attend to and digest the available information, they may well act *as if* that information were unavailable: rational inattention is akin to adding noise in the observation of the underlying shocks.

In the absence of strategic complementarity, the above point would have provided merely a new rationale for the type of informational frictions that were assumed by Lucas and Barro. The hallmark of Woodford's contribution, however, is the interaction between the informational friction and the strategic complementarity in pricing decisions, and the associated role of higher-order beliefs.

As we show in the next section, the setting studied in Lucas (1972) is akin to imposing the restriction $\alpha = 0$ (no strategic complementarity) and thereby shutting down the role of higher-order beliefs. In a nutshell, the theoretical mechanism in Lucas (1972) had to do only with the response of first-order beliefs—which is also why Barro went after measuring "unanticipated changes" in monetary policy.

By contrast, Woodford (2003) shifts the focus to strategic complementarity and higher-order beliefs. As explained in the previous section, the response of higher-order beliefs to innovations in fundamentals is both weaker and more sluggish than that of lower-order beliefs. It follows that, holding constant the precision of the available information and the speed of learning, a stronger complementarity translates into a more muted and more sluggish response of equilibrium prices to the underlying monetary shock—and hence also to larger and more persistent real effects.

Putting the above observations together, Woodford's first contribution is to explain why there can be significant nominal rigidity at the aggregate level even if (i) there is a lot of readily available information about the underlying monetary shocks; and (ii) each firm alone is only modestly unaware of, or confused about, the underlying monetary shocks. The cognitive friction explains why each firm may be confused in the first place; the strategic complementarity explains why there can be significant inertia at the aggregate level even if the aforementioned confusion is modest.

This kind of amplification effect distinguishes Woodford's contribution, not only from the earlier literature by Lucas and Barro, but also from some more recent literature that has proposed "observation costs" as a source of nominal rigidity but has abstracted from strategic interactions. Consider, in particular, Alvarez and Lippi (2014) and Alvarez et al. (2015). These papers use models in which firms update their information sets infrequently because they have to pay a fixed cost whenever they do so. Because the updating is asynchronized, this implies that at any given point of time, firms are differentially informed about the underlying shocks, opening the door to higher-order uncertainty. However, these papers abstract from strategic complementarity in pricing decisions, thus ultimately shutting down the effects of higher-order beliefs.[by] It follows that the macro-level nominal rigidity that is formalized and quantified in those papers is tied to micro-level rigidity. By contrast, Woodford's work highlights that the nominal rigidity can be large at the aggregate, even if the underlying micro-level rigidity is modest. An important open question for future research is how this elementary insight matters for the mapping between micro-level data and macro-level responses developed in Alvarez and Lippi (2014) and Alvarez et al. (2015), and the quantitative conclusions that are drawn with the help of those mappings.

Let us now return to the motivating evidence reviewed in the beginning of this section. How does incomplete information help the model match this evidence? And does it offer a superior structural interpretation of the evidence than that offered by the Calvo friction?

[by] Another difference is that these papers impose that an agent learns *perfectly* the entire state of nature once she updates her information. This implies that, an given date t, an agent who updated her information at $t-j$ ($j \geq 0$) does not face any uncertainty about the beliefs of the set of agents who updated their information prior to $t-j$; she only faces uncertainty about the beliefs of the set of agents who updated their information after she did (which is an empty set if $j=0$). The same point applies to Mankiw and Reis (2002), which we discuss in the sequel.

Woodford (2003) addresses these questions with the help of Fig. 7 (which copies figs. 3 and 4 from the original paper). These figures illustrate the impulse response of inflation and real output to a positive innovation in θ_t, under four alternative parameterizations of ρ, the autocorrelation of the growth rate of Nominal GDP. The right column corresponds to the incomplete-information model studied here; the left column gives, for comparison purposes, the impulse response of a New-Keynesian variant, which replaces the informational friction with standard, Calvo-like, sticky prices.[bz]

For the reasons already explained, incomplete information generates inertia in the response of the price level, P_t, to the underlying shocks. However, the evidence requires that the model delivers two stronger properties: first, we need inertia in the response of *inflation*, not just in that of the price level; second, we need the inertia of inflation to be more pronounced that that of real output.

As is evident in the top panel of Fig. 7, both the incomplete-information and the Calvo model fail to deliver the required empirical properties when $\rho = 0$, ie, when θ_t follows a random walk: in this case, both inflation and output peak on impact (at the moment that the innovation occurs). What is more, Woodford (2003) shows that the two models are isomorphic with regard to the response of inflation and output:

Proposition 29 *Suppose $\rho = 0$. For any parameterization of the Calvo model, there is a parameterization of the incomplete-information model that delivers identical impulse responses for output and inflation, and vice versa.*

This result complements the intuitions we developed in Section 6. Not only can the Calvo friction be thought of as a friction in coordination, we now have a concrete example in which the Calvo friction is observationally equivalent to an information-driven friction in coordination.

That said, this equivalence is exact *only* when $\rho = 0$. When instead $\rho > 0$, this equivalence breaks, indeed in a manner that helps incomplete information outperform the Calvo friction vis-a-vis the data.

Proposition 30 *Suppose that the output response is hump-shaped. For any parameterization, the Calvo model predicts that inflation peaks before output. By contrast, for some parameterizations, the incomplete-information model predicts that inflation peaks after output.*

The first part of the proposition follows directly from the property noted earlier that the New-Keynesian Philips Curve requires that inflation is the best predictor of the present value of future output gaps: insofar as a monetary (or any other) shock causes the output gap to follow a hump-shaped response like the one seen in the data, the NK model predicts that inflation has to peak before output.

[bz] See Woodford (2003) for how the two models are calibrated. The calibration of the Calvo model is based on standard criteria; the calibration of the incomplete-information model is more ad hoc. See, however, Melosi (2014) for an estimated version of Woodford's model, which delivers similar lessons.

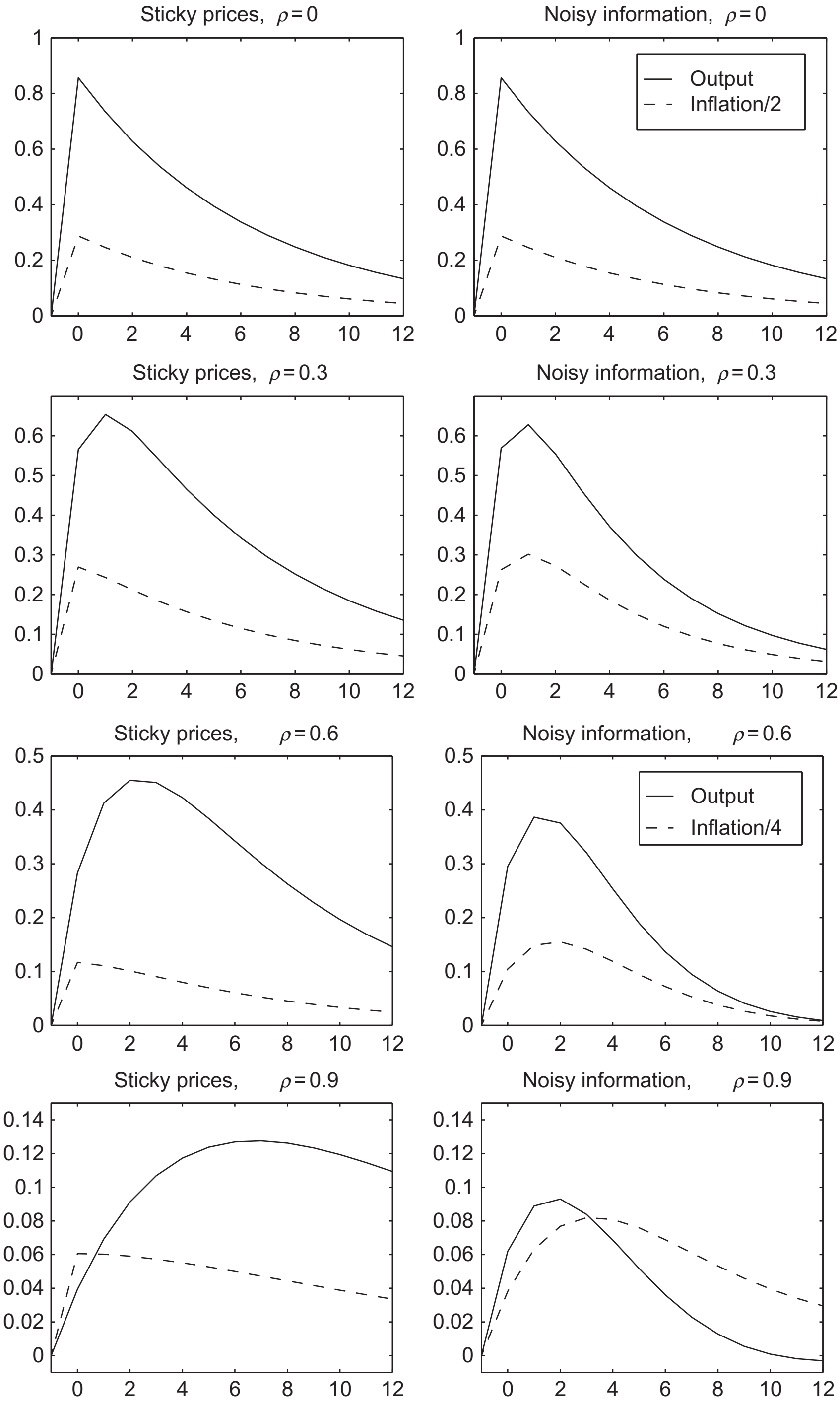


Fig. 7 Comparison of impulse response functions predicted by the textbook new-Keynesian model and by the incomplete-information model of Woodford (2003), for different values of ρ.

The second part follows from the numerical examples presented in Woodford (2003) and repeated in the lower panels of Fig. 7 here: once ρ is sufficiently high (such as $\rho = .6$ or $\rho = .9$ in the figure), the incomplete-information model delivers, not only a hump-shaped response for both inflation and output, but also a peak for inflation that comes later than the peak in output.

Let us explain why. A positive innovation v_t triggers a *gradual* increase in θ_t, from its initial level to a higher long-run level. This means that firms would find it optimal to increase their prices only gradually *even if* they become immediately aware of the shock and could perfectly coordinate their pricing decisions. The fact that each firm becomes only slowly aware about the shock causes each firm alone to delay its response. The fact that each firm expects other firms to do so adds further delay. This delay, however, is necessarily bounded. As time passes, not only first-order but also higher-order beliefs get closer to the true θ_t, which itself converges to its higher-long run level. It follows that the adjustment of prices accelerates after some point. Putting these observations together, we have that inflation is low early on and accelerates later on. By contrast, the growth rate of θ_t is, by assumption, high early on and slows down later on. It follows the difference between the two, which is the growth rate of output, can be high early on, exactly when inflation is still low. This explains why output can peak before inflation in the incomplete-information setting.

To recap, Woodford (2003) highlights how the inertia of higher-order beliefs can help rationalize nominal rigidity at the aggregate level in a manner that is empirically distinct from, and potentially superior to, the conventional Calvo-like formalization. Complementary subsequent work includes Nimark (2008) and Angeletos and La'O (2009), who study the interaction of the two frictions and find they tend to reinforce each other; Melosi (2014), who estimates Woodford's model on the basis of output and inflation data and finds that the model can match the data with a modest level of informational friction; and Hellwig (2005), who elaborates on the micro-foundations and the welfare implications of the model; and Angeletos and Lian (2016c), who study how lack of common knowledge can resolve the forward guidance puzzle. In what follows, we leave aside these extensions and instead relate Woodford's contribution to two other approaches: "sticky information" as in Mankiw and Reis (2002); and "rational inattention" as in Sims (2003) and Maćkowiak and Wiederholt (2009). But before doing this, we make an important parenthesis: we review a simplified version of Lucas (1972) and use this to clarify the novelty of the mechanism we have discussed in this section.

8.4 Parenthesis: Lucas (1972)

Consider the following simplified version of Lucas (1972). The economy consists of overlapping generations. Each agent lives for two periods, working when she is young and consuming when she is old. There is a continuum of islands, $i \in [0,1]$. In each period t, each island has a continuum of young and old agents.

Fix a period t and an island i. Consider the young agents who are born in that period and work in that island. There is a measure one of such agents and they are all identical. Each of them supplies $N_{i,t}$ units of labor to produce

$$Y_{i,t} = N_{i,t} \tag{32}$$

units of the island's good. Let $P_{i,t}$ denote the nominal price of that good in period t.

In period $t + 1$, the aforementioned agents become old; they are randomly relocated to different islands, according to an assignment rule that will be described shortly; and they each receive a monetary transfer that is proportional to the nominal income they made when young. At the aggregate level, the monetary transfer is pinned down by the growth rate of money supply. The latter follows a random walk in logs:

$$M_{t+1} = M_t e^{\nu_{t+1}},$$

where $\nu_{t+1} \sim N\left(0, \sigma_\nu^2\right)$ is an aggregate shock. It follows that the cash-in-hand of an old agent, who was born in island i and is currently (ie, in period $t + 1$) located in island j, is given by

$$M_{i,t+1} = P_{i,t} N_{i,t} e^{\nu_{t+1}}.$$

Her budget constraint is given by

$$P_{j,t+1} C_{i,j,t+1} = M_{i,t+1},$$

where $C_{i,j,t+1}$ denotes her consumption, Finally, the her realized utility is given by

$$\mathcal{U}_{i,j,t} = C_{i,j,t+1} - \frac{1}{1+\kappa} N_{i,t}^{1+\kappa}.$$

We now elaborate on the assignment of old agents to different islands and the resulting demand for the local good of each island. In each period t, the old agents on any given island are a representative sample of all the agents who were born in the previous period and work on different islands. However, different islands receive samples of different sizes. In particular, the mass of old agents that island i receives in period t is given by $\bar{\xi} e^{\xi_{i,t}}$, where $\xi_{i,t} \sim N\left(0, \sigma_\xi^2\right)$ is i.i.d. across islands and across periods, as well as independent from ν_t. We set $\bar{\xi}$ so that $\mathbb{E}[\bar{\xi} e^{\xi_{i,t}}] = 1$, that is, the average mass is one. It follows that the nominal demand on island i during period t is given by

$$D_{i,t} = \bar{\xi} e^{\xi_{i,t}} M_t,$$

where M_t is the aggregate quantity of money. Market clearing then imposes

$$P_{i,t} N_{i,t} = D_{i,t}. \tag{33}$$

The modeling role of $\xi_{i,t}$ is therefore to induce island-specific demand shocks—or, equivalently, variation in relative prices.

The above assumptions guarantee a simple characterization of the optimal labor supply of the young agents. From the budget constraint, we have that the consumption of a young agent who works on island i in period t and consumes on island j in period $t+1$ is given by

$$C_{i,j,t+1} = \frac{M_{i,t+1}}{P_{j,t+1}} = \frac{P_{i,t} N_{it} e^{\nu_{t+1}}}{P_{j,t+1}} = \frac{P_{i,t}}{M_t} \cdot \frac{M_{t+1}}{P_{j,t+1}} \cdot N_{i,t}.$$

It follows that the optimal labor supply of that agent in period t is given by

$$N_{i,t} = arg \max_N \mathbb{E}_{i,t}\left[\frac{P_{i,t}}{M_t} \cdot \frac{M_{t+1}}{P_{j,t+1}} \cdot N - \frac{1}{1+\kappa} N_{i,t}^{1+\kappa}\right] = \left(\mathbb{E}_{i,t}\left[\frac{P_{i,t}}{M_t} \cdot \frac{M_{t+1}}{P_{j,t+1}}\right]\right)^{\frac{1}{\kappa}}. \tag{34}$$

What remains then is to specify the information structure.

As in Lucas (1972), we assume that *previous-period* money supply, M_{t-1}, is public information, but *current* aggregate money supply, M_t, is unknown; equivalently, the current monetary shock v_t is unknown. We also assume that each agent observes the current nominal price in her island, but not those in other islands. It follows that the information set that enters the expectation in condition (34) is the pair $(M_{t-1}, P_{i,t})$.[ca] Finally, we restrict attention to equilibria in which $P_{i,t}$ is log-normally distributed.

Let lower-case variables denote the logarithm of the corresponding upper-case variables, measured as deviations from steady state, and restrict attention to rational-expectations equilibria in which p_{it} is Normally distributed. There exists a unique equilibrium of this type and is characterized as follows.

Proposition 31 (Lucas 72) *Consider the version of Lucas (1972) described above. There exist scalars $\beta, \lambda \in (0, 1)$ such that the following properties hold:*

(i) *The nominal price level is given by*

$$p_t = (1-\beta) m_t + \beta \bar{\mathbb{E}}[m_t]. \tag{35}$$

(ii) *Real output is given by*

$$y_t = m_t - p_t = \beta\left\{m_t - \bar{\mathbb{E}}[m_t]\right\}. \tag{36}$$

(iii) *The average forecast error of* $\mathrm{m_t}$ *(also known as the "unanticipated" change in the supply of money) is given by*

$$m_t - \bar{\mathbb{E}}[m_t] = \lambda v_t. \tag{37}$$

The combination of parts (ii) and (iii) reveal that monetary shocks have real effects. The intuition is simple. Young agents do not directly observe the shock in money supply.

[ca] We could also allow the agents to observe the entire history of money supply and nominal prices in all past periods. Given the rest of the assumptions we have made and the equilibrium we construct in the sequel, this history provides no additional information.

Instead, they only observe the movement in the local price, which confounds the aggregate monetary shock ν_t with the island-specific demand shock ξ_{it}. As a result, they (rationally) confuse nominal price movements for relative price movements. This confusion then explains why it is optimal for young agents to exert more effort and produce more output in response to a monetary shock.

For our purposes, however, it is more useful to focus on part (i). Contrast condition (35) with condition (31), that is, the condition that pins down the price level in Woodford (2003).[cb] It then becomes clear that the price level in Lucas's model depends only on first-order beliefs of the monetary shock (the "fundamental"), whereas in Woodford's model it also depends on higher-order beliefs. By the same token, the degree of monetary nonneutrality obtained in Lucas's model can be large only insofar as agents are uninformed about the monetary shock, which in turn explains why Barro sought to test the theory by measuring unanticipated monetary shocks. By contrast, insofar as there is strong complementarity in price-setting decisions, the degree of monetary nonneutrality obtained by Woodford can be large even if agents are well informed about the monetary shock—for the key is now the lack of common knowledge, as opposed to the lack of individual knowledge.

Remark 28 There are two reasons why higher-order beliefs are absent in Lucas's work. The one is the absence of strategic complementarity, which is emphasized above. The other is the assumption that the previous-period fundamental (m_{t-1}) is commonly known at the begin of each period. If the last assumption is appropriately relaxed, and in particular if different agents within any given island have differential information about the past fundamentals, then higher-order beliefs become relevant through the signal-extraction problem: the interpretation of the local price signal by any young agent depends, in general, on her beliefs regarding the information and the beliefs of old agents, as well as that of other young agents. This kind of mechanism—the role of higher-order beliefs in the interpretation of signals of the activity of others—was the topic of Townsend (1983) and of the subsequent works by Hansen and Sargent (1991) and Kasa (2000). But just as Lucas (1972) abstracted from strategic complementarity in actions, so did these works.

8.5 Sticky Information and Rational Inattention

Having clarified the key difference between Lucas (1972) and Woodford (2003), we now briefly comment on the connection of the latter with two other important recent contributions: that of Mankiw and Reis (2002), and that of Maćkowiak and Wiederholt (2009, 2015).

In Mankiw and Reis (2002), firms are free to adjust their prices continuously over time, but may update their information sets only infrequently. In particular, it is assumed that, in each period, and regardless of its history up to that point, a firm gets to see the

[cb] Note that θ_t in the previous section is the same as m_t presently.

underlying state with probability $\lambda \in (0, 1)$, and it is otherwise stuck with its previous-period information set. By the same token, a fraction λ of firms are perfectly informed within any given period, whereas the rest must set prices on the basis of outdated information. This friction is taken as given in Mankiw and Reis (2002), but it is microfounded on the basis of a fixed cost for observing the state of nature in Reis (2006), and Alvarez et al. (2011, 2015).[cc]

In their paper, Mankiw and Reis focus on the comparison of their model to the standard Calvo model. In particular, they show that their model imposes the following restriction on the joint dynamics of inflation and output, which can be interpreted as the Philips curve of the model:

$$\pi_t = \left(\frac{(1-\alpha)\lambda}{1-\lambda}\right) y_t + \lambda \Sigma_{j=0}^{\infty} (1-\lambda)^j \mathbb{E}_{t-1-j}(\pi_t + (1-\alpha)\Delta y_t),$$

where α is the degree of strategic complementarity in pricing decisions and λ is the aforementioned probability of observing the state. The above result indicates a certain backward-looking aspect, unlike the forward-looking nature of the New-Keynesian Philips Curve: inflation today depends on past information, simply because that past information is relevant for the current pricing choices of firms that have not updated their information.

Notwithstanding the distinct applied contribution of Mankiw and Reis (2002), we now proceed to clarify the manner in which this approach is similar to that of Woodford (2003).

Because the best responses in the two models are the same, the equilibrium price level can be expressed as the *same* function of the hierarchy of beliefs in both models. It follows that the two models can deliver quantitatively different predictions for the dynamics of prices, inflation, and output only to the extent that they happen to feature sufficiently different dynamics in higher-order beliefs.

As with Woodford (2003), we assume that $\Delta\theta_t$ follows an AR(1) process with autocorrelation coefficient $\rho \in [0, 1)$. A clear benchmark emerges when $\rho = 0$, that is, when θ_t follows a random walk. For this case, we already studied the dynamics of higher-order beliefs implied by Woodford's specification of the information structure. Turning to the dynamics implied by Mankiw and Reis's specification, we prove the following.

Proposition 32 *When $\rho = 0$, the first- and higher-order forecasts of θ_t are given by the following:*

[cc] See also the review in Mankiw and Reis (2011).

$$\bar{E}_t^h[\theta_t] = \sum_{j=0}^{+\infty} \left\{ \left(1-(1-\lambda)^{j+1}\right)^h \nu_{t-j} \right\} \quad \forall h \geq 1,$$

where λ ∈ (0, 1) is the probability that a firm updates its information in any given period.

This result gives a closed-form solution for the IRFs of the entire hierarchy of forecasts: the effect of an innovation on the h −th order forecast after j periods is $\left(1-(1-\lambda)^{j+1}\right)^h$, which is clearly increasing in λ, increasing in j, and decreasing in h.

In Woodford's specification such a closed-form solution was not feasible (with the exception of first-order beliefs). Nevertheless, it should now be evident that the qualitative features of the two specifications are the same. In both specifications, higher-order beliefs exhibit a weaker and more sluggish response to fundamentals than lower-order beliefs. Furthermore, the effect that λ has in the above IRFs is essentially the same as the one that the reciprocal of σ_ϵ has in Woodford's model. This is because in both cases these scalars relate to the speed of learning.

We further illustrate these points in Fig. 8. This figure depicts the impulse responses of first- and higher-order beliefs in the sticky-information model, under the maintained assumption that θ_t follows a random walk. Comparing this figure to Fig. 4, we see that the qualitative dynamics of the belief hierarchy is nearly indistinguishable—and therefore so are the impulse response of inflation and output.

Moving beyond the random-walk case, Fig. 9 revisits the last exercise we conducted for Woodford's contribution: it draws the impulse responses of inflation and output to a

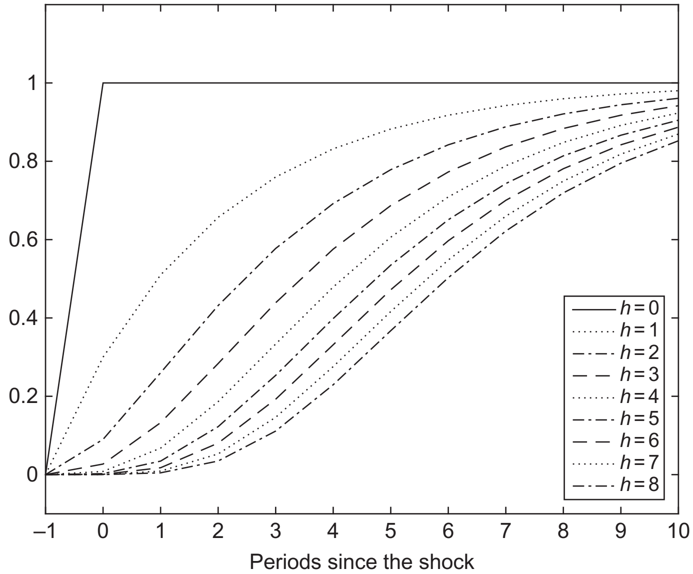


Fig. 8 Impulse response function of higher-order beliefs in sticky information model.

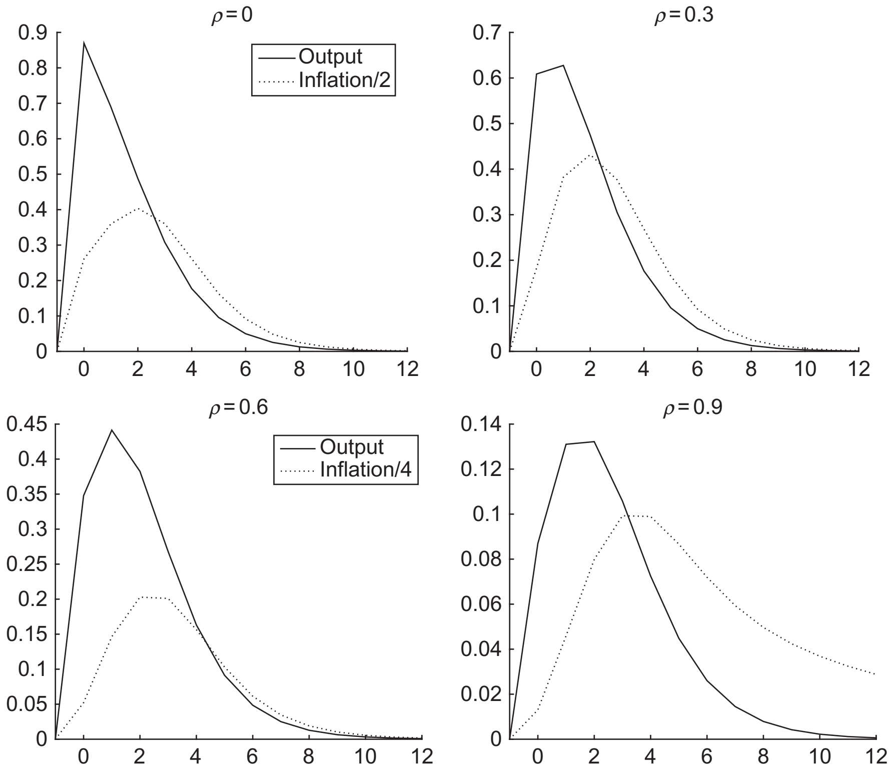


Fig. 9 Impulse response functions of inflation and output in sticky-information model.

positive monetary shock under different values for ρ. A comparison of this figure to Fig. 7 reveals that the predictions of the two models are closely related. The only noticeable difference is that, at least for the chosen parameterizations, the sticky-information model appears to have a relatively easier time in matching the fact that inflation peaks after output, even for low values of ρ.

To recap, Woodford (2003) and Mankiw and Reis (2002) appear to deliver similar dynamics for higher-order beliefs, and thereby also similar dynamics for inflation, output, and the average forecasts of any order. This, however, does not mean that the two models are observationally equivalent (or nearly equivalent) in *every* dimension or in every context. For instance, Woodford's specification implies that the cross-sectional dispersion of forecasts—whether of the exogenous shock or of inflation and all the other endogenous variables—is constant over time, whereas Mankiw and Reis's specification implies that the cross-sectional dispersion of forecasts *increases* following any innovation in the fundamentals.

Some authors have payed special attention to this last property.[cd] In our view, however, this property is not particularly interesting. We, economists, have little knowledge of the precise ways in which real-world people collect, exchange, and digest information. Accordingly, what we would like to retain as a lesson from *both*Woodford (2003) and Mankiw and Reis (2002) is their common predictions with regard to the inertia of higher-order beliefs and the consequent dynamics of inflation and output. These predictions are driven by the combination of strategic complementarity with private learning; they are therefore likely to be robust to other plausible specifications of the information structure.

Let us now consider yet another popular form of informational friction: the form of "rational inattention" proposed by Sims (2003). Because of space constraints, we will not review either the foundations of this approach or the extensive applied literature that followed Sims's original contribution.[ce] Instead, we limit our discussion to the relation of this approach to the central theme of this chapter.

Unlike the more ad hoc alternatives we have studied so far, Sims's formalization of rational inattention is grounded on the idea that, even if arbitrarily precise information is readily available, people may have limited capacity in processing all that information and they may therefore act *as if* their information were noisy. In short, noise is present, not because information is lacking, but rather because of a cognitive friction.

The basic idea is compelling. Applied to the context of monetary policy, it also allows us to avoid a critique of the earlier literature on imperfect information (Lucas, 1972), namely that information about monetary policy and the price level is readily available: even if this were true, rational inattention could explain why firms and consumers may act *as if* they did not have access to all the relevant data.

That said, we would also like to highlight that the foundations of rational inattention are decision theoretic. The issues that are central to this chapter—coordination, strategic or higher-order uncertainty, and solution concepts—were left completely out of the picture in Sims's original contribution.

Subsequent works, most notably those by Maćkowiak and Wiederholt (2009, 2015), applied Sims's approach to general-equilibrium models. The works impose information structures that rule out correlated noise: in the equilibria of those models, aggregate economic outcomes (such as prices and quantities) are pinned down by the aggregate

[cd] Mankiw et al. (2004) argue that this property helps explain certain time-varying patterns in the cross-sectional dispersion of inflation forecast. By contrast, Coibion and Gorodnichenko (2012) find no evidence of this property in the response of inflation forecasts to certain identified shocks (more on this in Section 8.6) and therefore conclude that the data is more in line with noisy information as in Woodford (2003) than with sticky information.

[ce] Important follow-up contributions include Luo (2008), Maćkowiak and Wiederholt (2009, 2015), Paciello and Wiederholt (2014), Matejka (2015a,b), Matejka and Sims (2011), Matejka and McKay (2015), Matejka et al. (2015), Sims (2006), Stevens (2015), Tutino (2013), and Woodford (2009). See also the review in Sims (2010).

fundamentals (such as monetary and productivity shocks). This assumes away the type of "animal spirits" identified earlier. Nevertheless, because agents act as if they observe the fundamentals with *idiosyncratic* noise, the following properties hold: first, information is incomplete (in the sense of Definition 10), not just imperfect (in the sense of Definition 9); second, as long as there is strategic complementarity, the mechanism we studied before regarding the inertia of higher-order beliefs is active.

In this respect, the approach taken in Maćkowiak and Wiederholt (2009, 2015) is closely related, and complementary, to Woodford (2003) and Mankiw and Reis (2002). Nevertheless, Mackowiak and Wiederholt's approach makes two distinct predictions, which are of independent interest.

The first distinct prediction is that the de-facto noise is likely to be much larger for aggregate shocks than for idiosyncratic shocks. This is because of the following. Empirically, idiosyncratic shocks are an order of magnitude more volatile than aggregate shocks. It follows that each individual agent finds it more worthwhile to allocate her attention (capacity) to idiosyncratic shocks rather than to aggregate shocks. What is more, strategic complementarity reinforces this decision-theoretic effect: when other firms pay less attention to aggregate shocks, the individual firm has an even small incentive to pay attention to such shocks.

The second distinct prediction has to do with comparative statics. In a setting where the signals are exogenously specified, one may find it natural to vary elements of the payoff structure—eg, the degree of strategic complementarity in pricing decision, or the specification of the policy rule followed by the monetary authority—keeping constant the precision of the available signals. In a rational-inattention setting, by contrast, the precision of the signals is tied to the underlying payoff characteristics. It follows that the two approaches may make different predictions about the effect of, say, a regime change in monetary policy, even if they deliver the same predictions about the effects of a monetary shock under a fixed policy regime.

Needless to say, the above point is not specific to rational inattention; it applies more generally to any model that endogenizes either the collection or the aggregation of information. Rational inattention imposes a particular structure on this kind of endogeneity. A plausible alternative is developed in Reis (2006) and Alvarez et al. (2015): this approach replaces rational-inattention with fixed costs in the observation of the underlying shocks, but shares the prediction that information is endogenous to policy.[cf]

[cf] This discussion pegs the question of how the endogeneity of information influences the nature of optimal monetary policy. For recent advances into answering this question, see Paciello and Wiederholt (2014) and Angeletos et al. (2016b).

8.6 Survey Evidence on Informational Frictions

In Section 7, we highlighted that complete-information and incomplete-information models have distinct predictions regarding the joint distribution of the aggregate action, the average expectation of the aggregate action, and the underlying fundamental. In the context of monetary models, these predictions regard, inter alia, the joint distribution of actual inflation, the average forecast of inflation, and the underlying shocks.

Consider the textbook New-Keynesian model, or any modern DSGE model that maintains the assumption of complete information. These models predict that firms face no uncertainty about the contemporaneous price level, or inflation, even if they face uncertainty about the underlying shock. It follows that, in these models, $\bar{\mathbb{E}}_t\pi_t$ coincides with π_t in all states of nature. Hence, the impulse response function of π_t to any shock coincides with that of $\bar{\mathbb{E}}_t\pi_t$. By contrast, incomplete-information models such as those in Woodford (2003), Mankiw and Reis (2002), and Maćkowiak and Wiederholt (2009) predict that the two IRFs are distinct, and in particular that $\bar{\mathbb{E}}_t\pi_t$ responds more sluggishly than π_t to innovations in the underlying fundamentals.

In an important recent paper, Coibion and Gorodnichenko (2012) provide evidence in favor of the latter prediction. This paper estimates the impulse response functions of actual inflation and of inflation forecasts to three distinct shocks recovered from the data: a technology shock identified as in Gali (1999); an oil shock identified as in Hamilton (1996); and a news shock identified as in Barsky and Sims (2011).[cg] Inflation forecasts are obtained from four sources: the Survey of Professional Forecasters (SPF); the University of Michigan Survey of Consumers; the Livingston Survey; and the FOMC blue book. For each of these sources, and for each of the aforementioned shocks, it is shown that the average inflation forecasts respond more sluggishly than actual inflation.

This finding is illustrated in Fig. 10, which is borrowed from Coibion and Gorodnichenko (2012). This figure depicts the IRFs of actual inflation, π_t, and of the average forecast error, $\bar{\mathbb{E}}\pi_t - \pi_t$, to each of the aforementioned shocks, using the SPF-based measure for $\bar{\mathbb{E}}\pi_t$. Note that the technology shock and news shock are found to be disinflationary, whereas the oil shock is found to be inflationary. These properties are consistent with previous empirical work, as well as with the predictions of standard macroeconomic models. More importantly for our purposes, note that the response of the average forecast error is found to be negative in the first two cases, positive in the last case, and always lower in absolute value than the actual inflation response. This means that, for all the three shocks, the average forecast $\bar{E}_t\pi_t$ moves in the same direction as

[cg] Given that Woodford (2003), Mankiw and Reis (2002), and Maćkowiak and Wiederholt (2009) were primarily interested in the response of prices to monetary shocks, it may seem peculiar that Coibion and Gorodnichenko (2012) do not study such shocks. Coibion and Gorodnichenko justify this on the basis that identified monetary shocks only drive a relatively small fraction of the business-cycle variation in inflation and economic activity.

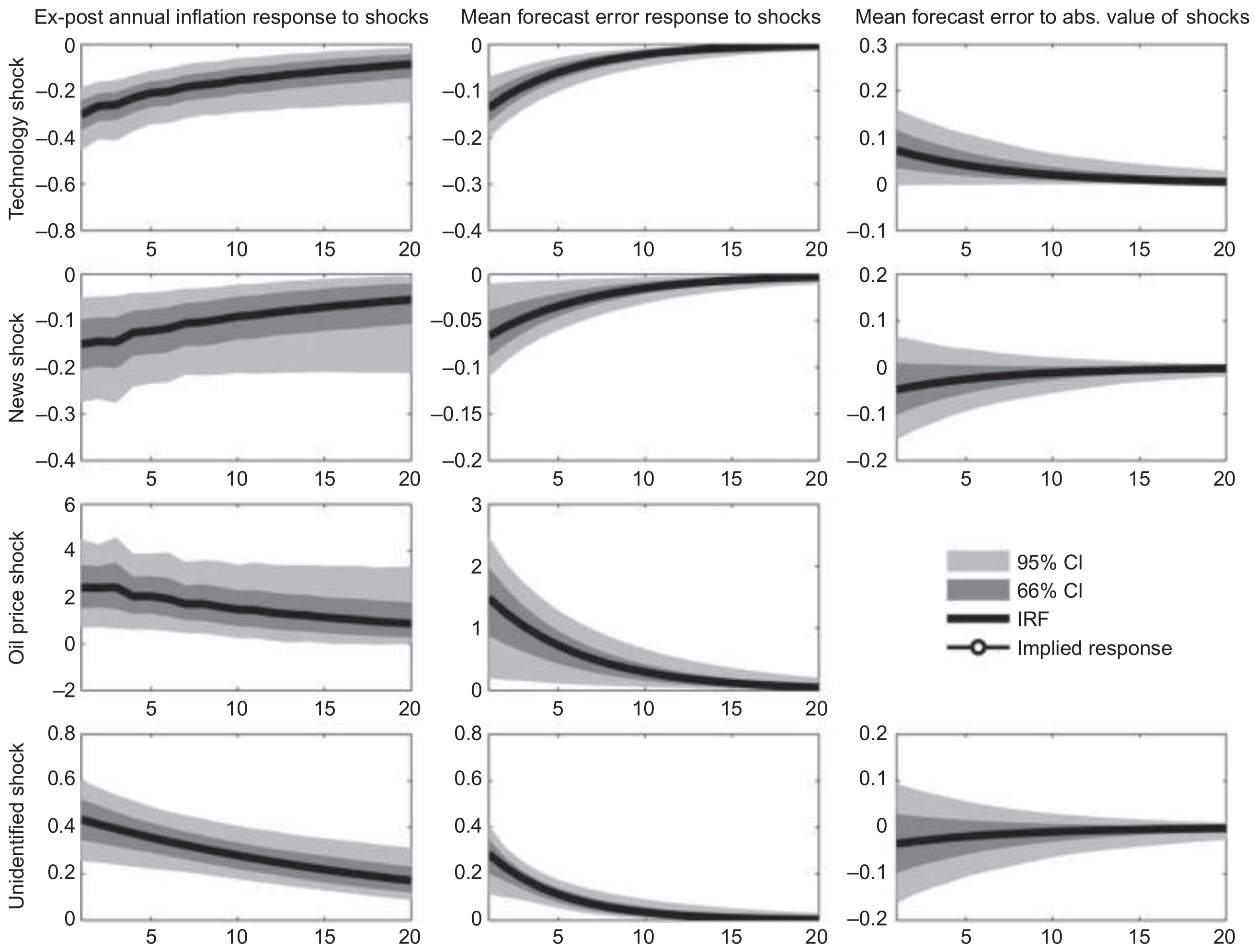


Fig. 10 Response of forecast errors to structural shocks in SPF.

actual inflation, π_t, but also in a more muted manner. Finally, note that the average forecast error vanishes with time.

Coibion and Gorodnichenko (2012) show that the same qualitative patterns hold if the SPF measure of inflation forecasts is replaced by forecasts measures from the University of Michigan Survey of Consumers the Livingston Survey, and the FOMC blue book. To the extent that these measures are good proxies for the average inflation forecast $\bar{\mathbb{E}}_t\pi_t$ in the theory, the above evidence provides prima-facie support against the standard framework and in favor of belief inertia like the one implied by incomplete information.

That said, there is an important disconnect between the empirical exercise and the theoretical models of interest. When Coibion and Gorodnichenko (2012) seek to map their evidence to the theory, they treat the observed inflation dynamics as exogenous to the observed forecast dynamics. By contrast, the models we have reviewed impose that the dynamics of inflation are endogenous to the dynamics of forecasts, indeed in a manner that is central to the predictions of the theory. To put it differently, Mankiw and Reis (2002), Woodford (2003), and others have use informational frictions, not just to explain why the agents' forecasts of inflation may adjust slowly to exogenous shocks, but also to explain why this kind of belief inertia may itself feed into inertia in actual inflation, which in turn may feed into further belief inertia, and so on.

To recap, Coibion and Gorodnichenko (2012) make a significant contribution by documenting salient comovement patterns in the joint responses of forecasts of inflation and of actual inflation to certain shocks. Nevertheless, the precise mapping from that evidence to the theory remains an open question: by treating inflation as an exogenous object, that paper has stopped short of quantifying the equilibrium mechanism that is at the core of the models we have reviewed.[ch]

We conclude by mentioning a few additional works that use survey data to information heterogeneity. Kumar et al. (2015) and Coibion et al. (2015) document widespread dispersion in firms' beliefs about macroeconomic conditions, especially inflation. Coibion and Gorodnichenko (2015) use the relationship between ex-post mean forecast errors and the ex-ante revisions in the average forecast to test the existence of information rigidities. Andrade and Le Bihan (2013), Branch (2007), Carvalho and Nechio (2014), Cavallo et al. (2015), and Sarte (2014) provide additional evidence in favor of information frictions. Combined, these papers provide ample motivations for studying the role of information frictions in the context of business cycles and monetary policy.

[ch] Relatedly, the finding in Coibion and Gorodnichenko (2012) that there is little evidence for strategic complementarity should not be misinterpreted. This finding refers to a rejection of the hypothesis that professional forecasters distort their reported forecasts in an attempt to conform to the forecasts of others. It has nothing to say about the role of strategic complementarity in price-setting behavior or, more generally, in business-cycle phenomena.

8.7 Demand-Driven Fluctuations

In Sections 8.3–8.6 we reviewed a literature that shows how informational frictions can offer a compelling micro-foundation for nominal rigidity. This literature complements the New-Keynesian paradigm, because nominal rigidity is central to this paradigm's ability to explain the observed business-cycles and to accommodate the notion of demand-driven fluctuations. We now turn attention to a different line of work, which goes against the New-Keynesian paradigm: we argue that informational frictions can achieve the aforementioned goals (ie, explain the business-cycle data and accommodate the demand-driven fluctuations) even in the absence of nominal rigidity.

There is a long tradition that formalizes demand-driven fluctuations as the product of "animal spirits" within multiple-equilibrium models. See, eg, Azariadis (1981), Benhabib and Farmer (1994), Cass and Shell (1983), Diamond (1982), Cooper and John (1988), and Guesnerie and Woodford (1993). This approach enjoys limited popularity in the modern business-cycle paradigm, in part because of debatable empirical foundations and in part because of the inconveniences that multiple-equilibrium models carry for estimation and policy evaluation. Nevertheless, what we find appealing with this tradition is that it disentangled the notion of demand-driven fluctuations from that of monetary non-neutrality. This contrasts with the New-Keynesian framework, which gives a central position to nominal rigidity and to monetary policies that fail to replicate flexible prices: in the absence of these features, the notion of demand-driven fluctuations evaporates, and so does the framework's ability to match salient features of the business-cycle data.

But now note that the modern business-cycle paradigm leaves no room for animal spirits, not only because it imposes a unique equilibrium, but also because it rules out imperfect coordination (in the sense we have defined in this chapter). This suggests that the introduction of incomplete information to otherwise canonical unique-equilibrium macroeconomic models may help accommodate a type of fluctuations in expectations and macroeconomic outcomes that resembles the sunspot fluctuations obtained in the older multiple-equilibrium literature, thus also providing a potent formalization of demand-driven fluctuations that does not rest on either nominal rigidity or "mistakes" in monetary policy.

This basic idea was pursued in Angeletos and La'O (2013). The authors consider a convex neoclassical economy in which agents are rational, markets are competitive, the equilibrium is unique, and there is no room for randomization devices. They also rule out aggregate shocks to preferences, technologies, or any other payoff-relevant fundamentals. This shuts down the type of higher-order uncertainty we have studied so far. And yet, the authors are able to obtain aggregate fluctuations, thanks to correlated higher-order beliefs about idiosyncratic trading opportunities.

Let us elaborate on the mechanics of that paper. The structure of the economy is similar to the one used in Section 8.1 to study real rigidity, except for two modifications:

islands are randomly matched in pairs; and in any given period, islands can trade and communicate only with their trading partners.

As in the model introduced in Section 8.1, each island specializes in the production of a specific good and consumes also the good produced by at least one other island. This gives rise to trade and, thereby, to a certain kind of strategic complementarity. Unlike the model of Section 8.1, however, trade is decentralized and takes place through random matching: in each period, each island meets and trades with only one other, randomly selected, island. This implies that game-theoretic representation of the general equilibrium of that model takes the following form:

$$y_{it} = (1-\alpha)\theta_i + \alpha\mathbb{E}_{it}[y_{m(i,t),t}], \tag{38}$$

where y_{it} is the output of island i in period t, θ_i is its exogenous and time-invariant productivity, $\alpha \in (0, 1)$ is the degree of strategic complementarity implied by the underlying preference and technology parameters, and $m(i, t)$ denotes the trading partner (or "match") of island i in period t. It follows that the output of an island depends, not only on its own productivity and the expected productivity of its likely trading partner, but also on what the later expects from its own trading partner, and so on. The authors then proceed to engineer aggregate fluctuations from this kind of higher-order uncertainty.

Formally, the fluctuations obtained in Angeletos and La'O (2013) are aggregate manifestations of exogenous correlated shifts in higher-order beliefs. As discussed before, however, these higher-order belief shifts need to be taken too literally. Rather, they can be interpreted as a device for accommodating nearly self-fulfilling fluctuations in expectations of "demand." This should be evident from equation (38): shifts in higher-order beliefs trigger shifts in actual output because, and only because, they rationalize shifts in the expectations (first-order beliefs) that each island forms about the demand for its product.[ci] What is more, even though the equilibrium is unique, the resulting fluctuations have a similar flavor as those sustained in multiple-equilibrium models: when an island expects more demand for its product, it produces more, which in turn raises the demand for other islands' products. Last but not least, these fluctuations are possible even if the aggregate fundamentals are constant, thus bypassing the limitation discussed in Proposition 21.

A related formalization of aggregate demand fluctuations appears in Benhabib et al. (2015b). As in Angeletos and La'O (2013), there are no aggregate shocks to fundamentals (preferences or technologies) and the equilibrium is unique when information is perfect. But unlike that paper, multiple equilibria obtain once a certain informational friction is introduced. What opens the door to multiplicity is the endogeneity of the information.

[ci] The higher an island's trading partner's output, $y_{m(i,t),t}$, the higher demand for the good produced by the island.

In the model, each firm observes an endogenous private signal about its demand, which in turn depends on the behavior of other firms. In one of the equilibria, which Benhabib et al. (2015b) argue is the most plausible one, aggregate activity is shown to vary with a sunspot. The sunspot is not publicly observable. Instead, the signal that each firm receives about its demand acts partly as an imperfect signal of an idiosyncratic demand shock and partly as an imperfect signal of the aggregate sunspot. What sustains the equilibrium is then the signal-extraction problem that firms face with regard to figuring out whether demand is driven by the one or the other shock.

Complementary are also the works of Gaballo (2015), Chahrour and Gaballo (2015), and Benhabib et al. (2015a, 2016). Gaballo (2015) studies an economy in which final producers are informed about aggregate conditions only through the equilibrium prices of their local inputs; shows that multiple equilibria can arise when idiosyncratic shocks to intermediate production are small; and documents the existence of an interesting equilibrium in which prices are rigid with respect to aggregate shock. Chahrour and Gaballo (2015) show that nontrivial aggregate fluctuations may originate with vanishingly small common shocks to either information or fundamentals. Benhabib et al. (2015a) introduce endogenous information acquisition in a model where firms face both idiosyncratic and aggregate demand shocks and show that endogenous information acquisition makes economic volatility time-varying and countercyclical. Benhabib et al. (2016) show that exuberant financial market sentiments can increase the price of capital, which signals strong fundamentals of the economy to the real side and consequently leads to an actual boom in real output and employment.

What all the aforementioned papers have in common with Angeletos and La'O (2013) is the central role played by incomplete information: if it were not for that feature, the models studied in all these papers would reduce to conventional, unique-equilibrium, neoclassical models, in which equilibrium outcomes would have been pinned down by fundamentals. What, however, distinguishes the aforementioned papers is the emphasis on the signal extraction problems that arise once information is incomplete and the additional volatility—including that in the form of multiple equilibria and sunspot fluctuations—which may obtain from such signal extraction problems.[cj]

Finally, Angeletos and Lian (2016a) study an environment that relates to the aforementioned papers in that it also rests on a signal-extraction problem between aggregate and idiosyncratic shocks, but shifts the emphasis away from multiple equilibria and sunspot fluctuations to a mechanism that helps formalize the notion of a Keynesian multiplier. In particular, a rational confusion between idiosyncratic and aggregate shocks—or at least the lack of common knowledge about aggregate shocks—is shown to explain why

[cj] At some abstract level, all these signal-extraction problems are similar to the one first formalized in Lucas (1972). However, not only are the applications very different, but also higher-order beliefs come into play.

a negative aggregate shock to consumer spending[ck] may lead firms to hire and produce less, which in turns leads consumers to spend less, and so on, ultimately leading to a recession. The same mechanism is then also shown to generate large fiscal multiplier: an exogenous increase in government spending may actually crowd in private consumption, leading to an increase in output that is higher than the exogenous increase in government spending, and helping undo the recession.

Combined, the papers we have discussed in this section indicate how incomplete information helps provide a set of complementary formalizations of the notion of demand-driven business cycles. Importantly, these formalizations do not require–but also do not preclude—either any form of nominal rigidity or any friction in monetary policy. Embracing these formalizations may thus affect, not only the structural interpretation of the available data, but also the policy implication one may wish to draw from the data. Indeed, if nominal rigidity is, by assumption, the only way one can make sense of demand-driven fluctuations, one is *forced* to think about frictions in monetary policy, such as the zero lower bound, as central to the observed phenomena. But if the key relevant mechanism is a friction in information and coordination, monetary policy could be a sideshow.[cl]

Angeletos et al. (2015) and Huo and Takayama (2015a) push this research agenda further by seeking to quantify the aforementioned kind of belief fluctuations. These papers share a similar objective, but take rather different approaches. The first one relaxes the assumption of a common prior in order to accommodate rich, higher-order belief dynamics in an arbitrary linear DSGE model and proceeds to conduct a wide range of horseraces between higher-order beliefs, nominal rigidity, and a variety of structural shocks commonly used in the literature. The second paper maintains the assumption of a common prior, thus disciplining the dynamics of higher-order beliefs, at the expense of a more limited range of quantitative explorations. Both papers nevertheless reach a similar conclusion: the kind of fluctuations in expectations and outcomes that are rationalized by shifts in higher-order beliefs can be quantitatively important and can offer a potent structural interpretation of salient features of the data.

An example of what this means is illustrated in Table 1, which is borrowed from Angeletos et al. (2015). This table compares the empirical performance of five alternative models. The first column in the table reports some key business-cycle moments of the US data; the other five columns report the corresponding moments of the five models. Let us explain what these models are.

[ck] Such shocks may reflect shifts in preferences or, more plausible, be proxies for shocks to consumer credit and/or consumer expectations.

[cl] To be clear, we do not question the empirical relevance of nominal rigidity; we only question whether nominal rigidities and Philips curves are central to understanding either the notion of demand-driven fluctuations or the key regularities of the business-cycle data.

Table 1 HOB shocks in the RBC model vs different kinds of demand shocks in the NK model

			NK with TFP shock plus…			
	Data	**RBC plus HOB**	***I* shock**	***C* shock**	**News shock**	***M* shock**
St. dev(y)	1.42	1.42	1.24	1.15	1.29	1.37
St. dev(h)	1.56	1.52	**1.18**	**0.97**	**1.02**	1.44
St. dev(c)	0.76	0.76	0.86	**0.95**	0.84	0.77
St. dev(i)	5.43	5.66	7.03	7.04	7.24	6.20
Corr(c, y)	0.85	0.77	**0.42**	**0.37**	**0.43**	0.73
Corr(i, y)	0.94	0.92	0.82	**0.75**	0.84	0.90
Corr(h, y)	0.88	0.85	0.80	0.77	0.86	0.84
Corr(c, h)	0.84	**0.34**	**−0.19**	**−0.29**	**−0.07**	**0.24**
Corr(i, h)	0.82	0.99	1.00	1.00	1.00	0.99
Corr(c, i)	0.74	**0.47**	**−0.17**	**−0.33**	**−0.13**	**0.35**

Boldface indicates significant difference between the model and the data.

The first model (column 2) is the authors' baseline model. This is a variant of the textbook RBC model that features only two sources of volatility: a persistent technology shock, A_t, which moves the production possibilities of the economy; and a transitory belief shock, ξ_t, which moves higher-order beliefs of A_t for given A_t. In the equilibrium of this model, variation in ξ_t manifests as waves of optimism and pessimism about the short-term economic outlook. The authors refer to ξ_t as a "confidence shock."

The remaining four models are versions of the New-Keynensian model. All four versions share the same RBC backbone as the baseline model of Angeletos et al. (2015). They differ from it only in two respects. First, they all add sticky prices a la Calvo and a Taylor rule for monetary policy, so as to accommodate the New-Keynesian transmission mechanism. Second, each one of them replaces the confidence shock with one of the following alternative structural shocks, which have been used in the literature as proxies for "demand shocks": an investment-specific shock; a consumption-specific, or discount-rate, shock; a news shock; and a monetary shock.

All five models are calibrated in a comparable manner: the technology and preference parameters, which are common to all models, are set at conventional values; and the stochastic properties of the shocks, which differ across models, are chosen so as to minimize the distance of the model's predicted volatilities for output, hours and investment from the corresponding volatilities in the US data. The models can then be judged in terms of how well they match these targeted moments and, most importantly, how well they match other salient features of the data.

As can be seen in the table, the belief-augmented RBC model does a good job in matching the US data. It also outperforms, in multiple fronts, the competing New-Keynesian models with either the investment shock, or the consumption shock, or

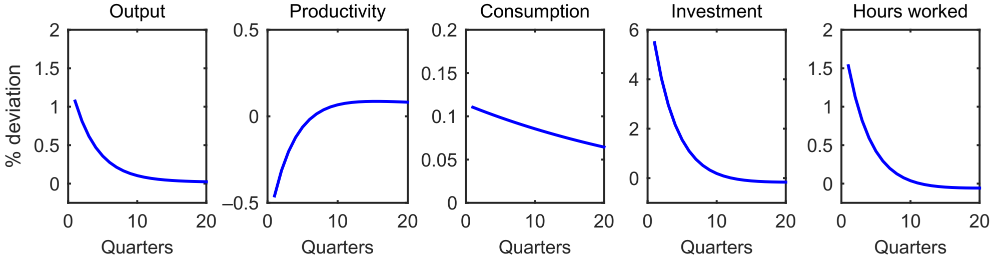


Fig. 11 IRFs to HOB shocks: A formalization of nonmonetary demand shocks.

the news shock. In this sense, the "confidence shock" is superior to these three conventional formalizations of "demand shocks."

The only New-Keynesian model that does as well as the RBC model with confidence shock is the one featuring monetary shock. The problem with that particular model is that its empirical fit rests on allowing for a size of monetary shocks that is an order of magnitude higher than standard estimates of monetary shocks.[cm] Nevertheless, the virtual tie between the two models is a measure of the ability of higher-order uncertainty to provide a potent substitute for the quintessential formalization of "demand shocks."

The reason for these findings can be found in Fig. 11. This figure shows the impulse response functions of the model's key endogenous outcomes to the exogenous shocks in higher-order beliefs. This shock triggers strong comovement in employment, output, consumption, and investment, without a strong comovement in either labor productivity (shown) or inflation (not shown). It is precisely these comovement patterns that are present in the data and that conventional structural shocks have difficulties in matching.

Angeletos et al. (2015) document that these lessons extend to an estimated medium-scale DSGE model that allows for richer propagation mechanisms (sticky prices, habit formation, and investment-adjustment costs) along with multiple structural shocks. In particular, the richer model includes both permanent and transitory TFP shocks, news shocks, investment shocks, consumption-specific shocks, and fiscal and monetary shocks. In the absence of the confidence shock, the estimated model delivers a similar picture as the one found in the extant DSGE literature. But once the confidence shock is included in the model, the picture changes dramatically: the confidence shock is estimated to account for more than half of the business-cycle volatility in output and other macroeconomic quantities.

Huo and Takayama (2015a) complement the above findings by studying a version of the RBC model that allows for a similar type of confidence shocks as in Angeletos and La'O (2013). Relative to Angeletos et al. (2015), the first key difference is the imposition

[cm] Relatedly, if we look at the model's predicted moments for inflation and the nominal interest rate, these moments are far away from their empirical counterparts.

of the common-prior assumption, which disciplines the magnitude and the persistence of the fluctuations in higher-order beliefs that can be entertained in the theory. The second key difference is the use of forecast data, which further discipline the aforementioned fluctuations. Despite these additional "constraints," Huo and Takayama (2015a) reach a similar bottom line as Angeletos et al. (2015): confidence shocks are shown to generate realistic business-cycle patterns and to account for a sizable component of the volatility in the data.

Whether one takes these quantitative findings at face value or not, they offer a very different message from Ramey (2016) in this *Handbook.*That chapter concludes: "we are much closer to understanding the shocks that drive economic fluctuations than we were twenty years ago." We contend that this conclusion hinges on structural interpretations of the data that rule out frictions in coordination and forces akin to market psychology and animal spirits. Once these elements are taken into account, both the interpretation of existing structural VAR evidence and the quantitative performance of existing DSGE models can be seriously upset. The state of our understanding is certainly *different* from what it was twenty years ago. But it is not necessarily closer to the "truth," at least not insofar as the "truth" assigns a prominent role to the type of frictions we have studied in this chapter.[cn]

Last but not least, the works we have reviewed in this section suggest that aggregate demand can be "deficient" during or in the aftermath of certain recessions, not only because of sticky prices and constraints on monetary policy such as the zero-lower bound, but also—and perhaps primarily—because of difficulties in the coordination of the decentralized choices of firms and consumers. How these ideas apply to the Great Recession is an important open question.

8.8 Financial Markets

In this section we discuss how incomplete information can help explain certain asset-pricing puzzles, such as the deviation of asset prices (or exchange rates) from fundamentals, or momentum. To this goal, we use a simple, forward-looking, asset-pricing model with incomplete information, which is the backbone of Futia (1981), Singleton (1987), Allen et al. (2006), Bacchetta and van Wincoop (2006), Kasa et al. (2007), Rondina and Walker (2014), and others. All these paper share the same key structural equation, namely condition (39) below, but make different assumptions about the information structure and the underlying stochastic processes.

[cn] Here is it important to clarify the following. Whenever macroeconomists talk about "shocks" and "propagation mechanisms," they do *not* talk directly about the data (the real world). Shocks and propagation mechanisms are theoretical objects, which are defined in specific models, or in associated structural VARs, and which are used to *interpret* the data.

There is a continuum of agents, or traders, who participate in a competitive market for a risky asset. The market operates for $T+1$ periods, where T can be either finite or infinite. In any period $t \leq T$, the demand for the asset of any given trader i is proportional to her expected excess return of the asset:

$$q_{it} = \mathbb{E}_{it} d_t + \beta \mathbb{E}_{it} p_{t+1} - p_t, \tag{39}$$

where q_{it} is the net position of trader i in period t, p_t is the price of the asset in period t, and d_t is the dividend the asset pays at the end of period t. In period $t = T+1$ (which is valid only if T is finite), we instead have $q_{iT+1} = \mathbb{E}_{iT+1} d_{T+1} - p_{T+1}$, because there is no further trading after that period. In what follows, we focus on what happens in $t \leq T$.

Condition (39), or a slight variation of it, is present in all the aforementioned papers. The papers differ only in the specification of the information structure and of the stochastic properties of d_t and s_t.

In what follows, we treat condition (39) as a primitive structural equation. It is worth noting, however, that this condition can be micro-founded in (at least) two ways.

First, suppose that the traders are risk neutral but need to pay a quadratic "holding" cost for any net position they hold in the risky asset. This means that the trader i's per-period payoff is given by

$$U_{it} = -p_t q_{it} - \frac{1}{2} q_{it}^2 + \left(d_t + \frac{p_{t+1}}{1+r} \right) q_{it}.$$

where r is the risk-free rate. It follows that the optimal demand is given by condition (39), with $\beta = \frac{1}{1+r}$.

Alternatively, suppose that the traders are myopic and have CARA preferences, meaning that their per-period payoff is given by

$$U_{it} = -\frac{1}{\gamma} \exp\left\{ -\gamma \left(-p_t q_{it} + \left(d_t + \frac{p_{t+1}}{1+r} \right) q_{it} \right) \right\},$$

where $\gamma > 0$ is the coefficient of absolute risk aversion. Suppose furthermore that the information structure is Gaussian and let us focus on equilibria in which the equilibrium price is itself a Gaussian signal of the dividend. Then, the optimal demand for the asset i is given by

$$q_{it} = \frac{\mathbb{E}_{it}[d_{it}(1+r) + p_{t+1}] - (1+r)p_t}{\gamma Var_{it}(d_{it}(1+r) + p_{t+1})},$$

where $V\,ar_{it}(X)$ denotes the variance of X conditional on the information of trader i in period t. Finally, suppose that all the underlying shocks have known time-invariant variances and let us focus on equilibria in which the conditional risk faced by the typical trader is also time-invariant. This means that $Var_{it}(d_{it}(1+r) + p_{t+1}) = V$, for some known constant V. The expression above then reduces to (39) if we let $\beta \equiv \frac{1}{1+r}$ and, without serious loss of generality, set $\gamma V = (1 + r)$.

We now put aside these micro-foundations and focus on how the equilibrium price dynamics are affected by the incompleteness of information. To start with, let us make the following basic, but important, observation. In any $t \leq T$, a trader faces two kinds of uncertainty in the return to her investment choice: one about the end-of-period dividend, d_t, and another about the next-period price, p_{t+1}. The first kind of uncertainty regards an exogenous variable; it therefore corresponds to what we have called fundamental uncertainty. The second kind of uncertainty has to do with the demand of future traders; it is therefore an example of strategic uncertainty.

Let s_t denote the exogenous supply of the asset, or equivalently the (negative of the) exogenous demand of any "noise traders." Market clearing requires $\int q_{it} di = s_t$. It follows that the equilibrium price satisfies

$$p_t = \beta \bar{\mathbb{E}}_t p_{t+1} + \bar{\mathbb{E}}_t d_t - s_t, \tag{40}$$

where $\bar{\mathbb{E}}_t$ denotes, as usual, the average expectation at t. Next, define the "fundamental" for this environment as $\theta_t \equiv \bar{\mathbb{E}}_t d_t - s_t$. This is somewhat at odds with the rest of our paper because θ_t now contains an average of first-order beliefs. If one finds this to be confusing, one can henceforth limit attention to the special case in which d_t is known, in which case one can also let $\theta_t \equiv d_t - s_t$. An additional advantage of this special case is that it isolates the role of strategic uncertainty: in the eyes of each trader, the only remaining uncertainty is the one about the behavior of future traders (as manifested in future prices).[co] Either way, condition (21) can be rewritten as

$$p_t = \beta \bar{\mathbb{E}}_t p_{t+1} + \theta_t, \tag{41}$$

which is essentially the same as the forward-looking condition we encountered in Section 7.8. This underscores that the type of strategic uncertainty that is relevant in the present context regards the *future* choices of other agents, as opposed to contemporaneous type of strategic interaction featured in either our abstract static framework or the business-cycles applications studied in Sections 8.1–8.5.

[co] The only minor caveat with this simplification is the following. If T is finite and the last-period dividend is known, the agents face no uncertainty in the last period of trading. Under the first of the two micro-foundations discussed above, the pricing condition $p_{T+1} = d_{T+1} - s_{T+1}$ is still valid. Under the second micro-foundation, however, the traders face no uncertainty and arbitrage now imposes $p_{T+1} = d_{T+1}$. For this case, we must therefore redefine the last-period fundamental as $\theta_{T+1} \equiv d_{T+1}$. (Note that we still have $\theta_t \equiv d_t - s_t$ for all $t \leq T$.)

As in Section 7.8, let

$$z_t^0 \equiv \theta_t, \qquad z_t^1 \equiv \bar{\mathbb{E}}_t z_{t+1}^0 = \bar{\mathbb{E}}_t \theta_{t+1},$$

and, for all $j \geq 2$,

$$z_t^j \equiv \bar{\mathbb{E}}_t z_{t+1}^{j-1} = \bar{\mathbb{E}}_t \{\bar{\mathbb{E}}_{t+1}\{\ldots\{\theta_{t+j}\}\ldots\}\},$$

Iterating (41) yields

$$p_t = \sum_{j=0}^{T+1} \beta^j z_t^j. \tag{42}$$

It follows that the equilibrium price in any given period depends, not only on today's fundamentals (z_t^0) and today's forecasts of tomorrow's fundamentals (z_t^1), but also on today's forecasts of tomorrow's forecasts of the fundamentals two periods ahead (z_t^2), and so on.

This result illustrates that higher-order beliefs of *future* fundamentals can be important determinants of asset prices. Importantly, this can be true even if the current fundamentals happen to be common knowledge: even if all traders are able to reach a common belief about θ_t in *every* period t, they may still fail to reach a common belief about the future demand-and-supply conditions, and can therefore disagree about future price movements. Finally, the relevant notion of the fundamentals now contains, not only the dividend, but also the supply of the asset or, equivalently, the residual demand of noise traders.

The related literature proceeds by making different special assumptions about the length of the horizon T, the stochastic process for the dividend and the supply, and the information that is available in each period. For example, Rondina and Walker (2014) assume that T is infinite, fix the dividend to a known constant (which can be normalized zero), and allow the supply shock to follow an autocorrelated process; they then focus on how higher-order uncertainty about the history of the underlying supply shock and therefore also about the future path of prices can persist even if the traders observe the entire history of past prices. By contrast, Allen et al. (2006) assume that T is finite, that the supply shock is i.i.d. over time, and that the asset pays out a dividend only in the last period; they then focus on the higher-order uncertainty the traders face with regard to the final-period dividend. In the rest of this section, we review some of the key findings of the latter paper as well as that of Bacchetta and van Wincoop (2006).

To nest Allen et al. (2006) in our setting we let the dividend be zero in all but the last period, namely $d_{T+1} = \vartheta \sim \mathcal{N}\left(\gamma, \sigma_\vartheta^2\right)$ and $d_t = 0 \forall 1 \leq t \leq T$; let the supply shock s_t be i.i.d. over time, drawn from $\mathcal{N}(0, \sigma_s^2)$; and finally assume that the traders receive no exogenous information about the supply shocks. The equilibrium price p_t always reveals some information about the current supply shock s_t. However, because the supply shock is i.i.d.

over time, and because the traders receive no exogenous information about future supply shocks, we have that $\bar{\mathbb{E}}_t s_{t+j} = 0$ for all t and all $j > 1$. As a result, condition (42) reduces to the following[cp]:

$$p_t = \beta^{T-t} \bar{\mathbb{E}}_t \bar{\mathbb{E}}_{t+1} \cdots \bar{\mathbb{E}}_{T+1}[\vartheta] - s_t.$$

In a nutshell, Allen et al. (2006) is a special case in which the higher-order uncertainty has a relatively small dimension: it regards only the last-period dividend, as opposed to the entire path of future dividends and supply shocks. Without any loss, we henceforth set $\beta = 1$.

Consider, as a benchmark, the scenario in which the supply is either fixed or commonly known. In equilibrium, the price p_t would perfectly reveal the value of $\bar{\mathbb{E}}_t \bar{\mathbb{E}}_{t+1} \cdots \bar{\mathbb{E}}_{T+1}[\vartheta]$. Under a common prior, the average belief of a random variable can be commonly known only if all the agents share a common belief about that variable. It follows that, in this benchmark scenario, the observation of the equilibrium price induces all traders to form the same belief about the underlying ϑ at all t, and therefore also the same belief about p_{t+1} at all $t < T$. By the same token, there is no speculative trading in this benchmark: all traders would choose the same position and would expect to make zero profits from their trades.

Consider then the alternative case in which the supply shock is present and unknown. In general, the equilibrium price is then only a noisy signal of the average belief $\bar{\mathbb{E}}_t \bar{\mathbb{E}}_{t+1} \cdots \bar{\mathbb{E}}_{T+1}[\vartheta]$. It follows that different traders may maintain different beliefs about this object in equilibrium, which in turn means that they can also maintain different beliefs about either the capital gain $p_{t+1} - p_t$ they can make at any $t < T$, or the final-period yield $\vartheta - p_t$ they can make at $t = T$. In other words, speculation obtains in equilibrium.

The above insights are general. To obtain sharper predictions, however, Allen et al. (2006) impose that all the information available to trader i in period t is given by the combination of the history of prices along with a private signal of the form $x_{it} = \vartheta + \epsilon_{it}$, with $\epsilon_{it} \sim \mathcal{N}(0, \sigma_\epsilon)$ for some $\epsilon > 0$. This in turn is justified by assuming that the traders are short-lived and do not observe any of the private signals of the earlier generations of traders, although they do observe the public history of prices.[cq]

With the assumed information structure, one can guess and verify the existence of an equilibrium in which the price in each period t conveys the same information as a Gaussian public signal of the form

[cp] Strictly speaking, the equilibrium in Allen et al. (2006) entails time-varying risk, which shows up in the demand for the asset (because of the assumption of CARA preferences) and thereby also in the equilibrium condition of the price. Here, we abstract from this nuisance.

[cq] For a discussion of this assumption and for the alternative of allowing long-term traders, see the original paper (Allen et al., 2006) and the references therein.

$$z_t = \vartheta - \chi_t s_t,$$

for some deterministic scalars $\chi = (\chi_t)_{t=0}^{T}$, which are themselves determined by the trading strategies of the agents. Together with the assumption that the private signal is also Gaussian, the above property guarantees a tractable dynamic structure for the posteriors of the traders—namely, a Kalman filter over the time-invariant variable ϑ. As it is standard in noisy rational-expectations settings, the scalars $\chi = (\chi_t)_{t=0}^{T}$ are then characterized by solving the fixed-point relation between (i) the forecasts that each trader forms on the basis of the available price signals and (ii) the price signal generated by the joint behavior of all the traders. The following two results emerge.

Proposition 33 *There exist weights* $\{\lambda_t\}$ *with* $0 < \lambda_T < \lambda_{T-1} < \cdots < \lambda_2 < \lambda_1 < 1$ *such that*

$$E_s(p_t) = \lambda_t \gamma + (1 - \lambda_t)\vartheta$$

where $E_s(.)$ denotes the average value over all the possible supply shocks (equivalently, the expectation conditional on γ and ϑ).

Proposition 34 *For all* $t < T$

$$E_s(|p_t - \vartheta|) > E_s(|\overline{E_t}(\theta) - \vartheta|).$$

Proposition 33 is reminiscent of the inertia effect we documented earlier: the equilibrium price is anchored to the initial prior γ and converges to the fundamental ϑ only sluggishly over time. By the same token, the price path exhibits "momentum": following certain innovations (namely, an innovation in ϑ), prices under-react in the short run relative to the long run; equivalently, they exhibit a drift reminiscent of the ones documented in empirical studies on momentum.[cr]

Proposition 33 complements the above lesson by establishing that the gap between the fundamental and the equilibrium price is more volatile than the gap between the fundamental and the average belief of it. This kind of "excess volatility" reflects the variation that obtains in higher-order beliefs for given first-order beliefs. Proposition 34 is therefore reminiscent of our earlier result regarding the role of incomplete information in sustaining animal spirits.

Combined, these results illustrate how incomplete information and higher-order uncertainty can help standard asset-pricing models accommodate momentum and other interesting volatility patterns in asset prices. In an insightful earlier contribution, Allen et al. (1993) show that higher-order uncertainty can also accommodate "bubbles": there can exist events in which the equilibrium price is higher, not only than the average expectation of the fundamental, but also than the expectation of *every* trader.

[cr] If we replace the private signals that the traders observe in each period t with a public signal of equal precision, then the weight λ_t remains positive but becomes smaller at each t. The dynamic effect documented above is therefore the product, not just of slow learning, but also of the heterogeneity of the information.

This particular possibility is ruled out in the present setting because the Gaussian specification implies that there always exists a trader whose private signal is arbitrarily high, and therefore whose forecast of ϑ is higher than the equilibrium price. Nevertheless, the excess volatility documented in Proposition 34 can be seen as a variant of the same possibility.

The paper by Allen et al. (1993) contains also an insightful discussion of the following issues: the appropriate definition of "fundamentals" in asset-pricing models; the close relation between incomplete-information common-prior settings like the one studied here and the kind of symmetric-information heterogeneous-prior settings studied in, inter alia, Harrison and Kreps (1978) and Scheinkman and Xiong (2003); and the potential richness of the relevant state space. The latter means that the revelation of information through prices can be quite limited in practice, or in sufficiently sophisticated models, even though it can be large in more simplistic models, therefore leaving significant room for higher-order uncertainty to drive asset prices. Complementary in this regard is also the message of Rondina and Walker (2014), although a quantitative evaluation of this insight is still missing.

Let us now shift attention to Bacchetta and van Wincoop (2006). This paper uses a version of the framework we have introduced in this section to argue that incomplete information may explain the disconnect of exchange rates from macroeconomic fundamentals at short to medium horizons. In their model, the asset price is the nominal exchange rate between two countries, which we denote by e_t.[cs] The key equilibrium condition is given by the following[ct]:

$$e_t = \beta \bar{\mathbb{E}}_t e_{t+1} + f_t - b_t,$$

where f_t is the observable difference in money supply between two countries—which is what the authors interpret as the macroeconomic fundamentals—and b_t is an unobserved hedging demand. It follows that Bacchetta and van Wincoop (2006) is nested in our setting simply by the following change in notation: $p_t = e_t$, $d_t = f_t$, and $s_t = b_t$.

The model is closed by assuming the following stochastic structure. The macroeconomic fundamental, f_t, follows a general MA process, $f_t = D(L)\epsilon_t^f$ where $D(L) = d_1 + d_2 L + \cdots$, L is the lag operator, and $\epsilon_t^f \sim N\left(0, \sigma_f^2\right)$. The hedging demand, b_t, follows an AR(1) process: $b_t = \rho_b b_{t-1} + \epsilon_t^b$, where $\epsilon_t^b \sim N\left(0, \sigma_b^2\right)$ and $\rho \in [0, 1)$. In each period, all traders observe the past and the current values of the fundamental and the exchange rate, but not of the hedging demand. In addition, each trader receives a private signal about the *future* fundamental. This is signal is given by $x_t^i = f_{t+\Delta} + \epsilon_t^{xi}$, where

[cs] In Bacchetta and van Wincoop (2006), the exchange rate is denoted by s_t; we use the alternative notation, e_t, in order to avoid any confusion with our earlier notation, which lets s_t denote the supply of the asset.

[ct] As in our discussion of Allen et al. (2006), we omit some potential time-varying coefficients on f_t and b_t.

$\epsilon_t^{xi} \sim \mathcal{N}(0, \sigma_x^2)$ is independent from f_τ for all τ, as well as of the noises in other agents' signals, and $\Delta \geq 1$.

The model's predictions are illustrated in Fig. 12, which we borrow from fig. 2 in Bacchetta and van Wincoop (2006).[cu] Panel A shows that the exchange rate responds sluggishly to innovations in the fundamental, a recurring theme of this chapter. At the same time, the exchange rate responds nontrivially to shocks in hedging demand, and more so than in the complete-information variant of the model. Both effects originate in rational confusion at the individual level, but get amplified by our familiar beauty-contest mechanism.

The amplification effect of hedging demand is evident in Panel C, which reports the contribution of the hedging demands to the variance of the exchange-rate change $e_{t+k} - e_t$ at different horizons $k \geq 1$. Although hedging shocks contribute to short-run volatility in both the incomplete-information and common-knowledge versions of the model, their contribution is far greater in the former case. The mirror-image of this result is Panel D, which reports the R^2 of the regression of $e_{t+k} - e_t$ on the realization of the fundamentals up to $t + k$. In the short run, the R^2 is far lower in the incomplete-information version of model than in its common-knowledge counterpart. In the long run, R^2 converges to 1 in both models, reflecting the fact that both first- and higher-order beliefs converge to the realized fundamental as time passes and more and more information is accumulated. However, the convergence is slower in the incomplete-information version, due to the fact that higher-order beliefs converge more slowly than first-order beliefs. These properties are consistent with empirical findings that macroeconomic fundamentals have weak explanatory power for exchange rates in the short to medium run (Meese and Rogoff, 1983), but play a much more important role over longer horizons.

Complementary to the above works are Biais and Bossaerts (1998), Kasa et al. (2007), Makarov and Rytchkov (2012), and Rondina and Walker (2014). The first paper is an early contribution that also touches on the beauty-contest aspect of asset markets and explores on the distinct positive implications of incomplete information and heterogeneous priors in a finite-horizon model with a finite number of agents. The other three papers use frequency-domain techniques to solve infinite-horizon models, where there are recurring shocks to fundamentals and the traders' beliefs remain perpetually heterogeneous. Combined, these papers show how higher-order belief dynamics can explain apparent violations of variance bounds and other cross-equation restrictions that guide the empirical test of representative-agent models; how they can accommodate momentum; and how can also generate endogenous boom-and-bust cycles.

Two other important contributions are Cespa and Vives (2012, 2015). These papers revisit the question of whether asset prices are closer or further away from the underlying fundamentals than the consensus (average) forecast of fundamentals. Whereas the aforementioned works by Allen et al. (2006), Bacchetta and van Wincoop (2006), and others answer this question always in the affirmative, Cespa and Vives (2012, 2015) show that

[cu] Panel B is omitted because it is not useful for our discussion.

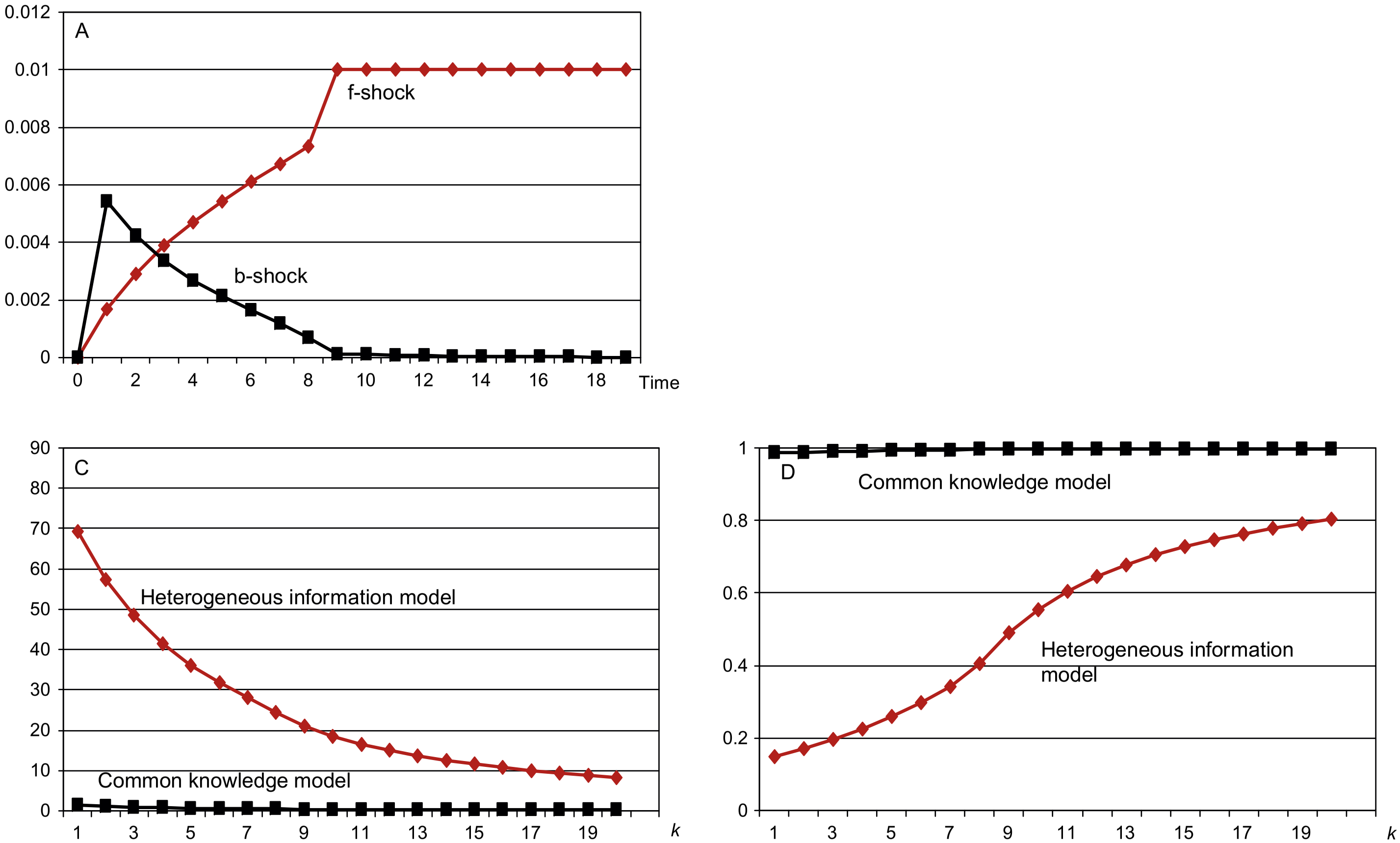


Fig. 12 Exchange rate disconnect in the short and long run. (A) Impulse response functions in heterogeneous information model. (C) Percent contribution b-shocks to $var(s_{t+k} - s_t)$. (D) Connection between exchange rate and observed fundamentals: R^2 of regression of $(s_{t+k} - s_t)$ on observed fundamentals.

the answer depends critically on the persistence of the shocks to liquidity (noise) trades, which is typically assumed to be zero in the literature.

More specifically, Cespa and Vives (2012) studies a model similar to that in Allen et al. (2006), except for two modifications: the informed, rational traders are long-lived; and the unobserved shocks to liquidity (noise) trades are allowed to be persistent. When this persistence is sufficiently low, equilibrium prices are systematically farther away from fundamentals than the average expectation, a pattern consistent with Allen et al. (2006) and Bacchetta and van Wincoop (2006). But when liquidity trades are sufficiently persistent, the opposite pattern emerges: rational traders chase long-term returns rather that short-term speculative gains, causing prices to be systematically closer to fundamentals than average expectations. Cespa and Vives (2015) provide a complementary result in a model where rational traders are short-lived. It nevertheless remains true that incomplete information can generate interesting time-series patterns, such as momentum.[cv]

All in all, the works we have reviewed in this section illustrate how higher-order uncertainty can offer a parsimonious explanation to various puzzles in asset pricing and international finance. These works are part of a broader literature that uses informational frictions—although not always higher-order uncertainty—in asset-pricing contexts. Grossman and Stiglitz (1980), Hellwig (1980), and Kyle (1985) are classics, which emphasize the aggregation of information through equilibrium prices. More recently, Albagli et al. (2014, 2015) develop an extension that accommodates more general specifications of the joint distribution of the dividend and the private signals. This helps shed light on the differential effect that the heterogeneity of information can have on different claims on the same underlying asset, such as the bonds and the stocks of the same firm. Hassan and Mertens (2014a), on the other hand, develop a framework for augmenting DSGE models with a noisy rational-expectations asset market, and use that to study the macroeconomic effects of the news contained in asset prices about future fundamentals. Notwithstanding the contributions of these papers, it is useful to note that none of them study the beauty-contest effects that are the core of the papers we reviewed above, for they study setting in which either the fundamentals become common knowledge at the end of each period (Hassan and Mertens, 2014a) or trading takes place only once (all the other papers mentioned in this paragraph).[cw]

More closely related to the papers we have reviewed above are, instead, Angeletos et al. (2010) and Hassan and Mertens (2011, 2014b). Angeletos et al. (2010) show that

[cv] What is more, Cespa and Vives (2015) show that the combination of incomplete information with short-term horizons may open the door to multiple equilibria, which are ranked both in terms of their informational efficiency and in terms of the strength of the momentum effect.

[cw] This statement should not be misinterpreted as a claim that that higher-order beliefs are entirely irrelevant. Because the agents extract a signal from prices, and because this signal is endogenous to the trades of other agents, higher-order beliefs matter in all the mentioned papers for essentially the same reason as the one suggested by Townsend (1983). But this mechanism, which has to do with the interpretation of endogenous signals, is different from the mechanism we discussed above, which has to do with speculation.

the information spillover between the real and the financial sector induces beauty-contest features in IPO activity and real investment, leading to excessive waves of optimism and pessimism in investment and asset prices. Hassan and Mertens (2011, 2014b) show how financial trading may amplify small correlated biases in traders' beliefs, while also reducing the individual incentive to correct one's beliefs from such biases, and how this mechanism can ultimately distort real investment. In these papers, strategic complementarity emerges endogenously and is key to the obtained results.

Finally, there is a growing strand of the asset-pricing literature that shifts the focus to the acquisition of information or the attention allocation (eg, Veldkamp, 2006; Van Nieuwerburgh and Veldkamp, 2009, 2010; Abel et al., 2007, 2013), but is beyond the scope of this chapter. For more thorough discussions of the broader asset-pricing literature on informational frictions, we refer the reader to Vives (1993) and Veldkamp (2011). A similar point applies to Goldstein et al. (2011), which focuses on trading frenzies; and to Goldstein et al. (2013), which focuses on the signal-extraction problem of a central bank in the context of self-fulfilling financial crises.

We conclude by mentioning a line of work that touches on similar mechanisms, albeit with different modeling techniques. This is the asset-pricing literature on heterogeneous beliefs and speculation, such as Harrison and Kreps (1978) and Scheinkman and Xiong (2003). What is common between this literature and the works we reviewed above is that equilibrium trades and asset prices are driven by higher-order beliefs. What is different is that variation in the relevant higher-order beliefs is engineered with the combination of heterogeneous priors and the arrival of public information over time, as opposed to dispersed private information. This means a great deal of tractability, the absence of which could have impeded the discovery of important results. But it also comes with three potential costs. First, this approach permits the theorist to choose higher-order beliefs at will, a freedom that must be exercised with great moderation. Second, this approach shuts down the role of the price mechanism in aggregating information and coordinating beliefs, a mechanism that is of interest on its own right. Last but not least, this approach invites tricky normative questions, such as which prior must be used to evaluate the welfare implications of different allocations.[cx] Whether one should use this approach or the more "conservative" one employed by the papers we reviewed in this section is ultimately a function of one's research objectives.[cy]

[cx] One possible answer to this last question is that the planner must evaluate the utility of each agent under the agent's *own* prior. This is effectively the assumption made in Debreu's classic proof of the two welfare theorems in economies with subjective probabilities. There are, however, other plausible, and potentially preferable, answers to the aforementioned question. See, for example, Brunnermeier et al. (2014) for a welfare criterion that instructs the planner to discard an allocation only if it is deemed inferior under multiple priors at once.

[cy] The discussion of this paragraph echoes our earlier discussion in Section 7.10 regarding solution methods and the heterogeneous-prior approach taken in Angeletos et al. (2015). Similar points apply to the recent literature on heterogeneous beliefs and leverage, such as Geanakoplos (2010) and Simsek (2013).

9. EFFICIENCY AND POLICY IMPLICATIONS

Although we often touched on policy implications in the preceding sections, we have not systematically studied either the normative content of the positive properties we have documented, or the implications of coordination frictions for the type of Ramsey policy problems that are commonly used in the study of optimal fiscal and monetary policy. In the RBC model considered in Section 8.1, we simply abstracted from policy altogether. In the monetary models studied in Section 8.3, nominal GDP was assumed to follow an exogenous stochastic process, bypassing the issue of what kind of monetary policies may or may not justify this assumption. Finally, in the applications of global games reviewed in Section 5, we restricted attention to specific policy instruments, such as a bailout from a lender of last resort in the context of bank runs and debt crises.

In this section, we take a different, and complementary, approach. We put aside any specific context and seek to identify a set of policy insights that may be robust across a variety of applications. In this regard, the approach taken in this section mirrors the one we have taken throughout this paper with regard to the study of the *positive* implications of frictions in information and coordination; the key difference is that we now shift attention to *normative* properties.

More specifically, we proceed as follows. First, we propose a notion of constrained efficiency, which suits our purposes, and discuss the ways in which the constrained efficient strategy may differ from either the first best or the equilibrium. Next, we compare the constrained efficient strategy to the equilibrium one within the class of beauty-contest games studied in Sections 7 and 8. This in turn paves the way to the last part of this section, where we consider a few applications. These applications regard the social value of information, the related argument in favor or against transparency in central bank communication, and the nature of optimal monetary policy in the presence of frictions in coordination.

Remark 29 Unlike the positive properties we have emphasized in this chapter, the normative properties of incomplete-information models are more sensitive to the underlying micro-foundations. This is naturally the case because the same strategic effects can be consistent with different externalities across the agents, and therefore the same positive properties can be consistent with different normative properties. A meaningful discussion of efficiency and policy therefore requires one to dig into the specifics of different applications, a task that is beyond the scope of this chapter with the exception of the few applications briefly considered in the last part of this section. What we nevertheless hope to accomplish in this section is to develop a useful way of posing the relevant normative questions.

9.1 Constrained Efficiency

The analysis follows Angeletos and Pavan (2007, 2009). The adopted efficiency concept corresponds to a benevolent planner who has the power to dictate to the agents how to act on the basis of their information, but can not collect information from one agent and send it to another. It therefore shares with Hayek (1945) and Radner (1962) the idea that information is dispersed and cannot be communicated to a "center." Similar efficiency concepts have been used to study the welfare properties of large Cournot games (Vives, 1988) and of business-cycle models with informational frictions (Angeletos and La'O, 2010; Lorenzoni, 2010).

To be concrete, consider the abstract framework introduced in Section 2 and, without serious loss of generality, restrict attention to settings in which the payoff of an agent depends on the actions of others only through the average actions. Next, for any strategy $k(\cdot): D_\omega \rightarrow D_k$, let[cz]

$$W(k(\cdot)) \equiv \mathbb{E}[U(k(\omega_i), K(\boldsymbol{\Omega}), \theta_i)]$$

denote the ex-ante utility attained by this strategy, with the understanding (i) that $K(\boldsymbol{\Omega}) \equiv \int k(\omega) d\boldsymbol{\Omega}(\omega)$ is the average action induced by the particular strategy when the distribution of information is $\boldsymbol{\Omega}$ and (ii) that the expectation is taken over the joint realizations of $s_i = (\omega_i, \theta_i)$ and $\mathbf{S}$. We can then define our efficiency benchmark as follows.[da]

Definition 16 A (constrained) efficient strategy is a strategy $k^*(\cdot)$ that maximizes $W(k(\cdot))$.

This concept lets the planner internalize how the action of each agent affects the payoff of other agents, which of course does not happen in equilibrium. Similarly to equilibrium, however, the action of any given agent is restricted to depend only on the information that Nature reveals to that particular agent, which means that the planner is precluded from transferring information from one agent to another.

If the planner were able to transfer information from one agent to another, then he would also be able to condition the action k_i of any agent i, not only on the agent's own ω_i, but also on $\boldsymbol{\Omega}$. And if that were true, the planner could instruct the agents to play the first-best actions. But as long as the planner is restricted from transferring information across agents, the first-best actions are generally not attainable. This explains the precise sense in which the notion of efficiency defined above is "constrained": the constraint is the measurability constraint that the action of each agent can not depend on the private information of other agents.

[cz] To avoid the confusion of strategies and realized actions, we henceforth denote strategies with $k(\cdot)$. That is, whereas k is an object in D_k, $k(\cdot)$ is an object in the space of functions that map D_ω to D_k.

[da] Note that Definition 16 imposes symmetry: the planner must choose the same strategy for all agents. This is without serious loss given the symmetry of the environment and the purposes of the exercises we conduct in the rest of this section, but it could matter in general. Relatedly, we envision a planner that maximizes ex-ante utility "behind the veil of ignorance," as opposed to a planner that favors any particular group of agents.

Although the planner is precluded from eliminating the heterogeneity in information, he can regulate the degree of strategic uncertainty by controlling the degree to which agents respond to their private information. For instance, the planner could eliminate strategic uncertainty altogether by instructing all agents to condition their actions only on publicly available information. Doing so, however, would mean that the agents completely disregard their private information, which is generally socially undesirable.

Studying the constrained efficient strategy therefore allows one to answer the following types of questions: What is the degree of coordination that is socially optimal given the underlying friction in communication? Should the planner use taxes, or other instruments that manipulate incentives, in an attempt to regulate the equilibrium use of information and the resulting strategic uncertainty? Do "animal spirits" justify policy intervention, and if yes of what kind?

A first step towards answering these questions is contained in the following proposition.

Proposition 35 (Efficiency) *Suppose that* $\mathrm{k}^*(\cdot)$ *is an efficient strategy, that* $k^*(\omega)$ *is in the interior of* D_k *for all* ω*, and that that* U *is differentiable. Then,*

$$\mathbb{E}[U_k(k^*(\omega),K^*(\Omega),\theta)+\Lambda^*(\Omega)|\omega]=0 \quad \forall\omega, \tag{43}$$

with

$$\Lambda^*(\Omega)\equiv\int U_K(k^*(\omega),K^*(\Omega),\theta)d\Omega(\omega) \quad \textit{and} \quad K^*(\Omega)=\int k^*(\omega)d\Omega(\omega) \quad \forall\Omega.$$

Condition (43) is simply the first-order condition of the planner's problem. The equilibrium counterpart is the best-response condition of the individual agent, which (assuming again an interior solution) is simply given by the following:

$$\mathbb{E}[U_k(k(\omega),K(\Omega),\theta)|\omega]=0 \quad \forall\omega.$$

The only essential difference between the two conditions is therefore the presence of Λ^* (Ω). This term captures the *average* externality that agents impose on one another when the cross-sectional distribution of information is $\mathbf{\Omega}$ and when all the agents follow the strategy k^*. In this regard, condition (43) is similar to the condition that characterizes the first best. What is different is that agents need not share a common belief about $\Lambda^*(\Omega)$, due to the incompleteness of the information.[db] We conclude that the efficient strategy is similar to the first best in the sense that the agents are instructed to internalize the externalities they impose on one another (if any), but departs from the first best in the sense that the agents hold different beliefs about the value of these externalities (because, and only because, they are unable to communicate with one another).

[db] Note that, as long as information is complete, $\Lambda(\Omega)$ remains commonly known even if information is *imperfect*: as with equilibrium, what matters is higher-order uncertainty, not first-order uncertainty.

A complementary interpretation of Proposition is the following. Note that we can represent the efficient strategy as the equilibrium of a fictitious game, in which the information structure remains the same as in the original game but the payoffs are appropriately modified, so that the best-response conditions of the fictitious game coincide with the first-order conditions of the planner's problem in the original game.[dc] Under this representation, condition (43) describes the socially optimal response of each agent to her beliefs of the actions of others. In this sense, the efficient strategy helps identify the degree of coordination that is socially optimal given the underlying friction in communication.

We develop below more concrete translations of these elementary insights by restricting attention to the class of beauty-contest games studied in Section 7 and to some of the related applications studied in Section 8. This focus is partly motivated by the fact that the efficiency concept defined above has not been sufficiently explored within the global-games literature.

A notable exemption to the last statement is Schaal and Taschereau-Dumouchel (2015), which characterizes, inter alia, the constrained efficient allocation within the global-games adaptation of an RBC model that we discussed in Section 5.6. See also Sakovics and Steiner (2012) and Frankel (2016), who study optimal subsidies within coordination games akin to the adoption of a network technology.

We conclude with two remarks regarding the applicability of the efficiency concept we have introduced.

First, it is straightforward to adapt this concept and the results developed in Section 9.2 to micro-founded macroeconomic models, at least insofar as one makes appropriate modeling choices. In particular, Angeletos and La'O (2010), Lorenzoni (2010), and Schaal and Taschereau-Dumouchel (2015) study business-cycle models in which the incompleteness of information introduces strategic uncertainty, without however introducing incomplete-risk sharing in consumption. This property is achieved by assuming that all relevant agents (firms, workers, etc.) belong to a "big family" whose income and consumption is fully diversified against the idiosyncratic noise in the information of different agents. All in all, one can find a direct mapping between the abstract analysis of this section and the specific results of those papers, a point that we revisit below.

Second, it is also possible to extend the notion of efficiency to accommodate certain forms of endogeneity in the information structure, whether this means endogenous *aggregation* of information or endogenous *collection* of information. For examples of the first type, we refer the reader to Amador and Weill (2012), Angeletos and Pavan (2009), Laffont (1985), Messner and Vives (2005), and Vives (2016); for example of the second type, see Angeletos et al. (2016a), Colombo et al. (2014), Llosa and Venkateswaran (2015), and Pavan (2015).

[dc] That is, the payoffs of the fictitious game are given by $\tilde{U}(k,K,\theta,\Omega) \equiv \mathbb{E}[U(k,K,\theta) + \Lambda^*(\Omega)k]$.

9.2 Efficient Coordination in Beauty Contests

We now turn to the study of constrained efficiency within the class of beauty-contests settings introduced in Section 7.1. To guarantee the existence and uniqueness of the efficient strategy, we complement Assumption 4 with the following additional restriction on the payoff structure.

Assumption 10 $U_{kk} + 2U_{kK} + U_{KK} < 0$.

We reach the following characterization of the efficient strategy.

Proposition 36 (Efficient Coordination) *There exist scalars* $(\kappa_0^*, \kappa_1^*, \kappa_2^*, \alpha^*)$*, pinned down by U, such that the following is true:*

(i) *Whenever information is complete, the efficient action is given by*

$$k_i = \kappa^*\left(\mathbb{E}_i\theta_i, \mathbb{E}_i\overline{\theta}\right) \equiv \kappa_0^* + \kappa_1^*\mathbb{E}_i\theta_i + \kappa_2^*\mathbb{E}_i\overline{\theta}. \tag{44}$$

(ii) *Whenever information is incomplete, the efficient strategy solves the following fixed-point relation:*

$$k_i = \mathbb{E}_i\left[\kappa^*\left(\theta_i, \overline{\theta}\right)\right] + \alpha^* \cdot \mathbb{E}_i\left[K - \kappa^*\left(\overline{\theta}, \overline{\theta}\right)\right]. \tag{45}$$

Parts (i) and (ii) are the normative counterparts of, respectively, Propositions 14 and 15. These propositions characterized the equilibrium under, respectively, complete and incomplete information.

By assuming that information is complete, part (i) identifies in effect the first-best action. The possibility that the scalars $(\kappa_0^*, \kappa_1^*, \kappa_2^*)$ differ from their equilibrium counterparts reflects the familiar reasons why equilibrium and first-best allocations can differ in complete-information models. But as with the equilibrium, the key observation here is that only first-order beliefs matter when information is complete.

Part (ii) then shows how the constrained efficient action differs from the first-best one when information becomes incomplete. As with equilibrium, the planner wants each agent to align her action with her forecast of the average deviation in the population. By direct implication, higher-order beliefs matter. But unlike equilibrium, the extent of the desired alignment in actions and the associated role on higher-order beliefs is parameterized by the new scalar α^*, as opposed to the scalar α we encountered before. This new scalar encapsulates the social value of coordination.

Angeletos and Pavan (2007, 2009) analyze how this value can be understood in terms of two more familiar concepts: the welfare cost of inefficiency volatility in the aggregate outcome K; and the welfare cost of inefficient dispersion in individual actions. In a business-cycle context, these objects map to the volatility of the "output gap" and the inefficient component of the cross-sectional dispersion in relative prices; see Angeletos et al. (2016b). To understand the general logic, note that if the planner induces the agents to coordinate more, then he also induces them to rely more on correlated sources of information, which in turn means more aggregate "mistakes" in actions (ie, more volatile gaps between the equilibrium and the first-best level of aggregate activity) but also fewer idiosyncratic mistakes (ie, less inefficient dispersion in the cross section). It follows that varying the level of coordination entails a trade-off between volatility and dispersion. The higher the social cost of the latter relative to the former, the higher the socially optimal degree of coordination (ie, the higher α^*).

9.3 Policy Implications

What does the result above imply for policy? If α and α^* happen to coincide, then the equilibrium and the efficient allocation exhibit the same sensitivity to higher-order beliefs. Therefore, although the equilibrium inertia and the animal spirits that we repeatedly encountered in Sections 7 and 8 are necessarily a symptom of departure from the first best, neither of them is necessarily a call for policy stabilization, at least insofar as the policy maker is unable to eliminate the incompleteness of information.

Now contrast this observation with conventional wisdom. In standard, complete-information models, animal spirits are tied to multiple, and often Pareto-ranked, equilibria. This has lead to the view that animal spirits are prima-facia rationale for policy intervention. For instance, the Keynesian tradition dictates that, if the business cycle is driven by animal spirits, then monetary and fiscal policy should be used to stabilize the economy. But if animal spirits are the product of incomplete information, and if α happens to coincide with α^*, then the aforementioned policy prescription is misguided.

Angeletos and La'O (2010) prove, in essence, that $\alpha = \alpha^*$ applies in the RBC model we studied in Section 8.1. In other words, the "aggregate demand externality" introduced by product differentiation gives rise to a private value for coordination that is perfectly aligned with its social counterpart.

Angeletos and La'O (2012) and Angeletos et al. (2016b) consider a richer framework that allows, inter alia, the type of real rigidity studied in Section 8.1 to coexist with the type of nominal rigidity studied in Section 8.3 and featured in Woodford (2003), Mankiw and Reis (2002), and Maćkowiak and Wiederholt (2009). Within that framework, it is shown that $\alpha = \alpha^*$ applies when monetary policy replicates flexible prices, as well as that such a policy is optimal when the business cycle is driven either by technology shocks or by shocks to the belief hierarchy about technology. The coincidence of the equilibrium and the efficient degrees of coordination is therefore an important benchmark for business-cycle analysis.

Angeletos and La'O (2012) further show that, similarly to the baseline New-Keynesian framework, the optimal monetary policy replicates flexible-price allocations as long as monetary policy does not have to substitute for missing tax instruments. But unlike that framework, replicating flexible-price allocations does not mean targeting price stability any more. Instead, because efficiency requires that both the quantity and the price of each firm move with its private information about the state of the economy, efficiency also requires that the aggregate price level moves with the average "sentiment" in the economy. This gives another concrete example of how accommodating realistic frictions in coordination can upset existing policy lessons.

Paciello and Wiederholt (2014) discuss additional implications in a variant setting in which firms are rationally inattentive. Monetary policy must substitute for missing tax instruments—there are markup shocks and no subsidies to correct them. For this kind of situation, the standard New-Keynesian model predicts that monetary policy should give up on price stability in order to mimic the missing counter-cyclical subsidy that would have offset the markup shocks. Paciello and Wiederholt (2014) instead show that the optimal policy may now target price stability in order to reduce the incentives of the firms to acquire information about the underlying markup shocks. Once again, existing policy lessons are upset.

A thorough analysis of the policy implications of higher-order uncertainty is beyond the scope of this chapter. Furthermore, the relevant literature is still relatively immature. For instance, all the aforementioned papers study settings in which firms are informationally constrained, but consumers are not; they also rule out the type of asset- and labor-market frictions that seem to be important in understanding business cycles. With the aforementioned examples we therefore only wish to indicate the significance of further investigations of the policy implications of incomplete information within the context of business cycles and monetary policy.

9.4 Welfare Effects of Public Information (and Central-Bank Transparency)

We conclude this section with a brief discussion of another topic, the welfare effects of the public information provided by policy makers, the media, or other sources.

As noted before, public signals can have a disproportionate effect on equilibrium outcomes in the presence of strategic uncertainty, because they become focal points and serve a role akin to coordination devices. In an influential article, Morris and Shin (2002b) relied on this observation to show that the provision of public information can contribute to more volatility and thereby also to lower welfare. They then argued that this raises questions about the social value of the information disseminated by the financial media, as well as that it justifies "constructive ambiguity" in central bank communications.

Subsequent work has questioned the applicability of Morris and Shin's welfare result to a macroeconomic context. Angeletos and Pavan (2004) document the opposite result in an investment game with production spillovers, illustrating how the welfare effects of information depend on the form of externalities, not just the form of complementarity. Angeletos and Pavan (2007) define and characterize the socially optimal degree of coordination in a flexible-class of linear-quadratic games and use this to show that Morris and Shin's welfare result hinges on assuming that coordination motives are socially unwarranted—a property that need not hold in workhorse macroeconomic models. Finally, starting with Hellwig (2005), a number of more applied works have studied the welfare effects of different types of information in micro-founded monetary models in which nominal rigidity originates from incomplete information; see, eg, Roca (2005), Walsh (2007), Baeriswyl and Cornand (2010a,b), and Lorenzoni (2010). This has lead to a variety of welfare lessons, some pro and some against central-bank transparency.

In a more recent paper, Angeletos et al. (2016b) develop a taxonomy of the welfare effects of information within a micro-founded business-cycle framework that encompasses the aforementioned applied works and disentangles the separate roles played by informational frictions and monetary policy. Importantly, the framework allows the incompleteness of information to be a source of *both* real and nominal rigidity, as in, respectively, Sections 8.1 and 8.3. Accordingly, the welfare effects of either public or private information can be decomposed into two channels: one working through the real rigidity, and another working through the nominal rigidity.

The first channel is present regardless of the conduct of monetary policy. It also contains a particularly sharp answer to the question of interest: through this channel, more information unambiguously contributes to higher welfare when the business cycle is driven by efficient forces such as technology shocks, and to lower welfare when the business cycle is driven by distortionary forces such as markup shocks.

The second channel hinges on the conduct of monetary policy. As in the baseline New-Keynesian framework, there is a policy that neutralizes the nominal rigidity; this is the same as replicating flexible-price allocations, recast in the context of models where the nominal rigidity originates from an informational friction as opposed to a Calvo-like friction. At this benchmark, the welfare effects of information are shaped solely by the real-rigidity channel. Away from it, they hinge on whether the provision of more information dampens or amplifies the deviation from the flexible-price benchmark and on whether that deviation was desirable to begin with.

When the business cycle is driven by technology shocks, the policy that replicates flexible prices is optimal. When, instead, the business cycle is driven by markup shocks, a deviation from this policy benchmark is desirable. In the latter case, more information in the hands of private agents can decrease welfare, not only because it exacerbates the inefficiency of the underlying flexible-price fluctuations, but also because it curtails the monetary authority's ability to combat these fluctuations.

These findings indicate that the welfare effects of information in baseline RBC and New-Keynesian models are closely connected to more familiar normative properties of RBC and New-Keynesian models, and have little, if anything, to do with the mechanism in Morris and Shin (2002b). That said, Morris and Shin's result may well apply to asset-markets applications insofar as speculative trading is a zero-sum game. If this is true, it would upset the conventional wisdom that the informational role of financial markets is welfare improving. It is also unclear—and therefore interesting to explore—whether the result applies to macroeconomic models that give a central role to financial markets, such as those discussed by Brunnermeier and Sannikov (2016), Gertler et al. (2016), and Guerrieri and Uhlig (2016) in this *Handbook*.

Amador and Weill (2010) identify a different reason for why the provision of public information can reduce welfare: the endogeneity of the information role of the price system. More specifically, they study the welfare effects of releasing public information about productivity and/or monetary shocks in a micro-founded model in which agents learn from the distribution of nominal prices. The release of such information induces private agents to rely less on their private sources of information. As this happens, the informational efficiency of the price system is reduced. The release of public information therefore has two competing effects: a beneficial direct one, and an adverse indirect one through the informativeness of the price system. Under certain conditions, the second effect dominates, yielding a negative overall welfare effect.

Vives (1993, 1997) and Amador and Weill (2012) develop related results within the context of a class of dynamic games with social learning. A key result is that, as long as some of the learning is private, the release of public information at some point may increase the overall level of information in the short run at the expense of lowering information in the long run. When agents are patient enough, this opens again the door to the possibility that public information is welfare deteriorating.

Whereas the above papers focus on the *aggregation* of information,[dd] Burguet and Vives (2000) focus on the *acquisition* of information. They study a model in which agents can collect private information at a cost and show that the release of public information may reduce the incentives to do so. This finding is reminiscent of a key observation made by Grossman and Stiglitz (1980), namely that the information revealed by asset prices may reduce the incentives for individual traders to collect valuable private information.

Notwithstanding the potential relevance of this last set of results, it is worth noting two facts. First, these results require that private and public information are *substitutes* for one another, which is not necessarily the case; if instead they are complements, the provision of public information can raise both the acquisition and the aggregation of private information.[de] Second, these results operate even in settings in which the payoff of each agent is independent of the actions of other agents; it is an open question whether there is sufficiently interesting interaction between these results and the mechanisms our chapter is preoccupied with.

[dd] Closely related is also the literature on herds and information cascades (Banerjee, 1992; Bikhchandani et al., 1992; Chamley, 2004). The results described above can indeed be seen as a smooth variant of a key lesson from that literature, namely the possibility that agents can *completely* cease to react to their private information after the revelation of sufficient information about the choices of other agents.

[de] This relates to the issue of whether there is strategic substitutability or complementarity in the acquisition of *private* information with the asset-market context; see Barlevy and Veronesi (2000, 2007), Veldkamp (2006), and the discussion in chapter 4 in Vives (2010).

Last but not least, Baeriswyl and Cornand (2010b) identify a potential trade off between the stabilization and the signaling roles of monetary policy. Suppose that the business cycle is driven by inefficient forces such as markup shocks. Suppose further that the monetary authority has some private information about these shocks and/or the level of economic activity.[df] To be more concrete, think of the monetary authority knowing something about either the sources or the severity of an inefficient recession. In such a situation, it is desirable for the monetary authority to withhold its information from the public, because disclosing it would only exacerbate the inefficient fluctuations. Clearly, it is also desirable to act on the basis of such information, by lowering interest rates or, perhaps, by engaging in unconventional policy measures. But note that acting means disclosing: the public will be able to extract a signal from the observed policy action about the state of the economy. It follows that the attempt to stabilize could backfire, which explains why there is a trade off between stabilization and signaling. Baeriswyl and Cornand (2010b) proceed to characterize the optimal resolution of this trade off under the assumption that the policy maker can commit on a state-contingent policy rule prior to the arrival of any information. An interesting open question is how this trade off plays out in shaping the optimal policy in the absence of such commitment.[dg]

We conclude with an obvious disclaimer. Although many of the papers we have reviewed here touch upon the topic of central-bank communication and transparency, there is a vast literature on the topic that is simply beyond the scope of our chapter. Important early contributions include Cukierman and Meltzer (1986) and Stein (1989); see also the review in Geraats (2002).

10. CONCLUSION

Since the publication of the first volume of the *Handbook of Macroeconomics*, a growing literature has emerged on the macroeconomic effects of various forms of informational frictions. In this chapter, we have tried to present and synthesize the results of this literature as it relates to coordination, strategic uncertainty, and higher-order beliefs. In the light of this synthesis, we invite the reader to consider two complementary

[df] This does not require the monetary authority's information to be more precise than that of any private agent; it only requires that the monetary authority's information is not a priori publicly available to the private agents.

[dg] Note here a certain parallel to the policy-traps setting of Angeletos et al. (2006), which we reviewed in Section 5.8. In both papers, there is a trade off between manipulating the incentives of the private agents and disclosing information to them. Apart from the apparent difference in the setting, what distinguishes the two papers is the assumption about commitment: in Angeletos et al. (2006), policy is determined in the interim state, after the policy maker has received her information.

interpretations of the particular departure from standard macroeconomic models undertaken in this chapter:

1. Insofar as signals represent "hard information," the theoretical advances we have reviewed in this chapter help understand the positive and normative implications of different kinds of information frictions within a variety of contexts, including financial crises and business cycles.
2. Insofar as signals represent "states of mind," incomplete information is more broadly a vehicle for operationalizing the notion that coordination is imperfect and for developing a more flexible framework of how agents form expectations about endogenous economic outcomes.

The mechanisms we have reviewed apply regardless of the interpretation. The interpretation, however, matters for the mapping of the theory to the real world. For certain issues, such as those that have to do with the aggregation of information or the welfare effects of public information, the first interpretation seems most appropriate. For other issues, such as interpreting the observed variation in expectations and economic outcomes, we would favor the second interpretation.

Underlying the mechanisms we have reviewed in this chapter is the difference between the strategic and the decision-theoretic aspects of informational frictions, a theme that we repeatedly visited in this chapter. We have argued that strategic uncertainty helps formalize frictions in coordination and have explored its distinct positive implications.

Throughout this chapter, we maintained the standard solution concept of rational-expectations equilibrium. While this is a standard practice, it is not necessarily the best one. We are sympathetic to other approaches that investigate plausible deviations from this practice. In fact, we are open to re-interpreting the departure considered in this chapter as a proxy for relaxing the equilibrium concept.

All in all, the following lesson emerges. Workhorse macroeconomic models, especially those used in the study of business cycles, abstract from the type of frictions in the coordination of expectations and behavior that we have studied in this chapter. As we have shown, this abstraction plays a central role in existing structural interpretations of the data and in the associated policy debate that is based on such structural interpretations. Once realistic frictions in coordination and expectations have been taken into account, our structural interpretation of the data, our view of the real world, and our prescriptions for policy can be significantly altered.

In this chapter we presented various applications of these ideas. We did not, however, offer a detailed empirical and quantitative assessment. This is because the relevant literature is still young and often confounds the decision-theoretic and the strategic aspects of informational frictions. By elucidating this distinction and the mechanisms that operate in a variety of applications, we hope to have offered, not only a sharper understanding and a certain synthesis of the literature, but also some guidance to future empirical and quantitative work.

APPENDIX. PROOFS

Proof of Propositions 1, 2, and 3 See the main text.

Proof of Proposition 4 The continuity of G together with the compactness of D_k guarantees that there always exists a solution to $K = G(K, \mathbf{B})$.

If the economy features either strategic substitutability or weak complementarity, G_K is uniformly bounded from above by 1, which guarantees that the solution is unique for all $\mathbf{B}$. With K thus determined, the optimal action of any agent i is pinned down by $k_i = g(K, b_i)$. It follows that there exists a unique equilibrium strategy, that is, a unique mapping from the realizations of b_i to those of k_i. By the same token, there is also a unique mapping from the realizations of the cross-sectional distribution of fundamentals, $\mathbf{B}$, to those of the cross-sectional distribution of actions, $\mathbf{K}$.

If instead the economy features strong complementarity, let K^* denote the solution of fixed point problem in the definition of strong complementarity. In a small neighborhood to the left of K^*, we have that $G(K,\mathbf{B}) < K$ due to the fact that the derivative $G_K(K, \mathbf{B})$ is locally higher than one. Let k_{low} be the lower bound of D_k. At $K = k_{low}$, we necessarily have that $G(K,\mathbf{B}) \geq K$. By the continuity of G, there must then exist another fixed point to G. In either case, once K is determined as a solution to $K = G(K, \mathbf{B})$, the optimal action of any agent i is given by $k_i = g(K, b_i)$. It follows that there exists multiple equilibrium strategies, each one associated with a different mapping from $\mathbf{B}$ to $\mathbf{K}$.

Proof of Proposition 5 See the main text.

Proof of Proposition 6 See Section 4.5.

Proof of Proposition 7 We start by studying the set of *monotone* or *threshold* equilibria, that is, equilibria in which the strategy of an agent satisfies the following property: for any realization of z, there is a threshold $x^*(z)$ such that an agent attacks if and only if $x \leq x^*(z)$.

When the agents follow such monotone strategies, the aggregate size of the attack is decreasing in θ, so that there is also a threshold $\theta^*(z)$ such that the status quo is abandoned if and only if $\theta \leq \theta^*(z)$. A monotone equilibrium is identified by thus identified by threshold functions $x^*$and θ^*.

The proof then proceeds in four steps. In step 1, we characterize the equilibrium θ^* for given x^*. In step 2, we characterize the equilibrium x^* for given θ^*. In step 3, we combine the two conditions to establish existence and to study the determinacy of monotone equilibria. In step 4, we conclude the proof by noting that, whenever the monotone equilibrium is unique, this gives also the unique rationalizable strategy.

Step 1. For given realizations of θ and z, the aggregate size of the attack is given by the mass of agents who receive signals $x \leq x^*(z)$. That is, letting $K(\theta, z)$ denote the size of attack when the fundamental is θ and the signal is z, we have

$$K(\theta,z) = Prob(x \leq x^*(z)|\theta) = \Phi(\sqrt{\alpha_\epsilon}(x^*(z) - \theta)),$$

where $\alpha_\epsilon = \sigma_\epsilon^{-2}$ and $\alpha_\zeta = \sigma_\zeta^{-2}$ denote, respectively, the precision of private and public information and Φ denotes the c.d.f. of the standardized Normal distribution.

Note that $K(\theta, z)$ is decreasing in θ, so that regime change occurs if and only if $\theta \leq \theta^*(z)$, where $\theta^*(z)$ is the unique solution to

$$K(\theta^*(z), z) = \theta^*(z).$$

Rearranging the above gives following relation between the thresholds x^* and θ^*:

$$x^*(z) = \theta^*(z) + \frac{1}{\sqrt{\alpha_\epsilon}}\Phi^{-1}(\theta^*(z)). \tag{A.1}$$

Step 2. Given that regime change occurs if and only if $\theta \leq \theta^*(z)$, the payoff of an agent is

$$\mathbb{E}[U(k, K(\theta, z), \theta)|x, z] = k(b\Pr[\theta \leq \theta^*(z)|x, z] - c).$$

The posterior of the agent is a Normal distribution with mean $\delta x + (1 - \delta)z$ and precision α, namely

$$\theta \ |x, z \sim \mathcal{N}\left(\delta x + (1-\delta)z, \ \ \alpha^{-1}\right),$$

where $\delta \equiv \alpha_\epsilon/(\alpha_\epsilon + \alpha_\zeta)$ captures the relative precision of private information and $\alpha \equiv \alpha_\epsilon + \alpha_\zeta$ captures the overall precision of information. Hence, the posterior probability of regime change is

$$\Pr[\theta \leq \theta^*(z)|x, z] = 1 - \Phi\left(\sqrt{\alpha}(\delta x + (1-\delta)z - \theta^*(z))\right),$$

which is monotonic in x. It follows that the agent attacks if and only if $x \leq x^*(z)$, where $x^*(z)$ solves the indifference condition

$$b\Pr[\theta \leq \theta^*(z)|x^*(z), z] = c.$$

Substituting the expression for the posterior and the definition of δ and α, we obtain:

$$\Phi\left(\sqrt{\alpha_\epsilon + \alpha_\zeta}\left(\frac{\alpha_\epsilon}{\alpha_\epsilon + \alpha_\zeta}x^*(z) + \frac{\alpha_\zeta}{\alpha_\epsilon + \alpha_\zeta}z - \theta^*(z)\right)\right) = \frac{b-c}{b}. \tag{A.2}$$

Step 3. Combining (A.1) and (A.2), we conclude that $\theta^*(z)$ can be sustained in equilibrium if and only if it solves

$$G(\theta^*(z)) = g(z), \tag{A.3}$$

where

$$G(\theta) \equiv -\frac{\alpha_\zeta}{\sqrt{\alpha_\epsilon}}\theta + \Phi^{-1}(\theta) \quad \text{and} \quad g(z) \equiv \sqrt{1 + \frac{\alpha_\zeta}{\alpha_\epsilon}}\Phi^{-1}\left(1 - \frac{c}{b}\right) - \frac{\alpha_\zeta}{\sqrt{\alpha_\epsilon}}z.$$

With $\theta^*(z)$ given by (A.3), $x^*(z)$ is then given by (A.1).

We are now in a position to establish existence and determinacy of the equilibrium by considering the properties of the function G.

Note that $G(\theta)$ is continuous in θ, with $G(\underline{\theta}) = -\infty$ and $G(\overline{\theta}) = \infty$, which implies that there necessarily exists a solution and any solution satisfies $\theta^*(z) \in (\underline{\theta}, \overline{\theta})$. This establishes existence; we now turn to uniqueness.

Next, note that

$$\frac{\partial G(\theta)}{\partial \theta} = \frac{1}{\phi(\Phi^{-1}(\theta))} - \frac{\alpha_\zeta}{\sqrt{\alpha_\epsilon}}.$$

Since $\max_{w \in \mathbb{R}} \phi(w) = 1/\sqrt{2\pi}$, the following properties are true. If $\alpha_\zeta/\sqrt{\alpha_\epsilon} \leq \sqrt{2\pi}$, we have that G is strictly increasing in θ, which implies a unique solution to (A.3) for *all* values of z. If, instead. $\alpha_\zeta/\sqrt{\alpha_\epsilon} > \sqrt{2\pi}$, then G is nonmonotonic in θ and there is an interval $(\underline{z}, \bar{z})$ such that (A.1) admits multiple solutions $\theta^*(z)$ whenever $z \in (\underline{z}, \bar{z})$ and a unique solution otherwise.

We conclude that the monotone equilibrium is unique if and only if $\alpha_\zeta/\sqrt{\alpha_\epsilon} \leq \sqrt{2\pi}$, or equivalently $\sigma_\epsilon \leq \sqrt{2\pi}\sigma_\zeta^2$.

Step 4. To complete the proof, we need to establish that, when the monotone equilibrium is unique, there is no other (nonmonotone) equilibrium. This follows from a similar argument as in the proof of Proposition 6 in the main text.

To see this, let $h(x, z)$ denote the threshold that an agent finds it optimal to follow when all other agents follow a threshold x and the public signal is z; this is the same as the function h used in the proof of Proposition 6, modulo the presence of the public signal. Next, note that there is a one-to-one mapping between the thresholds θ^* that solve the equation $G(\theta) = g(z)$ and the fixed points of h: the monotone equilibria are defined by the fixed points of h. Finally, fix an arbitrary z and re-consider the type of sequences we constructed in the proof of Proposition 6. One can always show that the sequence "from below" converges to the lowest fixed point of h, whereas the sequence "from above" converges to the higher fixed point of h. When h admits a unique fixed point, corresponding to a unique monotone equilibrium, then the same kind of argument as that in the proof of Proposition 6 implies that the unique monotone equilibrium is also the unique rationalizable outcome. And when h admits multiple fixed points, the same argument implies that the set of rationalizable outcomes is contained within the two extreme monotone equilibria.

Proof of Propositions 8 and 9 See Morris and Shin (2003).

Proof of Proposition 10 See Angeletos and Werning (2006).

Proof of Proposition 11 Similar to the proof of Proposition 5.

Proof of Proposition 12 See the argument in the main text and theorem 1 in Frankel and Pauzner (2000).

Proof of Proposition 13 Part (i) follows from theorem 3 in Frankel and Pauzner (2000) along with the following fact: that the limit considered in Frankel–Pauzner and the limit considered in Morris–Shin coincide with the risk-dominance criterion of Harsanyi and Selten (1988).

Part (ii) follows from the fact that, in the limit as $\lambda \to 0$, the drift in K_t explodes to plus infinity whenever (θ_t, K_t) is on the left of κ^* and to minus infinity on the right of it.

Proof of Proposition 14 First, consider the case of perfect information, which means that the agent knows $(\theta_i, \overline{\theta})$ and, in equilibrium, also knows K. His best response is pinned down by the following first-order condition:

$$U_k(k_i, K, \theta_i) = 0.$$

Using the fact that U is quadratic, we have:

$$U_k(0,0,0) + U_{kk}k_i + U_{kK}K + U_{k\theta}\theta_i = 0. \tag{A.4}$$

Aggregating gives

$$U_k(0,0,0) + (U_{kk} + U_{kK})K + U_{k\theta}\overline{\theta} = 0. \tag{A.5}$$

which proves that the equilibrium value of K is a linear function of $\overline{\theta}$:

$$K = -\frac{U_k(0,0,0)}{U_{kk} + U_{kK}} - \frac{U_{k\theta}}{U_{kk} + U_{kK}}\overline{\theta}.$$

Substituting this back into condition (A.4) gives the equilibrium value of k_i as a linear function of θ_i and $\overline{\theta}$:

$$k_i = \kappa\left(\theta_i, \overline{\theta}\right) \equiv \kappa_0 + \kappa_1\theta_i + \kappa_2\overline{\theta},$$

and by implication $K = \kappa(\overline{\theta}, \overline{\theta})$, where

$$\kappa_0 \equiv -\frac{U_k(0,0,0)}{U_{kk} + U_{kK}}, \quad \kappa_1 \equiv -\frac{U_{k\theta}}{U_{kk}}, \quad \kappa_2 \equiv \frac{U_{k\theta}U_{kK}}{U_{kk}(U_{kk} + U_{kK})}. \tag{A.6}$$

Next, consider the case that information is complete but not necessarily perfect. Relative to the previous case, θ_i and $\overline{\theta}$ are not necessarily known any more, yet K remains known in equilibrium. The preceding proof therefore continues to work if we simply replace θ_i with $\mathbb{E}_i[\theta_i]$ and we accordingly replace $\overline{\theta}$ with the cross-sectional average of these first-order forecasts. That is, we can now express the equilibrium action as

$$k_i = \kappa(\mathbb{E}_i[\theta_i], \vartheta^1),$$

where $\vartheta^1 \equiv \int \mathbb{E}_j[\theta_j]dj$. Because complete information means that every agent knows $\mathbf{\Omega}$, and because $\mathbf{\Omega}$ is a sufficient statistic for $\overline{\theta}$ with respect to $(\omega_i, \mathbf{\Omega})$, the following is true:

$$\mathbb{E}_i\left[\overline{\theta}\right] = \mathbb{E}\left[\overline{\theta}|\Omega\right] = \mathbb{E}\left[\theta_j|\Omega\right] = \mathbb{E}\left[\mathbb{E}_j\left[\theta_j|\omega_j, \Omega\right]|\Omega\right] = \mathbb{E}\left[\mathbb{E}_j\left[\theta_j|\omega_j\right]|\Omega\right] = \int \mathbb{E}_j\left[\theta_j\right]dj = \vartheta^1 \forall i. \tag{A.7}$$

It follows that we can express the equilibrium action as:

$$k_i = \kappa(\mathbb{E}_i[\theta_i], \mathbb{E}_i[\overline{\theta}]) = \mathbb{E}_i[\kappa(\theta_i, \overline{\theta})],$$

which proves the proposition.

Proof of Proposition 15 Let us start with a preliminary observation. From the definition of $\kappa(\cdot)$, the following is trivially true for all θ_i and $\overline{\theta}$:

$$U_k(0,0,0)+U_{kk}\kappa\left(\theta_i,\overline{\theta}\right)+U_{kK}\kappa\left(\overline{\theta},\overline{\theta}\right)+U_{k\theta}\theta_i=0.$$

Taking the expectation of both sides, we have the following is also true, no matter the information of the agent:

$$U_k(0,0,0)+U_{kk}\mathbb{E}_i\left[\kappa\left(\theta_i,\overline{\theta}\right)\right]+U_{kK}\mathbb{E}_i\left[\kappa\left(\overline{\theta},\overline{\theta}\right)\right]+U_{k\theta}\mathbb{E}_i[\theta_i]=0. \tag{A.8}$$

Turning now the agent's optimal choice of k_i under incomplete information, note that this is characterized by the following first-order condition:

$$\mathbb{E}_i[U_k(k_i,K,\theta_i)]=0.$$

Using the fact that U is quadratic, the above equation can be rewritten as follows:

$$U_k(0,0,0)+U_{kk}k_i+U_{kK}\mathbb{E}_i[K]+U_{k\theta}\mathbb{E}_i[\theta_i]=0 \tag{A.9}$$

Subtracting equation (A.8) from (A.9), we get:

$$U_{kk}\mathbb{E}_i\left[k_i-\kappa\left(\theta_i,\overline{\theta}\right)\right]+U_{kK}\mathbb{E}_i\left[K-\kappa\left(\overline{\theta},\overline{\theta}\right)\right]=0.$$

We thus arrive at equation (15) in Proposition (15), namely,

$$k_i=\mathbb{E}_i\left[\kappa\left(\theta_i,\overline{\theta}\right)\right]+\alpha\cdot\mathbb{E}_i\left[K-\kappa\left(\overline{\theta},\overline{\theta}\right)\right],$$

with

$$\alpha\equiv-\frac{U_{kK}}{U_{kk}}=\frac{\kappa_2}{\kappa_1+\kappa_2}. \tag{A.10}$$

Proof of Proposition 16 Substituting equation (14) and (A.10) into equation (15), we can write an agent's optimal action k_i under incomplete information as

$$k_i=\kappa_0(1-\alpha)+\kappa_1\theta_i+\alpha\mathbb{E}_i[K]+(\kappa_2-\alpha(\kappa_1+\kappa_2))\mathbb{E}_i\left[\overline{\theta}\right]=\kappa_0(1-\alpha)+\kappa_1\theta_i+\alpha\mathbb{E}_i[K]. \tag{A.11}$$

Aggregating over i, we have:

$$K=\kappa_0(1-\alpha)+\kappa_1\overline{\theta}+\alpha\overline{\mathbb{E}}[K].$$

Iterating, we have:

$$K=\kappa_0+\kappa_1\left\{\sum_{h=0}^{\infty}\alpha^h\overline{\mathbb{E}}^h\left[\overline{\theta}\right]\right\}.$$

where $\overline{\mathbb{E}}^0\left[\overline{\theta}\right]=\overline{\theta}$. Substituting the above into (A.11) gives:

$$k_i = \kappa_0 + \kappa_1\theta_i + \alpha\mathbb{E}_i\left[\kappa_1\left\{\sum_{h=0}^{\infty}\alpha^h\bar{\mathbb{E}}^h[\bar{\theta}]\right\}\right]. \tag{A.12}$$

From the definition of α in condition (A.10), we have $\alpha\kappa_1 = (1-\alpha)\kappa_2$. Substituting this into (A.12), we arrive at condition (16), namely:

$$k_i = \kappa_0 + \kappa_1\theta_i + \kappa_2\mathbb{E}_i\left\{\sum_{h=0}^{\infty}(1-\alpha)\alpha^h\bar{\mathbb{E}}^h[\bar{\theta}]\right\}.$$

Proof of Proposition 17 The posterior for θ conditional only on the public signal z is given by

$$\theta|z \sim N\left(\mu_{\theta|z}, \sigma^2_{\theta|z}\right), \tag{A.13}$$

where $\mu_{\theta|z} \equiv \eta_z z \equiv \dfrac{\sigma_\zeta^{-2}}{\sigma_\theta^{-2} + \sigma_\zeta^{-2}} z$ and $\sigma^{-2}_{\theta|z} \equiv \sigma_\theta^{-2} + \sigma_\zeta^{-2}$. Then we can derive the posterior for θ given both agent i's private signal x_i and public signal z as follows:

$$\theta|x_i, z \sim N\left(\lambda_x x_i + \lambda_\mu \mu_{\theta|z}, \sigma^2_{\theta|x,z}\right), \tag{A.14}$$

where $\lambda_x \equiv \dfrac{\left(\sigma_u^2 + \sigma_\epsilon^2\right)^{-1}}{\sigma^{-2}_{\theta|z} + \left(\sigma_u^2 + \sigma_\epsilon^2\right)^{-1}} = \dfrac{\left(\sigma_u^2 + \sigma_\epsilon^2\right)^{-1}}{\sigma^{-2}_{\theta|x,z}} > 0$, $\lambda_\mu \equiv \dfrac{\sigma^{-2}_{\theta|z}}{\sigma^{-2}_{\theta|z} + \left(\sigma_u^2 + \sigma_\epsilon^2\right)^{-1}} = \dfrac{\sigma_\theta^{-2} + \sigma_\zeta^{-2}}{\sigma^{-2}_{\theta|x,z}} > 0$ and $\sigma^{-2}_{\theta|x,z} \equiv \left(\sigma^2_{\theta|z}\right)^{-1} + \left(\sigma_u^2 + \sigma_\epsilon^2\right)^{-1}$.

We can also derive the posterior for the correlated error in private signals u given both agent i's private signal x_i and public signal z, distributing as:

$$u|x_i, z \sim N\left(\Lambda\left(x_i - \mu_{\theta|z}\right), \sigma^2_{u|x,z}\right), \tag{A.15}$$

where $\Lambda \equiv \dfrac{\left(\sigma^2_{\theta|z} + \sigma_\epsilon^2\right)^{-1}}{\sigma_u^{-2} + \left(\sigma^2_{\theta|z} + \sigma_\epsilon^2\right)^{-1}} = \dfrac{\left(\left(\sigma_\theta^{-2} + \sigma_\zeta^{-2}\right)^{-1} + \sigma_\epsilon^2\right)^{-1}}{\sigma_u^{-2} + \left(\left(\sigma_\theta^{-2} + \sigma_\zeta^{-2}\right)^{-1} + \sigma_\epsilon^2\right)^{-1}} = \dfrac{\left(\sigma_\theta^{-2} + \sigma_\zeta^{-2}\right)\left(\sigma_u^2 + \sigma_\epsilon^2\right)^{-1}}{\sigma_u^{-2}\left(\left(\sigma_u^2 + \sigma_\epsilon^2\right)^{-1} + \sigma_\theta^{-2} + \sigma_\zeta^{-2}\right)}$ > 0 and $\sigma^{-2}_{u|x,z} \equiv \sigma_u^{-2} + \left(\sigma^2_{\theta|z} + \sigma_\epsilon^2\right)^{-1}$.

We now state two useful properties about λ_x, λ_μ, and Λ, which will be used later:

$$\lambda_x + \Lambda = \frac{\left(\sigma_u^2 + \sigma_\epsilon^2\right)^{-1}}{\left(\sigma_u^2 + \sigma_\epsilon^2\right)^{-1} + \sigma_\theta^{-2} + \sigma_\zeta^{-2}}\left(\frac{\sigma_u^{-2} + \sigma_\theta^{-2} + \sigma_\zeta^{-2}}{\sigma_u^{-2}}\right) < 1, \tag{A.16}$$

$$\lambda_\mu - \Lambda = \frac{\sigma_\theta^{-2} + \sigma_\zeta^{-2}}{\sigma_\theta^{-2} + \sigma_\zeta^{-2} + \left(\sigma_u^2 + \sigma_\epsilon^2\right)^{-1}} \left(1 - \frac{\left(\sigma_u^2 + \sigma_\epsilon^2\right)^{-1}}{\sigma_u^{-2}}\right) > 0. \tag{A.17}$$

To express in closed form the best response of the agent given her information, we calculate each term in equation (16). From condition (A.14), we know

$$\mathbb{E}_i[\theta] = \lambda_x x_i + \lambda_\mu \mu_{\theta|z}.$$

Aggregating over i gives

$$\bar{\mathbb{E}}[\theta] = \lambda_x(\theta + u) + \lambda_\mu \mu_{\theta|z}.$$

From condition (A.14) and condition (A.15), we have:

$$\mathbb{E}_i[\bar{\mathbb{E}}[\theta]] = \lambda_x\left(\lambda_x x_i + \lambda_\mu \mu_{\theta|z} + \Lambda\left(x_i - \mu_{\theta|z}\right)\right) + \lambda_\mu \mu_{\theta|z}.$$

Aggregating over i, we have:

$$\bar{\mathbb{E}}[\bar{\mathbb{E}}[\theta]] = \lambda_x(\lambda_x + \Lambda)(\theta + u) + \left(\lambda_x\left(\lambda_\mu - \Lambda\right) + \lambda_\mu\right)\mu_{\theta|z}.$$

Similarly, for any $h \geq 0$, we have:

$$\begin{aligned} \mathbb{E}_i[\bar{\mathbb{E}}^h[\theta]] &= \lambda_x(\lambda_x + \Lambda)^h x_i + \left(\lambda_x\left(\lambda_\mu - \Lambda\right)\left(1 + (\lambda_x + \Lambda)^1 + \cdots + (\lambda_x + \Lambda)^{h-1}\right) + \lambda_\mu\right)\mu_{\theta|z} \\ &= \lambda_x(\lambda_x + \Lambda)^h x_i + \left(\lambda_x\left(\lambda_\mu - \Lambda\right)\frac{1 - (\lambda_x + \Lambda)^h}{1 - (\lambda_x + \Lambda)} + \lambda_\mu\right)\mu_{\theta|z}, \end{aligned} \tag{A.18}$$

and

$$\begin{aligned} \bar{\mathbb{E}}^{h+1}[\theta] &= \bar{\mathbb{E}}[\bar{\mathbb{E}}^h[\theta]] = \lambda(\mathbb{E}_i[\theta] + \mathbb{E}_i[u]) \\ &= \lambda_x(\lambda_x + \Lambda)^h(\theta + u) + \left(\lambda_x\left(\lambda_\mu - \Lambda\right)\frac{1 - (\lambda_x + \Lambda)^h}{1 - (\lambda_x + \Lambda)} + \lambda_\mu\right)\mu_{\theta|z}. \end{aligned} \tag{A.19}$$

As a result, for any $h \geq 1$, we have:

$$\omega_u^h = \lambda_x(\lambda_x + \Lambda)^{h-1},$$

$$\omega_\zeta^h = \left(\lambda_x\left(\lambda_\mu - \Lambda\right)\frac{1 - (\lambda_x + \Lambda)^{h-1}}{1 - (\lambda_x + \Lambda)} + \lambda_\mu\right)\eta_z,$$

$$\omega_\theta^h = \lambda_x(\lambda_x + \Lambda)^{h-1} + \left(\lambda_x\left(\lambda_\mu - \Lambda\right)\frac{1 - (\lambda_x + \Lambda)^{h-1}}{1 - (\lambda_x + \Lambda)} + \lambda_\mu\right)\eta_z = \omega_u^h + \omega_\zeta^h.$$

It is then straightforward to check that

$$0 < \omega_u^h < 1, \quad \omega_\zeta^h > 0, \quad \textit{and} \;\; \omega_\theta^h = \omega_u^h + \omega_\zeta^h > 0.$$

Now we use the fact that $\lambda_x + \lambda_\mu = 1$ to prove a useful property about the coefficients on u and ζ:

$$\omega_u^h + \frac{\omega_\zeta^h}{\eta_z} = 1 \tag{A.20}$$

This comes from

$$\begin{aligned}\omega_u^h + \frac{\omega_\zeta^h}{\eta_z} &= \lambda_x(\lambda_x + \Lambda)^{h-1} + \left(\lambda_x\left(\lambda_\mu - \Lambda\right)\frac{1-(\lambda_x+\Lambda)^{h-1}}{1-(\lambda_x+\Lambda)} + \lambda_\mu\right)\\ &= \frac{\lambda_x\left((\lambda_x+\Lambda)^{h-1} - (\lambda_x+\Lambda)^h + (1-(\lambda_x+\Lambda))\left(1-(\lambda_x+\Lambda)^{h-1}\right)\right)}{1-(\lambda_x+\Lambda)} + \lambda_\mu\\ &= \frac{\lambda_x(1-(\lambda_x+\Lambda))}{1-(\lambda_x+\Lambda)} + \lambda_\mu = 1.\end{aligned}$$

Together with the fact that $0 < \omega_u^h < 1$ and $0 < \eta_z < 1$, we have:

$$\omega_\theta^h = \eta_z\omega_u^h + \omega_\zeta^h + (1-\eta_z)\omega_u^h = \eta_z + (1-\eta_z)\omega_u^h < 1. \tag{A.21}$$

To prove

$$\omega_\theta^h > \omega_\theta^{h+1},$$

note that for any $h \geq 1$, we have:

$$\omega_u^h = \lambda_x(\lambda_x+\Lambda)^{h-1} > \lambda_x(\lambda_x+\Lambda)^h = \omega_u^{h+1}.$$

Together with equation (A.21) and the fact that $0 < \eta_z < 1$, we then have that $\omega_\theta^h > \omega_\theta^{h+1}, \;\; \forall h \geq 1$.

To prove

$$0 < \frac{Var\left(\bar{\mathbb{E}}^h[\theta]\middle|\theta\right)}{Var\left(\bar{\mathbb{E}}^h[\theta]\right)} < \frac{Var\left(\bar{\mathbb{E}}^{h+1}[\theta]\middle|\theta\right)}{Var\left(\bar{\mathbb{E}}^{h+1}[\theta]\right)} < 1,$$

note that

$$\frac{Var\left(\bar{\mathbb{E}}^h[\theta]\middle|\theta\right)}{Var\left(\bar{\mathbb{E}}^h[\theta]\right)} = \frac{\left(\omega_\zeta^h\right)^2\sigma_\zeta^2 + \left(\omega_u^h\right)^2\sigma_u^2}{\left(\omega_\zeta^h\right)^2\sigma_\zeta^2 + \left(\omega_u^h\right)^2\sigma_u^2 + \left(\omega_\theta^h\right)^2\sigma_\theta^2},$$

we obviously have

$$0 < \frac{Var\left(\bar{\mathbb{E}}^h[\theta]\middle|\theta\right)}{Var(\bar{\mathbb{E}}^h[\theta])} < 1.$$

We thus only need to prove

$$\frac{\left(\omega_\zeta^h\right)^2 \sigma_\zeta^2 + \left(\omega_u^h\right)^2 \sigma_u^2}{\left(\omega_\theta^h\right)^2 \sigma_\theta^2} < \frac{\left(\omega_\zeta^{h+1}\right)^2 \sigma_\zeta^2 + \left(\omega_u^{h+1}\right)^2 \sigma_u^2}{\left(\omega_\theta^{h+1}\right)^2 \sigma_\theta^2} \quad \forall h \geq 1.$$

Define $t(\omega) = \dfrac{\omega}{1-\omega} > 0, \quad \forall \omega \in (0,1)$, and use equation (A.20) and (A.21), we have:

$$\frac{\left(\omega_\zeta^h\right)^2 \sigma_\zeta^2 + \left(\omega_u^h\right)^2 \sigma_u^2}{\left(\omega_\theta^h\right)^2 \sigma_\theta^2} = \frac{t\left(\omega_u^h\right)^2 \sigma_u^2 + \eta_z^2 \sigma_\zeta^2}{\left(t\left(\omega_u^h\right) + \eta_z\right)^2 \sigma_\theta^2}.$$

Next, note that $1 > \omega_u^h > \omega_u^{h+1} > 0$ and $t(x)$ increases with $x \in (0,1)$, we have:

$$t\left(\omega_u^{h+1}\right) < t\left(\omega_u^h\right) \leq t\left(\omega_u^1\right) = \frac{\lambda_x}{1-\lambda_x} = \frac{\left(\sigma_u^2 + \sigma_\epsilon^2\right)^{-1}}{\sigma_\theta^{-2} + \sigma_\zeta^{-2}} < \frac{\sigma_u^{-2}}{\sigma_\theta^{-2} + \sigma_\zeta^{-2}}.$$

As a result, for $t \in \left(0, t\left(\omega_u^1\right)\right]$, we have:

$$t\sigma_u^2 - \eta_z \sigma_\zeta^2 = t\sigma_u^2 - \frac{1}{\sigma_\theta^{-2} + \sigma_\zeta^{-2}} < 0; \tag{A.22}$$

$$\frac{\partial \left[\frac{t^2 \sigma_u^2 + \eta_z^2 \sigma_\zeta^2}{(t+\eta_z)^2 \sigma_\theta^2}\right]}{\partial t} = \left(\frac{2t\sigma_u^2}{(t+\eta_z)^2 \sigma_\theta^2} - \frac{2\left(t^2 \sigma_u^2 + \eta_z^2 \sigma_\zeta^2\right)}{(t+\eta_z)^3 \sigma_\theta^2}\right) = 2\eta_z \left(\frac{t\sigma_u^2 - \eta_z \sigma_\zeta^2}{(t+\eta_z)^3 \sigma_\theta^2}\right) < 0. \tag{A.23}$$

Together with the fact that $t\left(\omega_u^{h+1}\right) < t\left(\omega_u^h\right)$, we have

$$\frac{t\left(\omega_u^h\right)^2 \sigma_u^2 + \eta_z^2 \sigma_\zeta^2}{\left(t\left(\omega_u^h\right) + \eta_z\right)^2 \sigma_\theta^2} < \frac{t\left(\omega_u^{h+1}\right)^2 \sigma_u^2 + \eta_z^2 \sigma_\zeta^2}{\left(t\left(\omega_u^{h+1}\right) + \eta_z\right)^2 \sigma_\theta^2}, \quad \forall h \geq 1,$$

and therefore

$$\frac{Var\left(\bar{\mathbb{E}}^h[\theta]\middle|\theta\right)}{Var(\bar{\mathbb{E}}^h[\theta])} < \frac{Var\left(\bar{\mathbb{E}}^{h+1}[\theta]\middle|\theta\right)}{Var(\bar{\mathbb{E}}^{h+1}[\theta])}, \quad \forall h \geq 1.$$

Proof of Proposition 18 From equation (17), we know that the individual's optimal action is given by:

$$k_i = \mathbb{E}_i\left\{\sum_{h=0}^{\infty}(1-\alpha)\alpha^h \bar{\mathbb{E}}^h[\bar{\theta}]\right\}.$$

Substituting $\mathbb{E}_i[\bar{\mathbb{E}}^h[\theta]]$ from equation (A.18), we obtain

$$\begin{aligned}
k_i &= (1-\alpha)\sum_{h=0}^{+\infty}\alpha^h\left(\lambda_x(\lambda_x+\Lambda)^h x_i + \left(\lambda_x\left(\lambda_\mu - \Lambda\right)\frac{1-(\lambda_x+\Lambda)^h}{1-(\lambda_x+\Lambda)} + \lambda_\mu\right)\mu_{\theta|z}\right)\\
&= \frac{\lambda_x(1-\alpha)}{1-\alpha(\lambda_x+\Lambda)}x_i + (1-\alpha)\left(\frac{\lambda_\mu}{1-\alpha} + \frac{\lambda_x\left(\lambda_\mu-\Lambda\right)}{1-(\lambda_x+\Lambda)}\left(\frac{1}{1-\alpha} - \frac{1}{1-\alpha(\lambda_x+\Lambda)}\right)\right)\mu_{\theta|z}\\
&= \frac{\lambda_x(1-\alpha)}{1-\alpha(\lambda_x+\Lambda)}x_i + \left(\lambda_\mu + \frac{\alpha\lambda_x\left(\lambda_\mu-\Lambda\right)}{1-\alpha(\lambda_x+\Lambda)}\right)\mu_{\theta|z}.
\end{aligned}$$

Aggregating over i, we get

$$K = \frac{\lambda_x(1-\alpha)}{1-\alpha(\lambda_x+\Lambda)}(\theta+u) + \left(\lambda_\mu + \frac{\alpha\lambda_x\left(\lambda_\mu-\Lambda\right)}{(1-\alpha(\lambda_x+\Lambda))}\right)\mu_{\theta|z}.$$

As a result,

$$\phi_u = \frac{\lambda_x(1-\alpha)}{1-\alpha(\lambda_x+\Lambda)}, \tag{A.24}$$

$$\phi_\zeta = \left(\lambda_\mu + \frac{\alpha\lambda_x\left(\lambda_\mu-\Lambda\right)}{(1-\alpha(\lambda_x+\Lambda))}\right)\eta_z, \tag{A.25}$$

$$\phi_\theta = \phi_u + \phi_\zeta. \tag{A.26}$$

Similarly to the proof of condition (A.20), using the fact that $\lambda_x + \lambda_\mu = 1$, we can prove:

$$\phi_u + \frac{\phi_\zeta}{\eta_z} = 1. \tag{A.27}$$

Together with the fact that $\lambda_x + \Lambda < 1$,$\lambda_\mu - \Lambda > 0$, and $0 < \eta_z < 1$, we have:

$$\begin{gathered}
0 < \phi_u, \phi_\zeta < 1\\
0 < \phi_\theta = \eta_z + (1-\eta_z)\phi_u < 1
\end{gathered} \tag{A.28}$$

We can also express the agents' average expectation of K as follows:

$$\begin{aligned}\bar{\mathbb{E}}[K] &= \frac{\lambda_x(1-\alpha)}{1-\alpha(\lambda_x+\Lambda)}(\bar{\mathbb{E}}[\theta]+\bar{\mathbb{E}}[u]) + \left(\lambda_\mu + \frac{\alpha\lambda_x(\lambda_\mu-\Lambda)}{(1-\alpha(\lambda_x+\Lambda))}\right)\mu_{\theta|z} \\ &= \frac{\lambda_x(1-\alpha)(\lambda_x+\Lambda)}{1-\alpha(\lambda_x+\Lambda)}(\theta+u) + \left(\lambda_\mu + \frac{\lambda_x(\lambda_\mu-\Lambda)}{1-\alpha(\lambda_x+\Lambda)}\right)\mu_{\theta|z}.\end{aligned}$$

As a result,

$$\psi_u = \frac{\lambda_x(1-\alpha)(\lambda_x+\Lambda)}{1-\alpha(\lambda_x+\Lambda)}, \tag{A.29}$$

$$\psi_\zeta = \left(\lambda_\mu + \frac{\lambda_x(\lambda_\mu-\Lambda)}{1-\alpha(\lambda_x+\Lambda)}\right)\eta_z, \tag{A.30}$$

$$\psi_\theta = \psi_u + \psi_\zeta. \tag{A.31}$$

Similarly to the proof of equation (A.20) and (A.27), using the fact that $\lambda_x + \lambda_\mu = 1$, we can prove the following:

$$\psi_u + \frac{\psi_\zeta}{\eta_z} = 1. \tag{A.32}$$

Together with the fact that $\lambda_x + \Lambda < 1, \lambda_\mu - \Lambda > 0$, and $\eta_z > 0$, we have:

$$\begin{gathered} 0 < \psi_u < \phi_u < 1, \\ 0 < \phi_\zeta < \psi_\zeta < 1, \\ 0 < \psi_\theta = \eta_z + (1-\eta_z)\psi_u < \phi_\theta < 1. \end{gathered}$$

To prove

$$0 < \frac{Var(K|\theta)}{Var(K)} < \frac{Var\left(\bar{\mathbb{E}}[K]\middle|\theta\right)}{Var(\bar{\mathbb{E}}[K])} < 1,$$

we proceed in the same way as the proof of Proposition 17. First, note that

$$\frac{Var(K|\theta)}{Var(K)} = \frac{\phi_u^2\sigma_u^2 + \phi_\zeta^2\sigma_\zeta^2}{\phi_u^2\sigma_u^2 + \phi_\zeta^2\sigma_\zeta^2 + \phi_\theta^2\sigma_\theta^2} \quad \text{and} \quad \frac{Var\left(\bar{\mathbb{E}}[K]\middle|\theta\right)}{Var(\bar{\mathbb{E}}[K])} = \frac{\psi_u^2\sigma_u^2 + \psi_\zeta^2\sigma_\zeta^2}{\psi_u^2\sigma_u^2 + \psi_\zeta^2\sigma_\zeta^2 + \psi_\theta^2\sigma_\theta^2}.$$

We thus have

$$0 < \frac{Var(K|\theta)}{Var(K)}, \frac{Var\left(\bar{\mathbb{E}}[K]\middle|\theta\right)}{Var(\bar{\mathbb{E}}[K])} < 1.$$

We only need to prove

$$\frac{\phi_u^2\sigma_u^2+\phi_\zeta^2\sigma_\zeta^2}{\phi_\theta^2\sigma_\theta^2}<\frac{\psi_u^2\sigma_u^2+\psi_\zeta^2\sigma_\zeta^2}{\psi_\theta^2\sigma_\theta^2}.$$

Defining again $t(x)=\dfrac{x}{1-x}>0,\ \ \forall x\in(0,1)$, we have

$$\frac{\phi_u^2\sigma_u^2+\phi_\zeta^2\sigma_\zeta^2}{\phi_\theta^2\sigma_\theta^2}=\frac{t(\phi_u)^2\sigma_u^2+\eta_z^2\sigma_\zeta^2}{(t(\phi_u)+\eta_z)^2\sigma_\theta^2},$$

$$\frac{\psi_u^2\sigma_u^2+\psi_\zeta^2\sigma_\zeta^2}{\psi_\theta^2\sigma_\theta^2}=\frac{t(\psi_u)^2\sigma_u^2+\eta_z^2\sigma_\zeta^2}{(t(\psi_u)+\eta_z)^2\sigma_\theta^2}.$$

$$\frac{\sigma_u^{-2}}{\sigma_\theta^{-2}+\sigma_\zeta^{-2}}>\frac{\left(\sigma_u^2+\sigma_\epsilon^2\right)^{-1}}{\sigma_\theta^{-2}+\sigma_\zeta^{-2}}=t(\lambda_x)>t(\phi_u)>t(\psi_u)>0,$$

where we use the fact that $\phi_u=\dfrac{\lambda_x(1-\alpha)}{1-\alpha(\lambda_x+\Lambda)}<\lambda_x$, which comes from $\lambda_x+\Lambda<1$, to prove $t(\lambda_x)>t(\phi_u)$. Similar as the proof of Proposition 17, for $t\in(0,t(\lambda_x)]$, we have equations (A.22) and (A.23) hold. As a result,

$$\frac{t(\phi_u)^2\sigma_u^2+\eta_z^2\sigma_\zeta^2}{(t(\phi_u)+\eta_z)^2\sigma_\theta^2}<\frac{t(\psi_u)^2\sigma_u^2+\eta_z^2\sigma_\zeta^2}{(t(\psi_u)+\eta_z)^2\sigma_\theta^2},$$

and therefore

$$\frac{Var(K|\theta)}{Var(K)}<\frac{Var\left(\bar{\mathbb{E}}[K]\middle|\theta\right)}{Var(\bar{\mathbb{E}}[K])}.$$

Proof of Proposition 19 Let $w\equiv\lambda_x u+(1-\lambda_x)\eta_z\zeta$ be the residual of projecting $\bar{\mathbb{E}}[\theta]$ on θ. Next, let $R\equiv-(1-\lambda_x)\eta_z\sigma_\zeta^2u+\lambda_x\sigma_u^2\zeta$. Clearly, we have both $w\perp\theta$ and $R\perp\theta$. Furthermore, we have $R\perp w$. To see this, note that $cov(R,w)=-(1-\lambda_x)\lambda_x\eta_z\sigma_\zeta^2\sigma_u^2+(1-\lambda_x)\lambda_x\eta_z\sigma_\zeta^2\sigma_u^2=0$, and both R and w are distributed Normally. It follows that θ, w, and R are an orthogonal basis of the space spanned by random variable θ, ζ and u.

Using the above, we can express u and ζ as linear transformations of w and R. Replacing them in to $K=\phi_\theta\theta+\phi_u u+\phi_\zeta\zeta$, we can then express K as a linear combination of θ, w, and R:

$$K=\phi_\theta\theta+\frac{\phi_u\lambda_x\sigma_u^2+(1-\phi_u)(1-\lambda_x)\eta_z^2\sigma_\zeta^2}{\lambda_x^2\sigma_u^2+(1-\lambda_x)^2\eta_z^2\sigma_\zeta^2}w+\frac{(\lambda_x-\phi_u)\eta_z}{\lambda_x^2\sigma_u^2+(1-\lambda_x)^2\eta_z^2\sigma_\zeta^2}R.\qquad\text{(A.33)}$$

Similarly, we can express $\bar{\mathbb{E}}[K]=\psi_\theta\theta+\psi_u u+\psi_\zeta\zeta$ as a linear combination of θ, w, and R:

$$\bar{\mathbb{E}}[K] = \phi_\theta \theta + \frac{\psi_u \lambda_x \sigma_u^2 + (1-\psi_u)(1-\lambda_x)\eta_z^2 \sigma_\zeta^2}{\lambda_x^2 \sigma_u^2 + (1-\lambda_x)^2 \eta_z^2 \sigma_\zeta^2} w + \frac{(\lambda_x - \psi_u)\eta_z}{\lambda_x^2 \sigma_u^2 + (1-\lambda_x)^2 \eta_z^2 \sigma_\zeta^2} R. \tag{A.34}$$

From the above, we then have

$$\begin{aligned} Var(K|\theta, \bar{\mathbb{E}}\theta) = Var(K|\theta, w) &= \left[\frac{(\lambda_x - \phi_u)\eta_z}{\lambda_x^2 \sigma_u^2 + (1-\lambda_x)^2 \eta_z^2 \sigma_\zeta^2}\right]^2 Var(R) \\ &= \frac{(\lambda_x - \phi_u)^2 \eta_z^2 \sigma_u^2 \sigma_\zeta^2}{\lambda_x^2 \sigma_u^2 + (1-\lambda_x)^2 \eta_z^2 \sigma_\zeta^2} \\ &= \left[\frac{\alpha(1-(\lambda_x + \Lambda))}{1-\alpha(\lambda_x + \Lambda)}\right]^2 \frac{\lambda_x^2 \eta_z^2 \sigma_u^2 \sigma_\zeta^2}{\lambda_x^2 \sigma_u^2 + (1-\lambda_x)^2 \eta_z^2 \sigma_\zeta^2} > 0, \end{aligned} \tag{A.35}$$

$$\begin{aligned} Var\left(\bar{\mathbb{E}}K\middle|\theta, \bar{\mathbb{E}}\theta\right) = Var\left(\bar{\mathbb{E}}K\middle|\theta, w\right) &= \left[\frac{(\lambda_x - \psi_u)\eta_z}{\lambda_x^2 \sigma_u^2 + (1-\lambda_x)^2 \eta_z^2 \sigma_\zeta^2}\right]^2 Var(R) \\ &= \frac{(\lambda_x - \psi_u)^2 \eta_z^2 \sigma_u^2 \sigma_\zeta^2}{\lambda_x^2 \sigma_u^2 + (1-\lambda_x)^2 \eta_z^2 \sigma_\zeta^2} \\ &= \left[\frac{1-(\lambda_x + \Lambda)}{1-\alpha(\lambda_x + \Lambda)}\right]^2 \frac{\lambda_x^2 \eta_z^2 \sigma_u^2 \sigma_\zeta^2}{\lambda_x^2 \sigma_u^2 + (1-\lambda_x)^2 \eta_z^2 \sigma_\zeta^2} > 0, \end{aligned} \tag{A.36}$$

and finally

$$\begin{aligned} Cov\left(K, \bar{\mathbb{E}}K\middle|\theta, \bar{\mathbb{E}}\theta\right) = Cov\left(K, \bar{\mathbb{E}}K\middle|\theta, w\right) &= \frac{(\lambda_x - \phi_u)(\lambda_x - \psi_u)\eta_z^2}{\left[\lambda_x^2 \sigma_u^2 + (1-\lambda_x)^2 \eta_z^2 \sigma_\zeta^2\right]^2} Var(R) \\ &= \frac{(\lambda_x - \phi_u)(\lambda_x - \psi_u)\eta_z^2 \sigma_u^2 \sigma_\zeta^2}{\lambda_x^2 \sigma_u^2 + (1-\lambda_x)^2 \eta_z^2 \sigma_\zeta^2} \\ &= \alpha \left[\frac{1-(\lambda_x + \Lambda)}{1-\alpha(\lambda_x + \Lambda)}\right]^2 \frac{\lambda_x^2 \eta_z^2 \sigma_u^2 \sigma_\zeta^2}{\lambda_x^2 \sigma_u^2 + (1-\lambda_x)^2 \eta_z^2 \sigma_\zeta^2} > 0, \end{aligned} \tag{A.37}$$

where we have used the facts that

$$\lambda_x - \phi_u = \lambda_x \left[1 - \frac{1-\alpha}{1-\alpha(\lambda_x + \Lambda)}\right] = \lambda_x \left[\frac{1-\alpha(\lambda_x + \Lambda) - (1-\alpha)}{1-\alpha(\lambda_x + \Lambda)}\right] = \lambda_x \left[\frac{\alpha(1-(\lambda_x + \Lambda))}{1-\alpha(\lambda_x + \Lambda)}\right]$$

and

$$\lambda_x - \psi_u = \lambda_x \left[1 - \frac{(1-\alpha)(\lambda_x + \Lambda)}{1-\alpha(\lambda_x+\Lambda)}\right] = \frac{\lambda_x(1-(\lambda_x+\Lambda))}{1-\alpha(\lambda_x+\Lambda)}.$$

Proof of Proposition 20 Using the fact that $\lambda_x + \Lambda < 1$, shown in condition (A.16), we have:

$$\begin{aligned}\frac{\partial\phi_u}{\partial\alpha} &= \frac{-\lambda_x}{1-\alpha(\lambda_x+\Lambda)} + \frac{\lambda_x(1-\alpha)(\lambda_x+\Lambda)}{(1-\alpha(\lambda_x+\Lambda))^2}\\ &= \frac{\lambda_x}{1-\alpha(\lambda_x+\Lambda)}\left(\frac{\lambda_x+\Lambda-1}{1-\alpha(\lambda_x+\Lambda)}\right) < 0.\end{aligned}$$

From condition (A.28), we have

$$\frac{\partial\phi_\theta}{\partial\alpha} = (1-\eta_z)\frac{\partial\phi_u}{\partial\alpha} < 0.$$

Similarly,

$$\begin{aligned}\frac{\partial\psi_u}{\partial\alpha} &= (\lambda_x+\Lambda)\left(\frac{-\lambda_x}{1-\alpha(\lambda_x+\Lambda)} + \frac{\lambda_x(1-\alpha)(\lambda_x+\Lambda)}{(1-\alpha(\lambda_x+\Lambda))^2}\right)\\ &= \frac{\lambda_x(\lambda_x+\Lambda)}{1-\alpha(\lambda_x+\Lambda)}\left(\frac{\lambda_x+\Lambda-1}{1-\alpha(\lambda_x+\Lambda)}\right) < 0\end{aligned}$$

and therefore also

$$\frac{\partial\psi_\theta}{\partial\alpha} = (1-\eta_z)\frac{\partial\psi_u}{\partial\alpha} < 0.$$

Note that

$$Cov(K,\theta) = \phi_\theta\sigma_\theta^2 \quad\text{and}\quad Cov(\bar{\mathbb{E}}K,\theta) = \psi_\theta\sigma_\theta^2.$$

It follows that

$$\frac{\partial Cov(K,\theta)}{\partial\alpha} < 0 \quad\text{and}\quad \frac{\partial Cov(\bar{\mathbb{E}}K,\theta)}{\partial\alpha} < 0.$$

To prove

$$\frac{\partial}{\partial\alpha}\left(\frac{Var(K|\theta)}{Var(K)}\right) > 0 \quad\text{and}\quad \frac{\partial}{\partial\alpha}\left(\frac{Var(\bar{\mathbb{E}}K|\theta)}{Var(\bar{\mathbb{E}}K)}\right) > 0,$$

note that

$$0<\frac{Var(K|\theta)}{Var(K)}=\frac{\phi_u^2\sigma_u^2+\phi_\zeta^2\sigma_\zeta^2}{\phi_u^2\sigma_u^2+\phi_\zeta^2\sigma_\zeta^2+\phi_\theta^2\sigma_\theta^2}<1 \text{ and } 0<\frac{Var(\bar{\mathbb{E}}[K]|\theta)}{Var(\bar{\mathbb{E}}[K])}=\frac{\psi_u^2\sigma_u^2+\psi_\zeta^2\sigma_\zeta^2}{\psi_u^2\sigma_u^2+\psi_\zeta^2\sigma_\zeta^2+\psi_\theta^2\sigma_\theta^2}<1,$$

To get the desired result, we therefore only need to prove

$$\frac{\partial}{\partial\alpha}\left(\frac{\phi_u^2\sigma_u^2+\phi_\zeta^2\sigma_\zeta^2}{\phi_\theta^2\sigma_\theta^2}\right)>0 \quad\text{and}\quad \frac{\partial}{\partial\alpha}\left(\frac{\psi_u^2\sigma_u^2+\psi_\zeta^2\sigma_\zeta^2}{\psi_\theta^2\sigma_\theta^2}\right)>0.$$

Defining again $t(x)=\frac{x}{1-x}>0,\quad \forall x\in(0,1)$ we have

$$\frac{\phi_u^2\sigma_u^2+\phi_\zeta^2\sigma_\zeta^2}{\phi_\theta^2\sigma_\theta^2}=\frac{t(\phi_u)^2\sigma_u^2+\sigma_\zeta^2}{(t(\phi_u)+\eta_z)^2\sigma_\theta^2},$$

$$\frac{\psi_u^2\sigma_u^2+\psi_\zeta^2\sigma_\zeta^2}{\psi_\theta^2\sigma_\theta^2}=\frac{t(\psi_u)^2\sigma_u^2+\sigma_\zeta^2}{(t(\psi_u)+\eta_z)^2\sigma_\theta^2},$$

$$\frac{\sigma_u^{-2}}{\sigma_\theta^{-2}+\sigma_\zeta^{-2}}>\frac{(\sigma_u^2+\sigma_\epsilon^2)^{-1}}{\sigma_\theta^{-2}+\sigma_\zeta^{-2}}=t(\lambda_x)>t(\phi_u)>t(\psi_u)>0,$$

where we used the fact that $\phi_u=\frac{\lambda_x(1-\alpha)}{1-\alpha(\lambda_x+\Lambda)}<\lambda_x$, which comes from $\lambda_x+\Lambda<1$, to prove $t(\lambda_x)>t(\phi_u)$. Because $\frac{\partial\phi_u}{\partial\alpha}<0$ and $\frac{\partial\psi_u}{\partial\alpha}<0$, we also have

$$\frac{\partial}{\partial\alpha}(t(\phi_u))<0 \quad \mathit{and} \quad \frac{\partial}{\partial\alpha}(t(\psi_u))<0.$$

Similar as the proof of Proposition 17 and 18, for $t\in(0,t(\lambda_x)]$, we have equations (A.22) and (A.23) hold. As a result,

$$\frac{\partial}{\partial\alpha}\left(\frac{Var(K|\theta)}{Var(K)}\right)>0 \quad\text{and}\quad \frac{\partial}{\partial\alpha}\left(\frac{Var(\bar{\mathbb{E}}K|\theta)}{Var(\bar{\mathbb{E}}K)}\right)>0,$$

We finally prove

$$\frac{\partial Var(K|\theta,\bar{\mathbb{E}}\theta)}{\partial\alpha}>0,\quad \frac{\partial Var(K|\theta,\bar{\mathbb{E}}\theta)}{\partial\alpha}>0 \quad\text{and}\quad \frac{\partial\left(Cov\left(K,\bar{\mathbb{E}}K|\theta,\bar{\mathbb{E}}\theta\right)\right)}{\partial\alpha}>0.$$

From conditions (A.35), (A.36), and (A.37), we have

$$Var(K|\theta,\bar{\mathbb{E}}\theta) = \left[\frac{\alpha(1-(\lambda_x+\Lambda))}{1-\alpha(\lambda_x+\Lambda)}\right]^2 \frac{\lambda_x^2\eta_\gamma^2\sigma_u^2\sigma_\zeta^2}{\lambda_x^2\sigma_u^2+(1-\lambda_x)^2\eta_\gamma^2\sigma_\zeta^2},$$

$$Var\left(\mathbb{E}K\middle|\theta,\bar{\mathbb{E}}\theta\right) = \left[\frac{1-(\lambda_x+\Lambda)}{1-\alpha(\lambda_x+\Lambda)}\right]^2 \frac{\lambda_x^2\eta_\gamma^2\sigma_u^2\sigma_\zeta^2}{\lambda_x^2\sigma_u^2+(1-\lambda_x)^2\eta_\gamma^2\sigma_\zeta^2},$$

$$Cov\left(K,\bar{\mathbb{E}}K\middle|\theta,\bar{\mathbb{E}}\theta\right) = \alpha\left[\frac{1-(\lambda_x+\Lambda)}{1-\alpha(\lambda_x+\Lambda)}\right]^2 \frac{\lambda_x^2\eta_\gamma^2\sigma_u^2\sigma_\zeta^2}{\lambda_x^2\sigma_u^2+(1-\lambda_x)^2\eta_\gamma^2\sigma_\zeta^2}.$$

Note that

$$\frac{\partial}{\partial\alpha}\left\{\frac{[1-(\lambda_x+\Lambda)]}{1-\alpha(\lambda_x+\Lambda)}\right\}>0 \quad\text{and}\quad \frac{\partial}{\partial\alpha}\left\{\frac{\alpha[1-(\lambda_x+\Lambda)]}{1-\alpha(\lambda_x+\Lambda)}\right\}>0.$$

The result then follows.

Proof of Proposition 21 Note that for any random variable X, and any information set I, according to law of total variance, we have:

$$Var(\mathbb{E}[X|I]) \le Var(X).$$

As a result, we have:

$$Var(\bar{\mathbb{E}}[\theta]) = Var(\mathbb{E}[\mathbb{E}[\theta|\omega]|\Omega]) \le Var(\mathbb{E}[\theta|\omega]) \le Var(\theta).$$

Similarly, for any $h \ge 2$, we have:

$$Var\left(\bar{\mathbb{E}}^h[\theta]\right) \le Var\left(\mathbb{E}\left[\mathbb{E}\left[\bar{\mathbb{E}}^{h-1}[\theta]|\omega\right]|\Omega\right]\right) \le Var\left(\mathbb{E}\left[\bar{\mathbb{E}}^{h-1}[\theta]|\omega\right]\right) \le Var\left(\bar{\mathbb{E}}^{h-1}[\theta]\right).$$

It follows that $Var\left(\bar{\mathbb{E}}^h[\theta]\right) \le Var(\theta)\forall h\ge 1$, and therefore

$$Var(K) = Var\left(\sum_{h=1}^{\infty}(1-\alpha)\alpha^{h-1}\bar{\mathbb{E}}^h[\theta]\right) \le \left(\sum_{h=1}^{\infty}(1-\alpha)\alpha^{h-1}\sqrt{Var(\theta)}\right)^2 = Var(\theta).$$

For the last result, we used the fact that, for any random variable X, Y and scalars a, $b \ge 0$, the following inequality is true:

$$\begin{aligned} Var(aX+bY) &= a^2 Var(X) + 2abCov(X,Y) + b^2 Var(Y) \\ &\le a^2 Var(X) + 2ab\sqrt{Var(X)Var(Y)} + b^2 Var(Y) \\ &= \left(a\sqrt{Var(X)} + b\sqrt{Var(Y)}\right)^2. \end{aligned}$$

Proof of Proposition 22 See the main text.

Proof of Proposition 23 The result follows essentially from Woodford (2003); see Angeletos and La'O (2010) for the particular version stated in Proposition 23. The exact characterization of the scalars $(\gamma_K, \gamma_\theta, \gamma_v)$ cannot be found in these earlier works, but can be obtained with the method developed in Huo and Takayama (2015a,b).

Proof of Proposition 24 See proposition 2 in Myatt and Wallace (2012), and proposition 1 and corollary 1 in Pavan (2015).

Proof of Proposition 25 See proposition 4 in Myatt and Wallace (2012).

Proof of Proposition 26 Under complete information, $\mathbb{E}_{it} Y_t = Y_t$ for all i. As a result, optimal output level in island i in equation (29) can be written as:

$$y_{it} = (1-\alpha)\chi a_{i,t} + \alpha Y_t.$$

Aggregating the above equation, we have $Y_t = \chi A_t$, and thus $N_t = (\chi - 1)A_t$, which completes the proof.

Proof of Proposition 27 Aggregating equation (29), we have:

$$Y_t = (1-\alpha)\chi A_t + \alpha \bar{\mathbb{E}}_t Y_t.$$

Iterating, we have:

$$Y_t = \chi \sum_{h=0}^{\infty} (1-\alpha)\alpha^h \bar{\mathbb{E}}_t^h [A_t].$$

Proof of Proposition 28 Given the information structure introduced in Assumption 9, we have:

$$\bar{\mathbb{E}}_t^h [A_t] = \left(\frac{\sigma_\epsilon^{-2}}{\sigma_u^{-2} + \sigma_\epsilon^{-2}} \right)^h u_t + A_{t-1} \quad \forall h \in \mathbb{N}.$$

Substituting it into equation (30) gives

$$Y_t = \chi A_{t-1} + \chi \frac{(1-\alpha)\left(\sigma_u^{-2} + \sigma_\epsilon^{-2}\right)}{\sigma_u^{-2} + (1-\alpha)\sigma_\epsilon^{-2}} u_t \equiv \chi A_{t-1} + \phi u_t,$$

where

$$\phi \equiv \chi \frac{(1-\alpha)\left(\sigma_u^{-2} + \sigma_\epsilon^{-2}\right)}{\sigma_u^{-2} + (1-\alpha)\sigma_\epsilon^{-2}}$$

measures the response of aggregate output to a innovation in the current fundamental. Note that $\lim_{\alpha \to 1^-} \phi = 0$. This means that, no matter how precise the information is, we can always make the aforementioned response to be arbitrarily small by boosting enough the degree of strategic complementarity.

Proof of Proposition 29 See section 4 in Woodford (2003).

Proof of Proposition 30 For the first property, see argument in Section 8.3. For the second property, see Fig. 7.

Proof of Proposition 31 Because m_{t-1} is common knowledge in the beginning of period t, it has no real effect in that period, which means that p_{it} must move one-to-one with m_{t-1}. Because local demand varies only with the sum $v_t + \xi_{it}$ and young agents have no information about the individual components of this sum, the variation in p_{it} conditional on m_{t-1} must be spanned by the variation in that sum. Along with the assumption that p_{it} is Normal, this means that $p_{it} - m_{t-1}$ is linear in the sum $v_t + \xi_{it}$. In what follows, we thus guess and verify the existence of an equilibrium in which

$$p_{it} = m_{t-1} + \phi(v_t + \xi_{it}), \tag{A.38}$$

for some scalar ϕ (which remains to be determined).

Taking the first-order condition of the problem of a young agent, we obtain the optimal labor supply, up to a constant, as follows:

$$n_{it} = \frac{1}{\kappa}\left(\mathbb{E}_{it}[p_{it} - m_t] - \mathbb{E}_{it}\left[p_{j,t+1} - m_{t+1}\right]\right). \tag{A.39}$$

Because $\mathbb{E}_{it}[m_{t+1}] = \mathbb{E}_{it}[m_t]$, we can rewrite the above as

$$n_{it} = \frac{1}{\kappa}\left(p_{it} - \mathbb{E}_{it}[p_{j,t+1}]\right).$$

Intuitively, labor supply depends on the perceived *relative* price. Using (A.38), we get that $\mathbb{E}_{it}\left[p_{j,t+1}\right] = \mathbb{E}_{it}[m_t + \phi(v_{t+1} + \xi_{it+1})] = \mathbb{E}_{it}[m_t]$, simply because no agent in period t has any information about the next-period shocks v_{t+1} and $\xi_{i,t+1}$. We conclude that the optimal labor supply can be express as follows:

$$n_{it} = \frac{1}{\kappa}(p_{it} - \mathbb{E}_{it}[m_t]).$$

Note next that the local demand is given by

$$d_{it} = m_t + \xi_{it} = m_{t-1} + v_t + \xi_{it},$$

and that market clearing imposes

$$p_{it} + n_{it} = d_{it}.$$

Combining the above we get

$$y_{it} = n_{it} = \frac{1}{\kappa + 1}\left(\xi_{i,t} + v_t - \mathbb{E}_{it}[v_t]\right),$$

and

$$p_{it} = v_t + \xi_{it} - \frac{1}{\kappa+1}\left(\xi_{i,t} + v_t - \mathbb{E}_{it}[v_t]\right) = m_{t-1} + \frac{\kappa}{\kappa+1}\left(\xi_{i,t} + v_t\right) + \frac{1}{\kappa+1}\mathbb{E}_{it}[v_t].$$

Aggregating the last equation, and letting $\bar{\mathbb{E}}_t$ denotes the average expectation of young agents in the cross section of islands, we reach the following result for the aggregate price level:

$$p_t = m_{t-1} + \frac{\kappa}{\kappa+1}v_t + \frac{1}{\kappa+1}\bar{\mathbb{E}}[v_t] = \frac{\kappa}{\kappa+1}m_t + \frac{1}{\kappa+1}\bar{\mathbb{E}}[m_t].$$

Using the fact that $m_t = m_{t-1} + v_t$ and that m_{t-1} is commonly known, we can rewrite the above as

$$p_t = (1-\beta)m_t + \beta\bar{\mathbb{E}}[m_t],$$

where $\beta \equiv \frac{1}{\kappa+1}$. This gives part (i) in the proposition.

Real output can then be expressed as

$$y_t = m_t - p_t = \beta\left\{m_t - \bar{\mathbb{E}}[m_t]\right\},$$

which gives part (ii) of the proposition. The above condition simply means that real output is pinned down by the average forecast error of the underlying money supply (also known as the "unanticipated" component of money growth).

Part (iii) follows from solving the signal-extraction problem of the young agents. In particular, condition (A.38) implies that the observation of p_{it} contains the same information as the observation of the signal

$$x_{it} \equiv \frac{1}{\phi}(p_{it} - m_{t-1}) = \frac{1}{\phi}(v_t + \xi_{it}),$$

which in turn means that $\mathbb{E}_{it}[m_t] = m_{t-1} + \delta x_{it}$ and therefore $\bar{\mathbb{E}}_t[m_t] = m_{t-1} + \delta v_t$, where $\delta \equiv \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\xi^2} \in (0,1)$. By implication, we can write the price level as $p_t = m_{t-1} + (1 - \beta + \beta\delta)v_t$, which in turn verifies our initial guess in (A.38) with $\phi = 1 - \beta + \beta\delta \in (0, 1)$. This completes the characterization of the equilibrium; it also proves condition (37) with $\lambda \equiv 1 - \delta = \frac{\sigma_\xi^2}{\sigma_v^2 + \sigma_\xi^2}$.

Proof of Proposition 32 The average first-order belief of θ_t is a simple weighted average of beliefs of the firms that updated their information at different lags:

$$\bar{\mathbb{E}}_t[\theta_t] = \lambda\Sigma_{j=0}^{\infty}(1-\lambda)^j\mathbb{E}_{t-j}[\theta_t],$$

where $\mathbb{E}_{t-j}[\theta_t]$ is the expectation conditional on all the information that is available at $t-j$. This proves that the following condition holds for $h = 1$:

$$\bar{\mathbb{E}}_t^h[\theta_t] = \sum_{j=0}^{+\infty}\left[\left(1-(1-\lambda)^{j+1}\right)^h - \left(1-(1-\lambda)^j\right)^h\right]\mathbb{E}_{t-j}[\theta_t]. \tag{A.40}$$

We then prove that the condition holds for all $h \geq 1$ by induction.

Thus suppose that (A.40) holds for some arbitrary h. It follows that

$$\begin{aligned}\mathbb{E}_{t-j}\left[\bar{\mathbb{E}}_t^h[\theta_t]\right] &= \left(1-(1-\lambda)^{j+1}\right)^h \mathbb{E}_{t-j}[\theta_t] \\ &\quad + \Sigma_{l=j+1}^{\infty}\left[\left(1-(1-\lambda)^{l+1}\right)^h - \left(1-(1-\lambda)^l\right)^h\right]\mathbb{E}_{t-l}[\theta_t].\end{aligned}$$

Then it also follows that

$$\begin{aligned}\bar{E}_t^{h+1}[\theta_t] &= \sum_{j=0}^{+\infty}\left[\lambda(1-\lambda)^j\left(1-(1-\lambda)^{j+1}\right)^h + \Sigma_{l=0}^{j-1}\lambda(1-\lambda)^l\left[\left(1-(1-\lambda)^{j+1}\right)^h - \left(1-(1-\lambda)^j\right)^h\right]\right]\mathbb{E}_{t-j}[\theta_t] \\ &= \sum_{j=0}^{+\infty}\left[\lambda(1-\lambda)^j\left(1-(1-\lambda)^{j+1}\right)^h + \left[1-(1-\lambda)^j\right]\left[\left(1-(1-\lambda)^{j+1}\right)^h - \left(1-(1-\lambda)^j\right)^h\right]\right]\mathbb{E}_{t-j}[\theta_t] \\ &= \sum_{j=0}^{+\infty}\left[\left(1-(1-\lambda)^{j+1}\right)^{h+1} - \left(1-(1-\lambda)^j\right)^{h+1}\right]\mathbb{E}_{t-j}[\theta_t],\end{aligned}$$

which means that condition (A.40) holds for $h + 1$ and proves the claim.

Now let us focus in the case in which $\rho = 0$, that is, $\theta_t = \theta_{t-1} + \nu_t$. As a result,

$$\mathbb{E}_{t-j}[\theta_t] - \mathbb{E}_{t-j-1}[\theta_t] = \nu_{t-j}, \quad \forall j \geq 0.$$

Then condition (A.40) can be written as

$$\bar{\mathbb{E}}_t^h[\theta_t] = \sum_{j=0}^{+\infty}\left\{\left(1-(1-\lambda)^{j+1}\right)^h\left(\mathbb{E}_{t-j}[\theta_t] - \mathbb{E}_{t-j-1}[\theta_t]\right)\right\} = \sum_{j=0}^{+\infty}\left\{\left(1-(1-\lambda)^{j+1}\right)^h \nu_{t-j}\right\}.$$

Proof of Propositions 33 and 34 See, respectively, propositions 2 and 1 in Allen et al. (2006).
Proof of Proposition 35 This follows from the same argument as in step 1 in the proof of propositions 1 and 2 in Angeletos and Pavan (2009). The only nonessential differences are (i) that Angeletos and Pavan (2009) uses a different notation and (ii) that their framework allows for externalities to obtain, not only from the mean action, but also from the dispersion of actions.
Proof of Proposition 36 See propositions 1 and 2 in Angeletos and Pavan (2009).

ACKNOWLEDGMENTS

For helpful comments and suggestions, we thank the editors, John Taylor, and Harald Uhlig; our discussant, Alp Simsek; various participants at the Handbook Conference hosted by the Becker Friedman Institute in Chicago; and especially Harris Dellas.

REFERENCES

Abel, A.B., Eberly, J.C., Panageas, S., 2007. Optimal inattention to the stock market. Am. Econ. Rev. 97 (2), 244–249.

Abel, A.B., Eberly, J.C., Panageas, S., 2013. Optimal inattention to the stock market with information costs and transactions costs. Econometrica 81 (4), 1455–1481.

Abreu, D., Brunnermeier, M.K., 2003. Bubbles and crashes. Econometrica 71 (1), 173–204.

Acharya, S., 2013. Dispersed beliefs and aggregate demand management. University of Maryland, Mimeo.

Akerlof, G.A., Shiller, R.J., 2010. Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism. Princeton University Press, Princeton.

Albagli, E., Hellwig, C., Tsyvinski, A., 2014. Risk-taking, rent-seeking, and investment when financial markets are noisy. Yale, Mimeo.

Albagli, E., Hellwig, C., Tsyvinski, A., 2015. A theory of asset prices based on heterogeneous information. Yale, Mimeo.

Allen, F., Morris, S., Postlewaite, A., 1993. Finite bubbles with short sale constraints and asymmetric information. J. Econ. Theory 61 (2), 206–229.

Allen, F., Morris, S., Shin, H.S., 2006. Beauty contests and iterated expectations in asset markets. Rev. Financ. Stud. 19 (3), 719–752.

Alvarez, F., Lippi, F., 2014. Price setting with menu cost for multiproduct firms. Econometrica 82 (1), 89–135.

Alvarez, F., Lippi, F., Paciello, L., 2011. Optimal price setting with observation and menu costs. Q. J. Econ. 126 (4), 1909–1960.

Alvarez, F., Lippi, F., Paciello, L., 2015. Phillips curves with observation and menu costs. Uchicago, Mimeo.

Amador, M., Weill, P.O., 2010. Learning from prices: public communication and welfare. J. Polit. Econ. 118 (5), 866–907.

Amador, M., Weill, P.O., 2012. Learning from private and public observations of others' actions. J. Econ. Theory 147 (3), 910–940.

Andrade, P., Le Bihan, H., 2013. Inattentive professional forecasters. J. Monet. Econ. 60 (8), 967–982.

Angeletos, G.M., La'O, J., 2009. Incomplete information, higher-order beliefs and price inertia. J. Monet. Econ. 56, 19–37.

Angeletos, G.M., La'O, J., 2010. Noisy business cycles. In: NBER Macroeconomics Annual 2009, vol. 24. University of Chicago Press, Chicago, pp. 319–378.

Angeletos, G.M., La'O, J., 2012. Optimal monetary policy with informational frictions. MIT, Mimeo.

Angeletos, G.M., La'O, J., 2013. Sentiments. Econometrica 81 (2), 739–779.

Angeletos, G.M., Lian, C., 2016a. A (real) theory of aggregate demand. MIT, Mimeo.

Angeletos, G.M., Lian, C., 2016b. Dampening general equilibrium: from micro elasticities to macro effects. MIT, Mimeo.

Angeletos, G.M., Lian, C., 2016c. Forward guidance without common knowledge. MIT, Mimeo.

Angeletos, G.M., Pavan, A., 2004. Transparency of information and coordination in economies with investment complementarities. Am. Econ. Rev. 94 (2), 91–98.

Angeletos, G.M., Pavan, A., 2007. Efficient use of information and social value of information. Econometrica 75 (4), 1103–1142.

Angeletos, G.M., Pavan, A., 2009. Policy with dispersed information. J. Eur. Econ. Assoc. 7 (1), 11–60.

Angeletos, G.M., Pavan, A., 2013. Selection-free predictions in global games with endogenous information and multiple equilibria. Theor. Econ. 8 (3), 883–938.

Angeletos, G.M., Werning, I., 2006. Crises and prices: information aggregation, multiplicity, and volatility. Am. Econ. Rev. 96 (5), 1720–1736.

Angeletos, G.M., Hellwig, C., Pavan, A., 2006. Signaling in a global game: coordination and policy traps. J. Polit. Econ. 114 (3), 452–484.

Angeletos, G.M., Hellwig, C., Pavan, A., 2007. Dynamic global games of regime change: learning, multiplicity, and the timing of attacks. Econometrica 75 (3), 711–756.

Angeletos, G.M., Lorenzoni, G., Pavan, A., 2010. Beauty contests and irrational exuberance: a neoclassical approach. MIT, Mimeo.

Angeletos, G.M., Collard, F., Dellas, H., 2015. Quantifying confidence. NBER Working Paper Series.

Angeletos, G.M., Iovino, L., La'O, J., 2016a. Efficiency and policy with endogenous learning. MIT, Mimeo.

Angeletos, G.M., Iovino, L., La'O, J., 2016b. Real rigidity, nominal rigidity, and the social value of information. Am. Econ. Rev. 106 (1), 200–227.

Arellano, C., Bai, Y., Kehoe, P.J., 2012. Financial frictions and fluctuations in volatility. University of Minnesota, Mimeo.

Atkeson, A., 2000. Discussion of Morris and Shin's 'rethinking multiple equilibria in macroeconomic modelling'. NBER Macroecon. Annu. 15, 162–171.

Aumann, R.J., 1974. Subjectivity and correlation in randomized strategies. J. Math. Econ. 1 (1), 67–96.

Aumann, R.J., 1987. Correlated equilibrium as an expression of Bayesian rationality. Econometrica 55, 1–18.

Aumann, R.J., Heifetz, A., 2002. Incomplete information. In: Aumann, R., Hart, S. (Eds.), Handbook of Game Theory with Economic Applications, vol. 3. Elsevier, Amsterdam, Netherlands, pp. 1665–1686.

Aumann, R.J., Peck, J., Shell, K., 1988. Asymmetric information and sunspot equilibria: a family of simple examples. CAE, Mimeo.

Azariadis, C., 1981. Self-fulfilling prophecies. J. Econ. Theory 25 (3), 380–396.

Bacchetta, P., van Wincoop, E., 2006. Can information heterogeneity explain the exchange rate determination puzzle? Am. Econ. Rev. 96 (3), 552–576.

Baeriswyl, R., Cornand, C., 2010a. Optimal monetary policy in response to cost-push shocks: the impact of central bank communication. Int. J. Cent. Bank. 6 (2), 31–52.

Baeriswyl, R., Cornand, C., 2010b. The signaling role of policy actions. J. Monet. Econ. 57 (6), 682–695.

Banerjee, A.V., 1992. A simple model of herd behavior. Q. J. Econ. 107, 797–817.

Bannier, C.E., 2006. The role of information disclosure and uncertainty in the 1994/95 Mexican Peso crisis: empirical evidence. Rev. Int. Econ. 14 (5), 883–909.

Barlevy, G., Veronesi, P., 2000. Information acquisition in financial markets. Rev. Econ. Stud. 67 (1), 79–90.

Barlevy, G., Veronesi, P., 2007. Information acquisition in financial markets: a correction. University of Chicago, Mimeo.

Barro, R.J., 1976. Rational expectations and the role of monetary policy. J. Monet. Econ. 2 (1), 1–32.

Barro, R.J., 1977. Unanticipated money growth and unemployment in the United States. Am. Econ. Rev. 67 (2), 101–115.

Barro, R.J., 1978. Unanticipated money, output, and the price level in the United States. J. Polit. Econ. 549–580.

Barsky, R.B., Sims, E.R., 2011. News shocks and business cycles. J. Monet. Econ. 58 (3), 273–289.

Barsky, R.B., Sims, E.R., 2012. Information, animal spirits, and the meaning of innovations in consumer confidence. Am. Econ. Rev. 102, 1343–1377.

Beaudry, P., Portier, F., 2006. Stock prices, news, and economic fluctuations. Am. Econ. Rev. 96 (4), 1293–1307.

Bebchuk, L.A., Goldstein, I., 2011. Self-fulfilling credit market freezes. Rev. Financ. Stud. 24 (11), 3519–3555.

Benhabib, J., Farmer, R.E.A., 1994. Indeterminacy and increasing returns. J. Econ. Theory 63 (1), 19–41.

Benhabib, J., Farmer, R.E.A., 1999. Indeterminacy and sunspots in macroeconomics. In: Taylor, J., Woodford, M. (Eds.), Handbook of Macroeconomics, vol. 1. Elsevier, Amsterdam, Netherlands, pp. 387–448.

Benhabib, J., Liu, X., Wang, P., 2015a. Endogenous information acquisition and countercyclical uncertainty. NYU mimeo.

Benhabib, J., Wang, P., Wen, Y., 2015b. Sentiments and aggregate demand fluctuations. Econometrica 83 (2), 549–585.

Benhabib, J., Liu, X., Wang, P., 2016. Sentiments, financial markets, and macroeconomic fluctuations. J. Financ. Econ. 120 (2), 420–443.

Bergemann, D., Morris, S., 2013. Robust predictions in games with incomplete information. Econometrica 81 (4), 1251–1308.

Bergemann, D., Heumann, T., Morris, S., 2015. Information and volatility. J. Econ. Theory 158, 427–465.

Biais, B., Bossaerts, P., 1998. Asset prices and trading volume in a beauty contest. Rev. Econ. Stud. 65, 307–340.

Bikhchandani, S., Hirshleifer, D., Welch, I., 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. J. Polit. Econ. 100, 992–1026.

Blanchard, O.J., L'Huillier, J.P., Lorenzoni, G., 2013. News, noise, and fluctuations: an empirical exploration. Am. Econ. Rev. 103 (7), 3045–3070.

Bloom, N., 2009. The impact of uncertainty shocks. Econometrica 77 (3), 623–685.

Bloom, N., Floetotto, M., Jaimovich, N., Saporta Eksten, I., Terry, S., 2014. Really uncertain business cycles. Stanford, Mimeo.

Branch, W.A., 2007. Sticky information and model uncertainty in survey data on inflation expectations. J. Econ. Dyn. Control 31 (1), 245–276.

Brunnermeier, M.K., Sannikov, Y., 2016. Macro, money and finance: a continuous time approach. In: Taylor, J.B., Uhlig, H. (Eds.), Handbook of Macroeconomics, vol. 2B. Elsevier, Amsterdam, Netherlands, pp. 1497–1545.

Brunnermeier, M.K., Simsek, A., Xiong, W., 2014. A welfare criterion for models with distorted beliefs. Q. J. Econ. 129 (4), 1753–1797.

Bryant, J., 1983. A simple rational expectations Keynes-type model. Q. J. Econ. 98 (3), 525–528.

Burdzy, K., Frankel, D.M., Pauzner, A., 2001. Fast equilibrium selection by rational players living in a changing world. Econometrica 69 (1), 163–189.

Burguet, R., Vives, X., 2000. Social learning and costly information acquisition. Econ. Theory 15 (1), 185–205.

Cabrales, A., Nagel, R., Armenter, R., 2007. Equilibrium selection through incomplete information in coordination games: an experimental study. Exp. Econ. 10 (3), 221–234.

Calvo, G.A., 1983. Staggered prices in a utility-maximizing framework. J. Monet. Econ. 12 (3), 383–398.

Calvo, G.A., 1988. Servicing the public debt: the role of expectations. Am. Econ. Rev. 78, 647–661.

Carlsson, H., Van Damme, E., 1993a. Equilibrium selection in stag hunt games. In: Binmore, K., Kirman, A., Tani, P. (Eds.), Frontiers of Game Theory. MIT-Press, Cambridge, USA, pp. 237–253.

Carlsson, H., Van Damme, E., 1993b. Global games and equilibrium selection. Econometrica 61, 989–1018.

Carvalho, C., Nechio, F., 2014. Do people understand monetary policy? J. Monet. Econ. 66, 108–123.

Cass, D., Shell, K., 1983. Do sunspots matter? J. Polit. Econ. 91, 193–227.

Cavallo, A., Cruces, G., Perez-Truglia, R., 2015. Inflation expectations, learning and supermarket prices: evidence from field experiments. NBER Working Paper Series.

Cespa, G., Vives, X., 2012. Dynamic trading and asset prices: Keynes vs. Hayek. Rev. Econ. Stud. 79 (2), 539–580.

Cespa, G., Vives, X., 2015. The beauty contest and short-term trading. J. Financ. 70 (5), 2099–2154.

Chahrour, R., 2014. Public communication and information acquisition. Am. Econ. J. Macroecon. 6 (3), 73–101.

Chahrour, R., Gaballo, G., 2015. On the nature and stability of sentiments. Boston College, Mimeo.

Chamley, C., 1999. Coordinating regime switches. Q. J. Econ. 114, 869–905.

Chamley, C., 2004. Rational Herds: Economic Models of Social Learning. Cambridge University Press, Cambridge, UK.

Chassang, S., 2010. Fear of miscoordination and the robustness of cooperation in dynamic global games with exit. Econometrica 78 (3), 973–1006.

Chen, Q., Goldstein, I., Jiang, W., 2010. Payoff complementarities and financial fragility: evidence from mutual fund outflows. J. Financ. Econ. 97 (2), 239–262.

Christiano, L.J., Eichenbaum, M., Evans, C.L., 1999. Monetary policy shocks: what have we learned and to what end? In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics. Elsevier, Amsterdam, Netherlands, pp. 65–148.

Christiano, L.J., Eichenbaum, M., Vigfusson, R., 2003. What happens after a technology shock? NBER Working Paper Series.

Christiano, L.J., Eichenbaum, M., Evans, C.L., 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. J. Polit. Econ. 113 (1), 1–45.

Christiano, L., Ilut, C., Motto, R., Rostagno, M., 2008. Monetary policy and stock market boom-bust cycles. Northwestern, Mimeo.

Coibion, O., Gorodnichenko, Y., 2012. What can survey forecasts tell us about information rigidities? J. Polit. Econ. 120 (1), 116–159.

Coibion, O., Gorodnichenko, Y., 2015. Information rigidity and the expectations formation process: a simple framework and new facts. Am. Econ. Rev. 105 (8), 2644–2678.

Coibion, O., Gorodnichenko, Y., Kumar, S., 2015. How do firms form their expectations? New survey evidence. NBER Working Paper Series.

Collard, F., Dellas, H., 2010. Monetary misperceptions, output, and inflation dynamics. *Journal of Money, Credit and Banking* 42 (2–3), 483–502.

Colombo, L., Femminis, G., Pavan, A., 2014. Information acquisition and welfare. Rev. Econ. Stud. 81 (4), 1438–1483.

Cooper, R., 1999. Coordination Games: Complementarities and Macroeconomics. Cambridge University Press, Cambridge.

Cooper, R., John, A., 1988. Coordinating coordination failures in Keynesian models. Q. J. Econ. 103, 441–463.

Cornand, C., 2006. Speculative attacks and informational structure: an experimental study. Rev. Int. Econ. 14 (5), 797–817.

Cornand, C., Heinemann, F., 2009. Speculative attacks with multiple sources of public information. Scand. J. Econ. 111 (1), 73–102.

Corsetti, G., Dasgupta, A., Morris, S., Shin, H.S., 2004. Does one soros make a difference? A theory of currency crises with large and small traders. Rev. Econ. Stud. 71 (1), 87–113.

Corsetti, G., Guimaraes, B., Roubini, N., 2006. International lending of last resort and moral hazard: a model of IMF's catalytic finance. J. Monet. Econ. 53 (3), 441–471.

Costain, J.S., 2007. A herding perspective on global games and multiplicity. BE J. Theor. Econ. 7 (1), 1–55.

Cukierman, A., Meltzer, A.H., 1986. A theory of ambiguity, credibility, and inflation under discretion and asymmetric information. Econometrica 54, 1099–1128.

Daníelsson, J., Pe naranda, F., 2011. On the impact of fundamentals, liquidity, and coordination on market stability. Int. Econ. Rev. 52 (3), 621–638.

Dasgupta, A., 2007. Coordination and delay in global games. J. Econ. Theory 134 (1), 195–225.

Dasgupta, A., Steiner, J., Stewart, C., 2012. Dynamic coordination with individual learning. Games Econ. Behav. 74 (1), 83–101.

David, J.M., Hopenhayn, H.A., Venkateswaran, V., 2014. Information, misallocation and aggregate productivity. NYU, Mimeo.

Davila, E., 2012. Does size matter? Bailouts with large and small banks. Harvard, Mimeo.

Dekel, E., Siniscalchi, M., et al., 2015. Epistemic game theory. In: Peyton, Y.H., Zamir, S. (Eds.), Handbook of Game Theory with Economic Applications, vol. 4. Elsevier, Amsterdam, Netherlands, pp. 619–702.

Denti, T., 2016. Games with unrestricted information acquisition. MIT, Mimeo.

Diamond, P.A., 1982. Aggregate demand management in search equilibrium. J. Polit. Econ. 90, 881–894.

Diamond, D.W., Dybvig, P.H., 1983. Bank runs, deposit insurance, and liquidity. J. Polit. Econ. 91, 401–419.

Duffy, J., Ochs, J., 2012. Equilibrium selection in static and dynamic entry games. Games Econ. Behav. 76 (1), 97–116.

Edmond, C., 2013. Information manipulation, coordination, and regime change. Rev. Econ. Stud. 80 (4), 1422–1458.

Ennis, H.M., Keister, T., 2009. Bank runs and institutions: the perils of intervention. Am. Econ. Rev. 99 (4), 1588–1607.

Evans, G.W., Honkapohja, S., 1999. Learning dynamics. In: Taylor, J., Woodford, M. (Eds.), Handbook of Macroeconomics, vol. 1. Elsevier, Amsterdam, Netherlands, pp. 449–542.

Evans, G.W., Honkapohja, S., 2009. Learning and macroeconomics. Annu. Rev. Econ. 1, 421–451.

Farhi, E., Tirole, J., 2012. Collective moral hazard, maturity mismatch, and systemic bailouts. Am. Econ. Rev. 102 (1), 60–93.

Farmer, R.E.A., 1996. A theory of business cycles. Finn. Econ. Pap. 9 (2), 91–109.

Farmer, R.E.A., Woodford, M., 1997. Self-fulfilling prophecies and the business cycle. Macroecon. Dyn. 1, 740–769.

Frankel, D.M., 2016. Optimal insurance for users of network goods. Iowa State University, Mimeo.

Frankel, D.M., Burdzy, K., 2005. Shocks and business cycles. Adv. Theor. Econ. 5(1).

Frankel, D.M., Pauzner, A., 2000. Resolving indeterminacy in dynamic settings: the role of shocks. Q. J. Econ. 115, 285–304.

Frankel, D.M., Morris, S., Pauzner, A., 2003. Equilibrium selection in global games with strategic complementarities. J. Econ. Theory 108 (1), 1–44.

Fuster, A., Laibson, D., Mendel, B., 2010. Natural expectations and macroeconomic fluctuations. J. Econ. Perspect. 24 (4), 67–84.

Futia, C.A., 1981. Rational expectations in stationary linear models. Econometrica 49 (1), 171–192.

Gabaix, X., 2014. A sparsity-based model of bounded rationality. Q. J. Econ. 129 (4), 1661–1710.

Gaballo, G., 2015. Price dispersion, private uncertainty and endogenous nominal rigidities. Banque De France, Mimeo.

Gali, J., 1999. Technology, employment, and the business cycle: do technology shocks explain aggregate fluctuations? Am. Econ. Rev. 89 (1), 249–271.

Gali, J., Rabanal, P., 2005. Technology shocks and aggregate fluctuations: how well does the real business cycle model fit postwar US data? In: Gertler, M., Rogoff, K. (Eds.), NBER Macroeconomics Annual 2004, vol. 19. MIT Press, Cambridge, USA, pp. 225–318.

Geanakoplos, J., 2010. The leverage cycle. In: Acemoglu, D, Rogoff, K., Woodford, M. (Eds.), NBER Macroeconomics Annual 2009, vol. 24. University of Chicago Press, Chicago, USA, pp. 1–65.

Gennaioli, N., Ma, Y., Shleifer, A., 2015. Expectations and investment. In: Eichenbaum, M., Parker, J. (Eds.), NBER Macroeconomics Annual 2015, vol. 30. University of Chicago Press, Chicago, USA.

Geraats, P.M., 2002. Central bank transparency. Econ. J. 112 (483), 532–565.

Gertler, M., Kiyotaki, N., Prestipino, A., 2016. Wholesale banking and bank runs in macroeconomic modelling of financial crises. In: Taylor, J.B., Uhlig, H. (Eds.), Handbook of Macroeconomics, vol. 2B. Elsevier, Amsterdam, Netherlands, pp. 1345–1425.

Giannitsarou, C., Toxvaerd, F., 2006. Recursive global games. University of Cambridge, Mimeo.

Goldstein, I., 2005. Strategic Complementarities and the Twin Crises. The Economic Journal 115, 368–390.

Goldstein, I., Ozdenoren, E., Yuan, K., 2011. Learning and complementarities in speculative attacks. The Review of Economic Studies 78, 263–292.

Goldstein, I., Ozdenoren, E., Yuan, K., 2013. Trading frenzies and their impact on real investment. Journal of Financial Economics 109, 566–582.

Goldstein, I., Pauzner, A., 2004. Contagion of self-fulfilling financial crises due to diversification of investment portfolios. Journal of Economic Theory 119, 151–183.

Goldstein, I., Pauzner, A., 2005. Demand-deposit contracts and the probability of bank runs. J. Financ. 60 (3), 1293–1327.

Greenwood, R., Shleifer, A., 2014. Expectations of returns and expected returns. Rev. Financ. Stud. 27 (3), 714–746.

Grossman, S.J., Stiglitz, J.E., 1980. On the impossibility of informationally efficient markets. Am. Econ. Rev. 70 (3), 393–408.

Guerrieri, V., Uhlig, H., 2016. Housing and credit markets: booms and busts. In: Taylor, J.B., Uhlig, H. (Eds.), Handbook of Macroeconomics, vol. 2B. Elsevier, Amsterdam, Netherlands, pp. 1427–1496.

Guesnerie, R., 1992. An exploration of the eductive justifications of the rational-expectations hypothesis. Am. Econ. Rev. 1254–1278.

Guesnerie, R., 2008. Macroeconomic and monetary policies from the eductive viewpoint. In: Schmidt-Hebbel, K., Walsh, C. (Eds.), Central Banking, Analysis, and Economic Policies Book Series, vol. 13. Central Bank of Chile, Santiago, Chile, pp. 171–202.

Guesnerie, R., Woodford, M., 1993. Endogenous fluctuations. In: Laffont, J.J. (Ed.), Advances in Economic Theory, vol. 2. Cambridge University Press, Cambridge, United Kingdom, pp. 289–412.

Guimaraes, B., 2006. Dynamics of currency crises with asset market frictions. J. Int. Econ. 68 (1), 141–158.

Guimaraes, B., Morris, S., 2007. Risk and wealth in a model of self-fulfilling currency attacks. J. Monet. Econ. 54 (8), 2205–2230.

Guimaraes, B., Machado, C., Ribeiro, M., 2014. A model of the confidence channel of fiscal policy. J. Money Credit Bank. (forthcoming).

Hamilton, J.D., 1996. This is what happened to the oil price-macroeconomy relationship. J. Monet. Econ. 38 (2), 215–220.

Hansen, L.P., Sargent, T.J., 1991. Exact linear rational expectations models: specification and estimation. In: Hansen, L.P., Sargent, T.J. (Eds.), Rational Expectations Econometrics. Westview Press, Boulder and Oxford, pp. 45–76.

Harrison, J.M., Kreps, D.M., 1978. Speculative investor behavior in a stock market with heterogeneous expectations. Q. J. Econ. 92, 323–336.

Harsanyi, J.C., 1967-1968. Games with incomplete information played by Bayesian players, Parts I, L, and III. Manag. Sci. 14, 159–182, 320-334, 486-502.

Harsanyi, J.C., Selten, R., 1988. A General Theory of Equilibrium Selection in Games, vol. 1. The MIT Press, Cambridge, USA.

Hassan, T.A., Mertens, T.M., 2011. Market sentiment: a tragedy of the commons. Am. Econ. Rev. 101 (3), 402–405.

Hassan, T.A., Mertens, T.M., 2014a. Information aggregation in a dynamic stochastic general equilibrium model. NBER Macroecon. Annu. 29 (1), 159–207.

Hassan, T.A., Mertens, T.M., 2014b. The social cost of near-rational investment. Uchicago, Mimeo.

Hayek, F.A., 1945. The use of knowledge in society. Am. Econ. Rev. 35, 519–530.

He, Z., Manela, A., 2016. Information acquisition in rumor-based bank runs. J. Financ. 71, 1113–1158.

He, Z., Xiong, W., 2012. Dynamic debt runs. Rev. Financ. Stud. 25 (6), 1799–1843.

He, Z., Krishnamurthy, A., Milbradt, K., 2016. A model of safe asset determination. UChicago, Mimeo.

Heidhues, P., Melissas, N., 2006. Equilibria in a dynamic global game: the role of cohort effects. Econ. Theory 28 (3), 531–557.

Heinemann, F., 2000. Unique equilibrium in a model of self-fulfilling currency attacks: Comment. Am. Econ. Rev. 90 (1), 316–318.

Heinemann, F., Nagel, R., Ockenfels, P., 2004. The theory of global games on test: experimental analysis of coordination games with public and private information. Econometrica 72 (5), 1583–1599.

Heinemann, F., Nagel, R., Ockenfels, P., 2009. Measuring strategic uncertainty in coordination games. Rev. Econ. Stud. 76 (1), 181–221.

Hellwig, C., 2005. Heterogeneous information and the welfare effects of public information disclosures. UCLA, Mimeo.

Hellwig, M.F., 1980. On the aggregation of information in competitive markets. J. Econ. Theory 22 (3), 477–498.

Hellwig, C., Veldkamp, L., 2009. Knowing what others know: coordination motives in information acquisition. Rev. Econ. Stud. 76 (1), 223–251.

Hellwig, C., Venkateswaran, V., 2009. Setting the right prices for the wrong reasons. J. Monet. Econ. 56, 57–77.

Hellwig, C., Mukherji, A., Tsyvinski, A., 2006. Self-fulfilling currency crises: the role of interest rates. Am. Econ. Rev. 96 (5), 1769–1787.

Holden, S., James, G., Vigier, N.A., 2014. An equilibrium theory of credit rating. University of Oslo, Mimeo.

Howitt, P., McAfee, R.P., 1992. Animal spirits. Am. Econ. Rev. 493–507.

Huang, C., 2014. Defending Against Speculative Attacks: Reputation, Learning, and Coordination. University of California, Irvine Mimeo.

Huo, Z., Takayama, N., 2015a. Higher order beliefs, confidence, and business cycles. Yale, Mimeo.

Huo, Z., Takayama, N., 2015b. Rational expectations models with higher order beliefs. Yale, Mimeo.

Iachan, F.S., Nenov, P.T., 2015. Information quality and crises in regime-change games. J. Econ. Theory 158, 739–768.

Izmalkov, S., Yildiz, M., 2010. Investor sentiments. Am. Econ. J. Microecon. 2 (1), 21–38.

Jackson, M., Peck, J., 1991. Speculation and price fluctuations with private, extrinsic signals. J. Econ. Theory 55 (2), 274–295.

Jaimovich, N., Rebelo, S., 2009. Can news about the future drive the business cycle? Am. Econ. Rev. 99 (4), 1097–1118.

Kajii, A., Morris, S., 1997a. Commonp-belief: the general case. Games Econ. Behav. 18 (1), 73–82.

Kajii, A., Morris, S., 1997b. The robustness of equilibria to incomplete information. Econometrica 1283–1309.

Kasa, K., 2000. Forecasting the forecasts of others in the frequency domain. Rev. Econ. Dyn. 3 (4), 726–756.

Kasa, K., Walker, T.B., Whiteman, C.H., 2007. Asset prices in a time series model with perpetually disparately informed, competitive traders. University of Iowa, Mimeo.

Kiley, M.T., 2007. A quantitative comparison of sticky-price and sticky-information models of price setting. J. Money Credit Bank. 39 (s1), 101–125.

Kiyotaki, N., Moore, J., 1997. Credit cycles. J. Polit. Econ. 105 (2), 211–248.

Kováč, E., Steiner, J., 2013. Reversibility in dynamic coordination problems. Games Econ. Behav. 77 (1), 298–320.

Kumar, S., Afrouzi, H., Coibion, O., Gorodnichenko, Y., 2015. Inflation targeting does not anchor inflation expectations: evidence from firms in New Zealand. NBER Working Paper Series.

Kurlat, P., 2015. Optimal stopping in a model of speculative attacks. Review of Economic Dynamics 18, 212–226.

Kurz, M., 1994. On the structure and diversity of rational beliefs. Econ. Theory 4 (6), 877–900.

Kurz, M., 2012. A new Keynesian model with diverse beliefs. Stanford, Mimeo.

Kyle, A.S., 1985. Continuous auctions and insider trading. Econometrica 53 (6), 1315–1335.

Laffont, J.J., 1985. On the welfare analysis of rational expectations equilibria with asymmetric information. Econometrica 53 (1), 1–29.

Liu, X., 2016. Interbank Market Freezes and Creditor Runs. Review of Financial Studies. forthcoming.

Llosa, L.G., Venkateswaran, V., 2015. Efficiency with endogenous information choice. NYU, Mimeo.

Lorenzoni, G., 2009. A theory of demand shocks. Am. Econ. Rev. 99 (5), 2050–2084. http://dx.doi.org/10.1257/aer.99.5.2050.

Lorenzoni, G., 2010. Optimal monetary policy with uncertain fundamentals and dispersed information. Rev. Econ. Stud. 77 (1), 305–338.

Lucas, R.E., 1972. Expectations and the neutrality of money. J. Econ. Theory 4 (2), 103–124.

Lucas, R.E., 1973. Some international evidence on output-inflation tradeoffs. Am. Econ. Rev. 63 (3), 326–334.

Luo, Y., 2008. Consumption dynamics under information processing constraints. Rev. Econ. Dyn. 11 (2), 366–385.

Luo, Y., Nie, J., Young, E.R., 2015. Slow information diffusion and the inertial behavior of durable consumption. J. Eur. Econ. Assoc. 13 (5), 805–840.

Maćkowiak, B., Wiederholt, M., 2009. Optimal sticky prices under rational inattention. Am. Econ. Rev. 99 (3), 769–803. http://dx.doi.org/10.1257/aer.99.3.769.

Maćkowiak, B., Wiederholt, M., 2015. Business cycle dynamics under rational inattention. Rev. Econ. Stud. 82 (4), 1502–1532.

Makarov, I., Rytchkov, O., 2012. Forecasting the forecasts of others: Implications for asset pricing. J. Econ. Theory 147, 941–966.

Mankiw, N.G., Reis, R., 2002. Sticky information versus sticky prices: a proposal to replace the New Keynesian Phillips Curve. Q. J. Econ. 1295–1328.

Mankiw, N.G., Reis, R., 2007. Sticky information in general equilibrium. J. Eur. Econ. Assoc. 5, 603–613.

Mankiw, N.G., Reis, R., 2011. Imperfect information and aggregate supply. In: Friedman, B.M., Woodford, M. (Eds.), Handbook of Monetary Economics. Elsevier, Amsterdam, Netherlands, pp. 183–229.

Mankiw, N.G., Reis, R., Wolfers, J., 2004. Disagreement about inflation expectations. In: Gertler, M., Rogoff, K. (Eds.), NBER Macroeconomics Annual 2003, vol. 18. The MIT Press, Cambridge, MA, pp. 209–270.

Matejka, F., 2015a. Rationally inattentive seller: sales and discrete pricing. Rev. Econ. Stud. 83 (3), 1125–1155.

Matejka, F., 2015b. Rigid pricing and rationally inattentive consumer. J. Econ. Theory 158, 656–678.

Matejka, F., McKay, A., 2015. Rational inattention to discrete choices: a new foundation for the multinomial logit model. Am. Econ. Rev. 105 (1), 272–298.

Matejka, F., Sims, C.A., 2011. Discrete actions in information-constrained tracking problems. CERGE-EI, Mimeo.

Matejka, F., Steiner, J., Stewart, C., 2015. Rational Inattention Dynamics: Inertia and Delay in Decision-Making. CERGE-EI mimeo.

Mathevet, L., 2010. A contraction principle for finite global games. Econ. Theory 42 (3), 539–563.

Mathevet, L., Steiner, J., 2013. Tractable dynamic global games and applications. J. Econ. Theory 148 (6), 2583–2619.

Matsuyama, K., 1991. Increasing returns, industrialization, and indeterminacy of equilibrium. Q. J. Econ. 104, 617–650.

Matsuyama, K., 1995. Complementarities and cumulative processes in models of monopolistic competition. J. Econ. Lit. 33 (2), 701–729.

McGrattan, E.R., 2004. Comment on Gali and Rabanal's "technology shocks and aggregate fluctuations: how well does the RBC model fit postwar US data?" NBER Macroecon. Annu. 19, 289–308.

Meese, R.A., Rogoff, K., 1983. Empirical exchange rate models of the seventies: do they fit out of sample? J. Int. Econ. 14 (1), 3–24.

Melosi, L., 2014. Estimating models with dispersed information. Am. Econ. J. Macroecon. 6 (1), 1–31.

Mertens, J.F., Zamir, S., 1985. Formulation of bayesian analysis for games with incomplete information. Int. J. Game Theory 14 (1), 1–29.

Messner, S., Vives, X., 2005. Informational and economic efficiency in REE with asymmetric information. IESE, Mimeo.

Mian, A., Sufi, A., 2014. What explains the 2007-2009 drop in employment? Econometrica 82 (6), 2197–2223.

Mian, A., Rao, K., Sufi, A., 2013. Household balance sheets, consumption, and the economic slump. Q. J. Econ. 128 (4), 1687–1726.

Monderer, D., Samet, D., 1989. Approximating common knowledge with common beliefs. Games Econ. Behav. 1 (2), 170–190.

Monderer, D., Samet, D., 1996. Proximity of information in games with incomplete information. Math. Oper. Res. 21 (3), 707–725.

Morris, S., Shin, H.S., 1997. Approximate common knowledge and co-ordination: recent lessons from game theory. J. Logic Lang. Inf. 6 (2), 171–190.

Morris, S., Shin, H.S., 1998. Unique equilibrium in a model of self-fulfilling currency attacks. Am. Econ. Rev. 88, 587–597.

Morris, S., Shin, H.S., 2001. Rethinking multiple equilibria in macroeconomic modeling. In: Bernanke, B.S., Rogoff, K. (Eds.), NBER Macroeconomics Annual 2000, vol. 15. MIT Press, Cambridge, MA, pp. 139–182.

Morris, S., Shin, H.S., 2002a. Measuring Strategic Uncertainty. Princeton University mimeo.

Morris, S., Shin, H.S., 2002b. Social value of public information. Am. Econ. Rev. 92 (5), 1521–1534.

Morris, S., Shin, H.S., 2003. Global games: theory and applications. In: Advances in Economics and Econometrics (Proceedings of the Eighth World Congress of the Econometric Society). Cambridge University Press.

Morris, S., Shin, H.S., 2004a. Coordination risk and the price of debt. Eur. Econ. Rev. 48 (1), 133–153.

Morris, S., Shin, H.S., 2004b. Liquidity black holes. Rev. Financ. 8 (1), 1–18.

Morris, S., Shin, H.S., 2006. Catalytic finance: when does it work? J. Int. Econ. 70 (1), 161–177.

Morris, S., Yang, M., 2016. Coordination and the relative cost of distinguishing nearby states. Princeton, Mimeo.

Morris, S., Postlewaite, A., Shin, H.S., 1995. Depth of knowledge and the effect of higher order uncertainty. Econ. Theory 6 (3), 453–467.

Morris, S., Shin, H.S., Yildiz, M., 2016. Common belief foundations of global games. J. Econ. Theory 163, 826–848.

Murphy, K.M., Shleifer, A., Vishny, R.W., 1989. Industrialization and the big push. J. Polit. Econ. 97, 1003–1026.

Myatt, D.P., Wallace, C., 2012. Endogenous information acquisition in coordination games. Rev. Econ. Stud. 79 (1), 340–374.

Nagar, V., Yu, G., 2014. Accounting for crises. Am. Econ. J. Macroecon. 6 (3), 184–213.

Nakamura, E., Steinsson, J., 2014. Fiscal stimulus in a monetary union: evidence from US regions. Am. Econ. Rev. 104 (3), 753–792.

Nimark, K., 2008. Dynamic pricing and imperfect common knowledge. J. Monet. Econ. 55 (2), 365–382.

Nimark, K., 2011. Dynamic higher order expectations. Universitat Pompeu Fabra, Mimeo.

Nimark, K.P., Pitschner, S., 2015. Beliefs, coordination and media focus. Cornell, Mimeo.

Obstfeld, M., 1986. Rational and self-fulfilling balance-of-payments crises. Am. Econ. Rev. 76 (1), 72–81.

Obstfeld, M., 1996. Models of currency crises with self-fulfilling features. Eur. Econ. Rev. 40 (3), 1037–1047.

Ozdenoren, E., Yuan, K., 2008. Feedback effects and asset prices. J. Finance 63, 1939–1975.

Paciello, L., Wiederholt, M., 2014. Exogenous information, endogenous information, and optimal monetary policy. Rev. Econ. Stud. 81 (1), 356–388.

Pavan, A., 2015. Attention, coordination, and bounded recall. Northwestern, Mimeo.

Pearlman, J.G., Sargent, T.J., 2005. Knowing the forecasts of others. Rev. Econ. Dyn. 8 (2), 480–497.

Peck, J., Shell, K., 1991. Market uncertainty: correlated and sunspot equilibria in imperfectly competitive economies. Rev. Econ. Stud. 58 (5), 1011–1029.

Prati, A., Sbracia, M., 2002. Currency crises and uncertainty about fundamentals. IMF, Mimeo.

Radner, R., 1962. Team decision problems. Ann. Math. Stat. 33 (3), 857–881.

Ramey, V.A., 2016. Macroeconomic shocks and their propagation. In: Taylor, J.B., Uhlig, H. (Eds.), Handbook of Macroeconomics. vol. 2A. Elsevier, Amsterdam, Netherlands, pp. 71–162.

Reis, R., 2006. Inattentive producers. Rev. Econ. Stud. 73 (3), 793–821.

Reis, R., 2009. Optimal monetary policy rules in an estimated sticky-information model. Am. Econ. J. Macroecon. 1, 1–28.

Roca, M., 2005. Transparency and monetary policy with imperfect common knowledge. Columbia, Mimeo.

Rochet, J.C., Vives, X., 2004. Coordination failures and the lender of last resort: was Bagehot right after all? J. Eur. Econ. Assoc. 2 (6), 1116–1147.

Rondina, G., Walker, T., 2014. Dispersed information and confounding dynamics. Indiana University, Mimeo.

Rubinstein, A., 1989. The electronic mail game: strategic behavior under "almost common knowledge" Am. Econ. Rev. 79, 385–391.

Sakovics, J., Steiner, J., 2012. Who matters in coordination problems? Am. Econ. Rev. 102 (7), 3439–3461.

Sargent, T.J., 2008. Evolution and intelligent design. Am. Econ. Rev. 98 (1), 3–37.

Sarte, P.D., 2014. When is sticky information more information? J. Money Credit Bank. 46 (7), 1345–1379.

Schaal, E., Taschereau-Dumouchel, M., 2015. Coordinating business cycles. NYU, Mimeo.

Scheinkman, J.A., Xiong, W., 2003. Overconfidence and speculative bubbles. J. Polit. Econ. 111 (6), 1183–1220.

Shell, K., 1977. Monnaie et allocation intertemporelle. CNRS, Mimeo.

Shell, K., 1987. Sunspot equilibrium. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave: A Dictionary of Economics, vol, 4. Macmillan, New York, pp. 549–551.

Shimer, R., 2005. The Assignment of Workers to Jobs in an Economy with Coordination Frictions. J. Polit. Econ. 113 (5), 996–1025.

Shurchkov, O., 2013. Coordination and learning in dynamic global games: experimental evidence. Exp. Econ. 16 (3), 313–334.

Sims, C.A., 2003. Implications of rational inattention. J. Monet. Econ. 50 (3), 665–690.

Sims, C.A., 2006. Rational inattention: beyond the linear-quadratic case. Am. Econ. Rev. 96 (2), 158–163.

Sims, C.A., 2010. Rational inattention and monetary economics. Handbook of Monetary Economics , vol 3, pp. 155–181.

Simsek, A., 2013. Belief disagreements and collateral constraints. Econometrica 81 (1), 1–53.

Singleton, K.J., 1987. Asset prices in a time-series model with disparately informed, competitive traders. In: Barnett, W.A., Singleton, K.J. (Eds.), New Approaches to Monetary Economics. Cambridge University Press, Cambridge, United Kingdom, pp. 249–272.

Sockin, M., Xiong, W., 2015. Informational frictions and commodity markets. J. Financ. 70 (5), 2063–2098.

Solomon, R.H., 2003. Anatomy of a twin crisis. Bank of Canada, Mimeo.

Stein, J.C., 1989. Cheap talk and the fed: a theory of imprecise policy announcements. Am. Econ. Rev. 79, 32–42.

Steiner, J., 2008. Coordination cycles. Games Econ. Behav. 63 (1), 308–327.

Stevens, L., 2015. Coarse pricing policies. Federal Reserve Bank of Minneapolis, Mimeo.

Tarashev, N.A., 2007. Speculative attacks and the information role of the interest rate. J. Eur. Econ. Assoc. 5 (1), 1–36.

Taylor, J.B., 1979. Staggered wage setting in a macro model. Am. Econ. Rev. 69 (2), 108–113.

Taylor, J.B., 1980. Aggregate dynamics and staggered contracts. J. Polit. Econ. 1–23.

Tillmann, P., 2004. Disparate information and the probability of currency crises: empirical evidence. Econ. Lett. 84 (1), 61–68.

Townsend, R.M., 1983. Forecasting the forecasts of others. J. Polit. Econ. 91, 546–588.

Toxvaerd, F., 2008. Strategic merger waves: a theory of musical chairs. J. Econ. Theory 140 (1), 1–26.

Tutino, A., 2013. Rationally inattentive consumption choices. Rev. Econ. Dyn. 16 (3), 421–439.

Van Nieuwerburgh, S., Veldkamp, L., 2009. Information immobility and the home bias puzzle. J. Financ. 64 (3), 1187–1215.

Van Nieuwerburgh, S., Veldkamp, L., 2010. Information acquisition and under-diversification. Rev. Econ. Stud. 77 (2), 779–805.

Van Zandt, T., Vives, X., 2007. Monotone equilibria in bayesian games of strategic complementarities. J. Econ. Theory 134 (1), 339–360.

Veldkamp, L.L., 2006. Media frenzies in markets for financial information. Am. Econ. Rev. 96, 577–601.

Veldkamp, L.L., 2011. Information Choice in Macroeconomics and Finance. Princeton University Press, Princeton.

Venkateswaran, V., 2014. Heterogeneous information and labor market fluctuations. NYU, Mimeo.

Vives, X., 1988. Aggregation of information in large cournot markets. Econometrica 56 (4), 851–876.

Vives, X., 1993. How fast do rational agents learn? Rev. Econ. Stud. 60 (2), 329–347.

Vives, X., 1997. Learning from others: a welfare analysis. Games Econ. Behav. 20 (2), 177–200.

Vives, X., 2005. Complementarities and games: new developments. J. Econ. Lit. 43 (2), 437–479.

Vives, X., 2010. Information and Learning in Markets: The Impact of Market Microstructure. Princeton University Press, Princeton.

Vives, X., 2016. Endogenous Public Information and Welfare in Market Games. Review of Economic Studies. forthcoming.

Walsh, C.E., 2007. Optimal economic transparency. Int. J. Cent. Bank. 3 (1), 5–36.

Weinstein, J., Yildiz, M., 2007a. Impact of higher-order uncertainty. Games Econ. Behav. 60 (1), 200–212.

Weinstein, J., Yildiz, M., 2007b. A structure theorem for rationalizability with application to robust predictions of refinements. Econometrica 75 (2), 365–400.

Woodford, M., 1986. Stationary sunspot equilibria in a finance constrained economy. J. Econ. Theory 40 (1), 128–137.

Woodford, M., 1991. Self-fulfilling expectations and fluctuations in aggregate demand. The New Keynesian Macroeconomics. MIT Press, Cambridge, USA.

Woodford, M., 2003. Imperfect common knowledge and the effects of monetary policy. In: Aghion, P., Frydman, R., Stiglitz, J., Woodford, M. (Eds.), Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps. Princeton University Press, Princeton.

Woodford, M., 2009. Information-constrained state-dependent pricing. J. Monet. Econ. 56, 100–124.

Woodford, M., 2013. Macroeconomic analysis without the rational expectations hypothesis. Annu. Rev. Econ. 5, 303–346.

Yang, M., 2015. Coordination with flexible information acquisition. J. Econ. Theory 158, 721–738.

Zabai, A., 2014. Managing Default Risk. Bank of International Settlements, mimeo.

CHAPTER 15

New Methods for Macro-Financial Model Comparison and Policy Analysis

V. Wieland*, E. Afanasyeva*, M. Kuete*, J. Yoo*,†
*IMFS, Goethe University Frankfurt, Frankfurt, Germany
†Bank of Korea, Seoul, South Korea

Contents

Handbook of Macroeconomics, Volume 2A
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.04.004

Abstract

The global financial crisis and the ensuing criticism of macroeconomics have inspired researchers to explore new modeling approaches. There are many new models that deliver improved estimates of the transmission of macroeconomic policies and aim to better integrate the financial sector in business cycle analysis. Policy making institutions need to compare available models of policy transmission and evaluate the impact and interaction of policy instruments in order to design effective policy strategies. This chapter reviews the literature on model comparison and presents a new approach for comparative analysis. Its computational implementation enables individual researchers to conduct systematic model comparisons and policy evaluations easily and at low cost. This approach also contributes to improving reproducibility of computational research in macroeconomic modeling. Several applications serve to illustrate the usefulness of model comparison and the new tools in the area of monetary and fiscal policy. They include an analysis of the impact of parameter shifts on the effects of fiscal policy, a comparison of monetary policy transmission across model generations and a cross-country comparison of the impact of changes in central bank rates in the United States and the euro area. Furthermore, the chapter includes a large-scale comparison of the dynamics and policy implications of different macro-financial models. The models considered account for financial accelerator effects in investment financing, credit and house price booms and a role for bank capital. A final exercise illustrates how these models can be used to assess the benefits of leaning against credit growth in monetary policy.

Keywords

Model comparison, Model uncertainty, Monetary policy, Fiscal policy, Policy robustness, Macro-financial models

JEL Classification Codes

E17, E27, E32, E44, E52, E58

1. INTRODUCTION

The global financial crisis and the ensuing criticism of macroeconomics have inspired researchers to develop new modeling approaches. There are many new models that aim to better integrate the financial sector in business cycle analysis. In these models, financial disturbances can have major macroeconomic consequences and the financial sector can amplify disturbances emanating from other sectors. They have potentially important implications concerning the effects of monetary and fiscal policy and the role of macroprudential and regulatory policy instruments. Thus, it is essential to be able to compare model implications for business cycle and policy analysis, and inform policy makers about policy strategies that are robust to model uncertainty. In fact, macroeconomic model comparison has a long tradition in the fields of monetary and fiscal policy. Central banks and international organizations have made much use of academic research on macroeconomic modeling, and they have invested staff resources in practical policy applications and sometimes large-scale comparison exercises.

In this chapter, we report on new developments and new techniques for comparative analysis in macroeconomic modeling. It illustrates the usefulness of model comparison with practical applications concerning the effects of fiscal measures and the transmission of monetary policy. Furthermore, it gives an overview of recent macro-financial models and compares new propagation mechanisms that arise from different financial frictions.

The chapter is meant to provide researchers, graduate students, economists at policy institutions, as well as business cycle analysts with access to a variety of macroeconomic models that reflect different theoretical paradigms, and to new hands-on tools for comparative analysis of policy and business cycle implications. A new online macroeconomic model archive together with a computational platform for model comparison takes center stage. It builds on and extends recent work on model comparison by Taylor and Wieland (2012), Wieland et al. (2012), and Schmidt and Wieland (2013). The computational platform for model comparison, the Macroeconomic Model Data Base (MMB), enables individual researchers to conduct systematic model comparisons and policy evaluations easily and at low cost.[a] Furthermore, it is straightforward to include new models and compare their empirical and policy implications to a large number of established benchmark models from academia and policy-making institutions.

Thus, the chapter and associated tools should help users to investigate questions such as: What are key features of well-known macroeconomic models? How do they influence the implications these models have for policy making and business cycle analysis? To what extent have these implications changed over time, as new models have been developed? How can I easily replicate models other researchers have developed, so as to apply them to new policy questions, or to extend them with new theoretical insights? How can I include my model in a comparison with other models in order to show what is new about it and where it improves upon earlier work? Which policy prescriptions would perform well across a range of models and economies?

Next, we review some key contributions to the literature on model comparison and highlight a recent example concerning fiscal stimulus in the Great Recession. Section 3 briefly describes a formal approach for making sure that comparisons across different models focus on comparable objects. Section 4 deals with practical issues in conducting comparisons including reproducibility and user-friendliness. Illustrative applications and extensions of earlier comparisons are presented in section 5. The in-depth comparison of financial propagation mechanisms is discussed in section 6. Section 7 gives an example of how to evaluate policy robustness and section 8 concludes.

2. LITERATURE ON MODEL COMPARISON AND POLICY ROBUSTNESS

2.1 How Model Comparison Has Been Done So Far

Model comparison has a long tradition in the fields of monetary economics and macroeconomic modeling. Typically, comparisons were not undertaken by individual researchers or small teams. Rather, comparative studies brought together several teams of researchers multiple times to obtain results. In such a setting, each team usually just works with the model they have developed.

Interest in medium- to large-scale comparisons has invariably been supported by central banks and international organizations which have been building and using macro models for decades. Decision makers at central banks and finance ministries typically rely on their staff economists to inform them about likely macroeconomic consequences of various policy actions. Furthermore, they have a strong interest in projections conditional on different policy measures. Macroeconomic models are central to fulfilling this task. And, since policy makers are interested in many scenarios and want to know about effects on different markets and sectors of the economy, the staff of their institutions are often

[a] The model archive and software are available for download at www.macromodelbase.com.

asked to build a fairly large model of the economy, or to maintain a suite of models that are useful for addressing different questions.

In the following, we review some of the contributions to this literature, the questions that were investigated, the methodological problems that presented themselves, results obtained as well as some currently hot topics.

2.1.1 1980s to Early 1990s: Standardizing Experiments and Comparing Policy Multipliers

Between 1984 and 1993, a number of large-scale comparisons were undertaken, several of them coordinated by the Brookings Institution in Washington, DC. Results were made available in the form of books with chapters being contributed by many well-known researchers in the field. These include Bryant et al. (1988, 1989), Klein (1991), and Bryant et al. (1993).

Bryant et al. (1988) aimed at improving the empirical understanding of cross-border macroeconomic and policy interactions. The study focused on the effects of monetary and fiscal policy and compared policy multipliers. To this end, participants developed and implemented a set of standardized exercises within 12 multicountry econometric models. These included models from the International Monetary Fund (IMF), the Organization for Economic Cooperation and Development (OECD), the Federal Reserve and other policy institutions as well as models developed by leading academic macroeconomists. As emphasized by Hughes-Hallett (1989), this was a very impressive undertaking at the time. The objective was to investigate to what extent the different models produce different and potentially conflicting policy implications.

A key focus was to implement common policy experiments that are comparable across models. To this end, common baseline paths of variables and common shocks were constructed across the models. Furthermore, methods were proposed to derive standardized comparison objects, such as estimates of policy multipliers. Interestingly, the study proposed computational procedures to recover estimates of the coefficients of policy variables in "final-form" equations, to cast models to IS-LM relations, and to summarize model performances in simple analytic constructs (slopes of IS-LM curves, inflation–output-tradeoffs, partial policy multipliers). The study helped identifying many challenges for standardized model comparisons and addressed some of them. It also revealed substantial differences in dynamic policy multipliers across models. Frankel and Rockett (1988) used the results from the Brookings model comparison project to explore how important uncertainty about the *true* model is for policy.

The follow-up study by Bryant et al. (1989) investigated the effects of changes in U.S. government spending and U.S. monetary policy. The majority of models participating in the exercises featured adaptive expectations. The authors computed averages and standard deviations of domestic and cross-border effects of U.S. policies across 20 global econometric models. Predicted effects from individual models, however, varied considerably. This study raised the question whether such model averages could be used as guideline for robust policy design. Another lesson of this comparison was the need to understand and evaluate the effects of policies on more variables, eg, by decomposing the effects on output into its components rather than concentrating on this variable alone.

Another round of model comparisons was documented by Klein (1991). The study focused on dynamic multipliers of fiscal spending increases, monetary expansion, and supply shocks. Similarly to earlier studies, the authors found significant variation in the behavior of models, resulting in very different policy multipliers. They aimed to put forth a common policy experiment and to apply common methods to understand the sources of variability in the multipliers. Yet, the study also made clear the difficulties in achieving comparability. The participating model teams admitted that "*it required more than a year of repeated meetings to agree on a set of inputs to be used in all models and to be sure that each model operator made the appropriate arithmetic calculations*" (Chapter 1, page 8).

Bryant et al. (1993) continued the Brookings comparison project that began in 1984. This study aimed at evaluating alternative regimes for the conduct of national monetary policies and understanding the stabilization properties of alternative operating regimes. Policy regimes were represented by simple policy rules. Taylor (1993a) credits the comparative exercise of Bryant et al. (1993) as the testing grounds for what would later become known as the Taylor rule. In contrast with previous comparative studies that focused on policy multipliers alone, Bryant et al. (1993) was the first large-scale project to compare stabilization properties of monetary policy regimes across models. The editors concluded: "*A principal conclusion of the book is that some simplified regimes for monetary policy are markedly less promising than others for achieving the stabilization objectives customarily sought by policy makers. Most notably, for a wide variety of circumstances, neither money targeting nor exchange rate targeting performs as well as a regime that targets either nominal GNP or the sum of real GNP and the inflation rate.*"

Of course, all these comparison exercises were performed at a time when simulation techniques were much less developed than today. This made the comparisons such a challenging task. In particular, shocks and exogenous processes were very different across models, which made it difficult to disentangle the effects of different patterns of stochastic shocks from different model structure and different policy regimes. In many cases, it was not possible to use a common structure of stochastic disturbances, common transition paths, and terminal conditions. Different techniques in stochastic simulation were used to perform policy experiments, which again affected the comparability of results. Furthermore, quarterly and annual models were not fully comparable. Annual models would identify monetary policy very differently by construction. Therefore, one of the lessons from this wave of comparative studies concerned the need for more standardization in methodology. Also, they highlighted the importance of expectations formation with models assuming either adaptive or rational expectations. Importantly, these comparison exercises typically concluded with an urgent call for improving empirical model validation and estimation techniques.

2.1.2 Late 1990s and Early 2000s: New Models, Policy Rules, and Robustness

Large-scale model comparison resumed with Taylor (1999). First, there was a new generation of New Keynesian models with a microfoundation built around a representative agent framework in which a household maximizes utility over time. Yet, these models were still fairly small such as the models of Rotemberg and Woodford (1997) and McCallum and Nelson (1999). They were being compared to models from the earlier generation of New Keynesian models that also featured nominal rigidities and rational expectations but a microeconomic foundation that consisted of separate decision rules for a household's consumption or a firm's investment and production problems, rather than a consistent representative agent framework. These included Fuhrer (1997), one model from Bank of England staff economists, and four models developed by staff at the Federal Reserve Board (FRB). Also, there were some models with adaptive expectations such as Rudebusch and Svensson (1999) and Ball (1999).

In terms of modeling and numerical solution techniques, there had been much progress since the earlier studies. As pointed out in the introduction of the volume, participating models had certain common features that made it easier to compute key statistics such as the variances of inflation and output under different monetary policy rules. For example, it was possible to derive linear systems determining the endogenous variables as functions of lags of themselves, the policy rate, and exogenous shocks.

A central objective was to present econometric evidence on which type of monetary policy rule is likely to be both efficient and robust when used as a guideline for the conduct of monetary policy in the United States. The stabilization performance of selected interest rate rules was evaluated across nine models. Exploiting the improvements in modeling solution techniques, Levin et al. (1999) were able to optimize over classes of policy rules using four different models, including the large-scale FRB-US model that was heavily used to inform policy makers at the Fed. Taylor (1999) concluded that simple policy rules worked well, their performance was surprisingly close to that of fully optimal policies. Furthermore, simple rules turned out to be more robust than complex rules across a variety of models.

There was disagreement about whether the central bank should react to the exchange rate and whether policy should respond to the lagged interest rate (interest rate smoothing). Furthermore, there was disagreement whether the interest rate should respond solely to a measure of expected future inflation. Follow-up work by Levin et al. (2003) found that rules that respond to forecasts with a horizon of more than one year are less robust and more prone to generating equilibrium indeterminacy than rules that respond to current observations or near-term forecasts.

With the creation of the euro area many new models were built to inform policy makers at the European Central Bank (ECB) and other European and international institutions. A special issue of *Economic Modeling* was put together by Hughes-Hallett and Wallis (2004) to present and compare models for the euro area. It was preceded by conferences bringing together modelers from central banks, international institutions, and academia to discuss estimates from different models. The paper by Wallis (2004) presents comparative results from four models, the ECB's area-wide model, and three established multicountry models (IMF's MULTIMOD model, NIGEM from the National Institute of Economic and Social Research in London, and the QUEST model from the European Commission). He found the principal source of differences across the four models to be the different degree of forward-looking behavior incorporated in the treatment of consumption and investment decisions and the setting of wages and prices. Of course, at that point models for the euro area had to be estimated on pre-EMU macroeconomic data. Hence substantial uncertainty remained about the stability of established empirical regularities.

2.1.3 Building a Model Archive to Provide Easy Access to Model Comparison

The last 15 years have experienced a massive surge in macroeconomic model building. A new generation of medium-size New Keynesian Dynamic Stochastic General Equilibrium (DSGE) models emerged following the contribution by Christiano et al. (2005), which was first circulated as a working paper in 2001. These models extended the microfoundations of the representative agent framework with additional rigidities, adjustment costs, and behavioral economics features such as habit formation. Smets and Wouters (2003) estimated a version of such a medium-size model for the euro area and helped popularize the use of Bayesian estimation methods. Widely used solution and estimation techniques were implemented in the DYNARE software package developed by Juillard (2001) (see also Adjemian et al., 2011).

These advances in model building, model solution, model estimation, and software implementations prepared the ground for more easy access to model comparison and the analysis of policy robustness by small teams of researchers. Taylor and Wieland (2012) extended the earlier model comparisons with U.S. models to the new medium-size DSGE models. They compare three such models built and estimated for the U.S. economy with the earlier-generation multicountry model of Taylor (1993b). Somewhat surprisingly, despite all the differences in structural assumptions, estimation techniques and data sample, all four models considered produce strikingly similar output responses to monetary policy shocks. Here, we extend this analysis further in Section 5.2.

Kuester and Wieland (2010) and Orphanides and Wieland (2013) studied the robustness of simple monetary policy rules, the latter study using 11 new models estimated on euro area data. Orphanides and Wieland (2013) find that rules optimized for just one model are not robust, as they often result in substantially worse performance in other models. Yet, they show that a simple (not optimized) difference rule reacting to current inflation and output growth performs quite well across models.

Wieland et al. (2012) brought together models from these and earlier comparative studies to build an archive of macroeconomic models for easy simulation and comparison. This archive together with a new platform for performing standardized comparative exercises will be presented and used in subsequent sections of this chapter.

2.1.4 Hot Topics: Fiscal Policy, Macro-Financial Modeling, and Macroprudential Policy

Following the global financial crisis and Great Recession, there is high demand for new and improved models. Issues of great interest include the impact of fiscal stimulus and consolidation, the effects of unconventional monetary policy measures, and the interaction of the real and financial sectors of the economy. Furthermore, there are new policy instruments to evaluate in banking regulation and macroprudential policy making. As a result, many new macro-financial models are being developed.

There have been several model comparison studies regarding fiscal policy, among them Cogan et al. (2010), Cwik and Wieland (2011), Cogan et al. (2013), and two large-scale comparisons of fiscal multipliers across models and countries by Coenen et al. (2012) and Kilponen et al. (2015). Section 2.2 reviews results from this debate concerning the likely effects of fiscal policy near zero interest rates.

Of course, policy makers and modelers alike have been preoccupied with the role of the financial sector as a source of disruptions and as an amplifier of other economic disturbances for some time. There are many new modeling approaches. Thus, comparative research can generate useful insights. As a first step, Gerke et al. (2013) have considered five models of the European economies that are based on theoretical contributions from the precrisis period and are currently employed by central banks in the Eurosystem. They compare open-economy models featuring the financial accelerator mechanism as in Bernanke et al. (1999) and/or collateral constraints in the spirit of Iacoviello (2005). The focus of the study is on qualitative comparison of impulse responses of macroeconomic and financial variables to a range of common shocks (eg, monetary policy shocks, net worth shocks, loan-to-value shocks). The study concludes that models display qualitatively similar features, reflecting a common understanding of macro-financial linkages preceding the financial crisis and the Great Recession. The authors, however, emphasize the need for a new generation of macro-financial models.

Guerrieri et al. (2015) is one of the first comparative studies of models that explicitly consider risks emanating from the banking system itself. Five groups of modelers from the Federal Reserve Board participated in the study. The authors compare macroeconomic spillovers from a (standardized) shortfall in bank capital across five DSGE models. The shortfall in bank capital is modeled as a gradually decaying pure transfer from the banking sector to the household sector.

The models under consideration exhibit many differences. There are nominal and real models, models solved with linear and nonlinear techniques, models featuring complementary approaches to modeling financial intermediation. The financial shock is carefully standardized. Responses of macroeconomic and financial variables vary substantially. Noteworthy, the range of model-based outcomes is contained in the confidence bands of a bivariate vector autoregression (VAR), suggesting a similar degree of uncertainty in response to the financial shocks in the models as in the VAR. The authors identify several sources of differences in responses across models. For instance, modeling of different sources of bank funding (eg, inside vs. outside equity) and interactions between alternative sectors, which can provide credit in the economy, are found to be particularly relevant for the results.

Section 6 provides an overview of different approaches for modeling macro-financial interactions. It also presents a range of examples and new findings from model comparisons. The model archive to be presented allows individual researchers to conduct such comparisons fairly easily themselves, and to include their own model so as to identify its contributions relative to more established benchmarks.

2.2 A Recent Example: Comparing Effects of the 2009–10 Fiscal Stimulus

The ongoing debate about the benefits and drawbacks of discretionary fiscal stimulus provides an excellent example of the need for model comparisons. The Great Recession of 2008 and 2009 has triggered substantial interest in assessing the likely impact of fiscal measures. In response to the financial market meltdown and the sharp contraction of real GDP, central banks in advanced economies have first slashed interest rates for central bank liquidity and then resorted to quantitative easing in order to further expand their balance sheets as policy rates remained near zero percent. At the same time, governments have launched large-scale fiscal stimulus packages.

In the United States, for example, the American Recovery and Reinvestment Act (ARRA) of February 2009 comprised US$ 787 billion of additional government purchases, transfers, and tax reductions. The lion's share was planned to be spent over a period of five years reaching a peak in 2010. The European Union initiated the European Economic Recovery Plan (EERP) and euro area member states launched fiscal stimulus packages on the scale of €175 billion to be spent in the years 2009 and 2010. Clearly, when deciding on such large programs, policy makers should be informed of the likely quantitative impact.

2.2.1 Determinants of Keynesian Multiplier Effects

Advocates of fiscal stimulus refer to the Keynesian multiplier effect and emphasize that it would increase in strength with constant interest rates. The multiplier effect arises in the textbook IS-LM model due to the static nature of the Keynesian consumption function, which assumes a positive relationship between consumption and current household income. Additional government spending results in more aggregate demand, more production and more income, which in turn feeds additional household consumption and hence yet another increase in income and so on. This multiplication suggests that an increase in government spending would induce a greater than one-for-one increase in overall GDP.

However, there are several countervailing forces. An increase in government borrowing to finance spending puts upward pressure on interest rates and exchange rates, which tends to reduce domestic consumption and investment as well as foreign demand for domestic goods. Future tax increases needed to pay off the debt act to reduce current and future consumption of households that consider their life-time income. Thus, the increase in government demand tends to crowd out private sector demand. Yet, if central banks keep interest rates unchanged, there is less crowding out and more room for multiplication.

Whether GDP ultimately goes up and by how much is a quantitative question. Answering it requires an empirically estimated macroeconomic model, which accounts for key structural features of the economy that impact on the relative magnitudes of the multiplier and crowding-out effects. Furthermore, the particular timing and path of government spending and taxes, the reaction of monetary policy, and the expectations of households and firms regarding the paths of fiscal and monetary measures exert influence on the ultimate effects of fiscal stimulus.

2.2.2 Controversy About Model-Based Evaluations of ARRA and the Zero Bound

Several studies employed macroeconomic model comparisons in order to provide policy makers with quantitative estimates of the likely impact of the above-mentioned stimulus measures. In January 2009, Christina Romer, then Chair of the President's Council of Economic Advisers, and Jared Bernstein, Chief Economist of the Office of the Vice-President, estimated that a lasting increase in government purchases of 1% of GDP would lead to a rapid increase in GDP of 1.6% persisting for at least five years. This multiplier effect was obtained by averaging the effects in two macroeconomic models—a model from an unnamed private sector forecasting firm and a model from the staff of the Federal Reserve Board—while assuming constant interest rates for the full simulation period. On this basis, Romer and Bernstein (2009) anticipated that the ARRA would raise GDP by 3.6% by the fourth quarter of 2010 and employment by 3.5 million. Their report served as important quantitative policy advice for U.S. President Obama and the Members of U.S. Congress.

In a study first circulated in March 2009, Cogan et al. (2010) questioned the validity of the Romer–Bernstein estimate and reported much smaller GDP effects for simulations with the multicountry model of Taylor (1993b) and the model of the U.S. economy by Smets and Wouters (2007). The Smets–Wouters model is representative of current thinking in macroeconomics. It is largely based on another empirically estimated and widely-known medium-size New Keynesian model developed by Christiano et al. (2005). On this basis, Cogan et al. (2010) conclude that the likely impact of the ARRA on U.S. GDP would only be around 1/6 of the Romer–Bernstein estimate.

Crowding-out effects are more important in these models because they take into account the forward-looking behavior of households and firms. Regarding fiscal policy, the path for government purchases is anticipated based on the information published with the ARRA. Regarding monetary policy, the simulations assume that the interest peg lasts between one and two years, which is more consistent with market expectations at the time than the Romer–Bernstein assumption. Afterwards, the policy rate responds again to economic conditions as in the simple policy rule of Taylor (1993a). The period of constant interest rates is motivated by the lower bound on interest rates. Due to availability of cash, a zero interest rate asset, savers need not accept negative rates on bank deposits. Thus, in a situation where the central bank reaction function calls for a negative policy rate, the rate would be constrained near zero. As a result, fiscal stimulus that raises GDP would not be followed right away by tighter monetary policy as in normal times.

Cogan et al. (2010) account for the negative effect of increased future (lump sum) taxes on household income and current consumption, but not for the negative effect of distortionary taxation on potential growth (see Drautzburg and Uhlig, 2015). Furthermore, they extend the Smets–Wouters model by including "rule-of-thumb" households. Such households consume their current income as prescribed by the Keynesian consumption function. The empirically estimated share of Keynesian consumers is about 27%. The presence of rule-of-thumb consumers and the anticipation that interest rates remain constrained at zero for two years raise the GDP impact to about 1/4 of the Romer–Bernstein estimate.

In contrast to these findings, Christiano et al. (2011) suggest that under certain conditions Keynesian multiplier effects can be much larger than one, even in modern New Keynesian models, when the zero-bound constraint on monetary policy rates binds. They present simulations of particular shocks in a small New Keynesian model and in a version of the medium-size model of Christiano et al. (2005). Thus, there appears to be stark disagreement between Cogan et al. (2010) and Christiano et al. (2011) even though both studies rely on fairly similar modern New Keynesian macroeconomic models estimated on U.S. macro data and both try to account for implications of the zero lower bound.

2.2.3 A Large-Scale Comparison Study

To illustrate how additional model comparisons can help improve policy advice in light of such disagreements, it is instructive to take a peak at a large comparison exercise that was organized by the IMF. This exercise involved several teams of researchers from central banks and international institutions which met at an IMF conference to compare a set of standardized simulations of fiscal stimulus that each team implemented in its

Table 1 Models participating in the comparison of Coenen et al. (2012)

Notation	Description
CEE	Christiano et al. (2011) model
CCTW	Cogan et al. (2010) extension of Smets and Wouters (2007) model with rule-of-thumb households
IMF-GIMF	The IMF's Global Integrated Monetary and Fiscal Policy model
FRB-US	The Federal Reserve Board's U.S. model
SIGMA	The Federal Reserve Board's two-country DSGE model
BoC-GEM	The Bank of Canada's Global Economy Model
EC-QUEST	The European Commission's QUEST model
ECB-NAWM	The European Central Bank's New Area-Wide Model
OECD	The OECD's macroeconomic model

institution's model. Key findings were summarized in the journal article by Coenen et al. (2012). The article involves 17 authors and nine different macroeconomic models. The authors are staff members of six different institutions: the International Monetary Fund, the Federal Reserve Board, the Bank of Canada, the European Central Bank, the Organization for Economic Cooperation and Development, and the European Commission. Seven models were developed and used by staff members from these institutions, while the other two models are from Cogan et al. (2010) and Christiano et al. (2011) (see Table 1).

Here, we review just one particular set of simulations from Coenen et al. (2012). In this comparison, all participating models are simulated with the same fiscal stimulus. The stimulus corresponds to the increase in government purchases as planned according to the ARRA and previously studied by Cogan et al. (2010). Technically, the spending path is simulated as an anticipated sequence of discretionary shocks. Regarding monetary policy, three different scenarios are considered that differ according to the importance of the zero bound.

The findings are presented in Fig. 1. This figure is identical to Fig. 7 of Coenen et al. (2012). The bars shown in each panel are identical and show the time profile for government purchases planned under the ARRA. The simulations assume that market participants anticipate the execution of the announced purchases over the coming years according to this plan. The different lines shown in the panels indicate the estimated impact of these government purchases on GDP in different macroeconomic models. Models estimated with euro area data are reported in the right column of panels, labeled "Europe," whereas the left column displays results for the models of the U.S. economy.

Regarding monetary policy, three different scenarios are considered. The first row of panels in Fig. 1 displays results for the case of no monetary accommodation, ie, nominal interest rates in each model are set according to a model-specific interest rate rule.

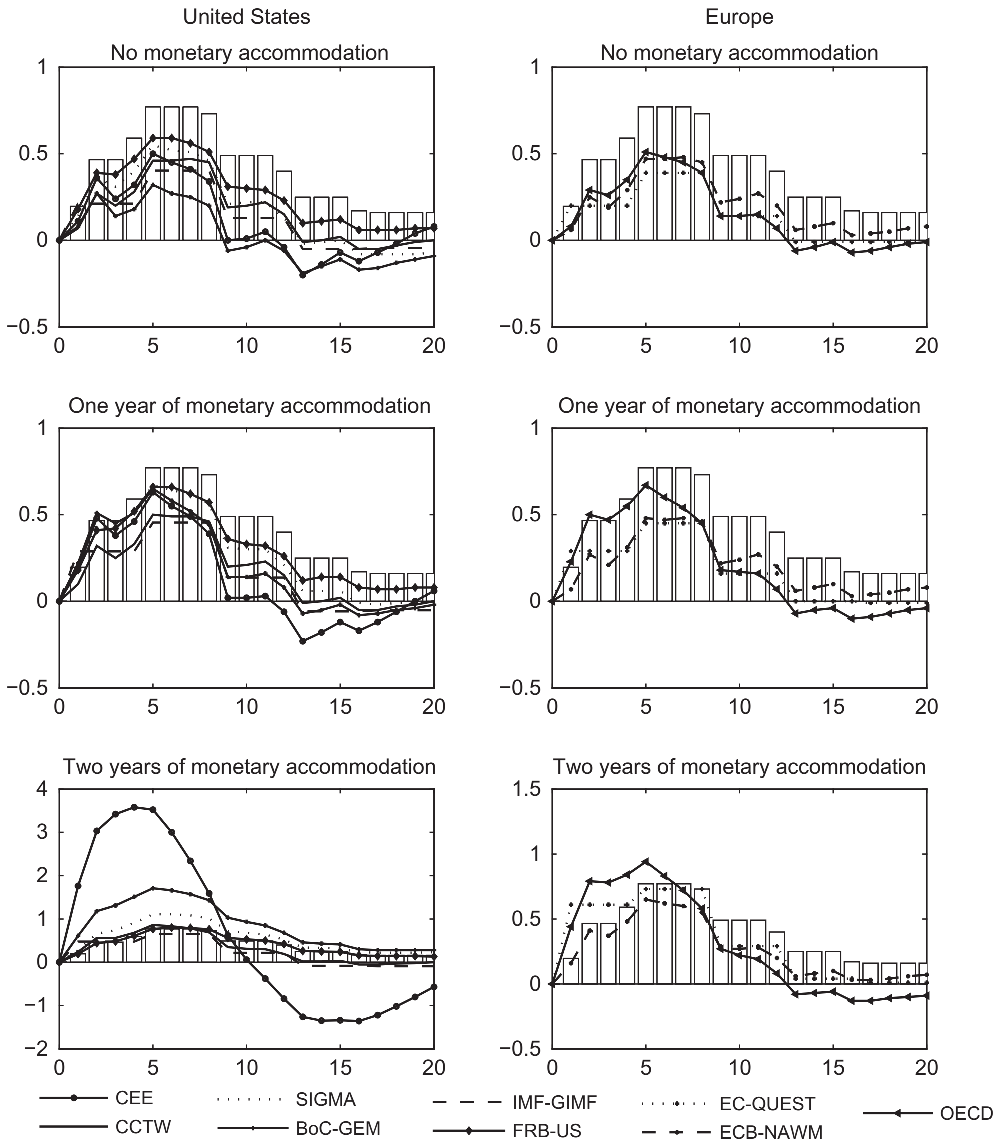


Fig. 1 Estimated GDP effects of planned ARRA spending. *Notes*: Horizontal axis represents time horizon in quarters. Units of the vertical axis are the percentage of GDP. Shown are estimated GDP effects of government purchases in the February 2009 U.S. stimulus legislation for nine macroeconomic models. The bars shown in each panel are identical and indicate the time profile of the planned ARRA government spending. CEE is the model of Christiano et al. (2011); CCTW is the model of Cogan et al. (2010); IMF-GIMF is the IMF's Global Integrated Monetary and Fiscal Policy model; FRB-US is the Federal Reserve's U.S. model; SIGMA is the Federal Reserve's two-country model; BoC-GEM is the Bank of Canada's Global Economy Model; EC-QUEST is the European Commission's QUEST model; ECB-NAWM is the European Central Bank's New Area-Wide Model; and OECD refers to the OECD's macroeconomic model.

Thus, interest rates rise along with the increase in GDP and inflation and dampen the stimulative effects of government spending. In this scenario, all models under consideration deliver an increase in GDP over the first 2.5 years of the stimulus. However, the increase in GDP remains well below the associated increase in government spending, as private demand is being crowded out by government demand. Some of the models even predict an overall negative effect on GDP in the fourth year of the stimulus. The simulation outcome of the CCTW model lies in between the other outcomes displayed in the left panel of Fig. 1. This finding provides further support for the CCTW results, in particular, since several of the other models incorporate a more detailed fiscal sector.

For the simulations shown in the second row of panels, nominal interest rates are held constant for one year and follow a model-specific rule thereafter. For the results shown in the third row, nominal interest rates are held constant for two years. These simulations illustrate the role of the degree of monetary accommodation in the effectiveness of fiscal stimulus. If nominal interest rates are initially held constant, fiscal multipliers increase. With one year of anticipated monetary accommodation, multipliers still remain below one in all of the models. With two years of monetary accommodation, the increase in GDP exceeds the increase in government spending a little bit in some of the models, due to crowding-in of private demand. There is one outlier. The CEE model exhibits a very large effect for two years, followed by a recession. As suggested in Christiano et al. (2011), government spending multipliers may be large in this model. Yet, all the other models considered by Coenen et al. (2012) imply much smaller effects on GDP. Thus, the larger multiplier effect in the CEE model is not just due to the monetary accommodation resulting from the presence of the zero bound. There are other features of the CEE model that lead to greater effectiveness of fiscal stimulus in this scenario. Coenen et al. (2012) suggest that the CEE model is an outlier, because it exhibits a much lower degree of price rigidity.

In sum, the large-scale comparison exercise confirms the cautionary assessment of the likely impact of ARRA provided by Cogan et al. (2010) and helps identify outliers. It would certainly have been useful to have such a large-scale comparison available in 2009 in order to provide policy advice to the Obama Administration and the members of U.S. Congress. Thus, it is of great interest to explore how model comparisons can be implemented more easily and more frequently whenever such analysis can help inform policy makers in real time.

3. A SYSTEMATIC APPROACH TO MODEL COMPARISON

One important goal of model comparison exercises is to identify policy implications that are due to different structural features of the respective models. Yet, quantitative simulation results may also differ because the economic concepts and variables to be compared are not defined consistently across models. Furthermore, different outcomes may be due to different assumptions about policy rather than different structures of the economy.

This section briefly describes how macroeconomic models can be augmented systematically with a few equations to produce comparable objects concerning policy implications for key macroeconomic aggregates, while keeping the total number of modifications quite small. This formal approach is elaborated on in Wieland et al. (2012).

3.1 Notation for a General Nonlinear Model

The following notation is used to define a general nonlinear model of the economy. The superscript $m=(1,2,3,\ldots,M)$ denotes the equations, variables, parameters, and shocks of a specific model m that is to be included in the comparison. These model-specific objects need not be comparable across models. They are listed in Table 2. In the computational implementation, m corresponds to an abbreviated model name rather than simply a number.

Two types of model equations are distinguished. Policy rules are denoted by $g_m(.)$ while all other equations and identities are denoted by $f_m(.)$. Together, they determine the endogenous variables denoted by the vector x_t^m. These variables are functions of each other, of model-specific shocks, $[\epsilon_t^m \quad \eta_t^m]$, and of model-specific parameters $[\beta^m \quad \gamma^m]$. A particular model m is then defined by:

$$E_t[g_m(x_t^m, x_{t+1}^m, x_{t-1}^m, \eta_t^m, \gamma^m)] = 0 \tag{1}$$

$$E_t[f_m(x_t^m, x_{t+1}^m, x_{t-1}^m, \epsilon_t^m, \beta^m)] = 0 \tag{2}$$

The superscript m refers to the version of the respective model originally presented by its authors. The model may include current values, lags, and the expectation of leads of endogenous variables. In Eqs. (1) and (2), the lead and lag lengths are set to unity for notational convenience. Additional leads and lags can be accommodated with auxiliary variables. Even so, our software implementation does not restrict the lead and lag lengths of participating models.

The model may also include innovations that are random variables with zero mean and covariance matrix, Σ^m:

Table 2 Model-specific variables, parameters, shocks, and equations

Notation	Description
x_t^m	Endogenous variables in model m
$x_t^{m,g}$	Policy variables in model m (also included in x_t^m)
η_t^m	Policy shocks in model m
ϵ_t^m	Other economic shocks in model m
$g_m(.)$	Policy rules in model m
$f_m(.)$	Other model equations in model m
γ^m	Policy rule parameters in model m
β^m	Other economic parameters in model m
Σ^m	Covariance matrix of shocks in model m

$$E([\eta_t^m \ \epsilon_t^m]') = 0 \tag{3}$$

$$E([\eta_t^{m\prime} \ \epsilon_t^{m\prime}]'[\eta_t^{m\prime} \ \epsilon_t^{m\prime}]) = \Sigma^m = \begin{pmatrix} \Sigma_\eta^m & \Sigma_{\eta\epsilon}^m \\ \Sigma_{\eta\epsilon}^m & \Sigma_\epsilon^m \end{pmatrix} \tag{4}$$

We refer to innovations interchangeably as shocks. Some models include serially correlated economic shocks that are themselves functions of random innovations. In our notation, such serially correlated economic shocks would appear as elements of the vector of endogenous variables x_t^m, only the innovations would appear as shocks. Eq. (4) distinguishes the covariance matrices of policy shocks and other economic shocks as Σ_η^m and Σ_ϵ^m. The correlation of policy shocks and other shocks is typically assumed to be zero, $\Sigma_{\eta\epsilon}^m = 0$.

3.2 Introducing Common Variables, Parameters, Equations, and Shocks

In order to compare policy implications from different models, it is necessary to define a set of comparable variables, shocks, and parameters. They are common to all models considered. Policies can then be expressed in terms of such common parameters, variables and policy shocks, and their consequences can be calculated for a set of common endogenous variables. Our notation for comparable endogenous variables, policy instruments, policy shocks, policy rules, and parameters is given in Table 3.

Every model to be included in the comparison has to be augmented with common variables, parameters, and shocks. Augmenting the model requires adding some equations. These additional equations serve to define the common variables in terms of model-specific variables. We denote these definitional equations or identities by $h_m(.)$. They are necessarily model-specific. Additionally, the original model-specific policy rules need to be replaced with common policy rules. Of course, these common rules could be defined generally enough such that they nest many of the model-specific policy rules. Furthermore, there are many interesting questions that would require comparing model implications for common variables of interest when policy follows the respective model-specific rule. An example is provided in Section 5.3.

All the other equations, variables, parameters, and shocks may be preserved in the original notation of the model's authors. Consequently, the augmented model consists

Table 3 Comparable common variables, parameters, shocks, and equations

Notation	Description
z_t	Common variables in all models
z_t^g	Common policy variables in all models (also included in z_t)
η_t	Common policy shocks in all models
$g(.)$	Common policy rules
γ	Common policy rule parameters

of three components: (i) the common policy rules, $g(.)$, expressed in terms of common variables, z_t, policy shocks, η_t, and policy parameters, γ; (ii) the model-specific definitions of common variables in terms of original model-specific endogenous variables, $h_m(.)$, with parameters θ^m; (iii) the original set of model-specific equations $f_m(.)$ that determine the endogenous variables. It corresponds to:

$$E_t[g(z_t, z_{t+1}, z_{t-1}, \eta_t, \gamma)] = 0 \tag{5}$$

$$E_t[h_m(z_t, x_t^m, x_{t+1}^m, x_{t-1}^m, \theta^m)] = 0 \tag{6}$$

$$E_t[f_m(x_t^m, x_{t+1}^m, x_{t-1}^m, \epsilon_t^m, \beta^m)] = 0 \tag{7}$$

Models augmented accordingly are ready for comparing policy implications. For example, it is then straightforward to compare the consequences of a particular policy rule for the dynamic behavior of consistently defined endogenous variables across models. This approach requires only a limited number of common elements. The rest of each model remains unchanged in the authors' original notation. This includes the variable names and definitions of endogenous variables, x_t^m, the other economic shocks ϵ_t^m, the equations $f_m(.)$ with model parameters β^m and the covariance matrix of shocks Σ_ϵ^m. The covariance matrix of policy shocks Σ_η may be treated as an element of the vector of policy parameters or set to zero.

Wieland et al. (2012) provide some concrete examples for the model augmentation step, which includes setting up the additional definitional equations, $h_m(.)$, and determining their parameters, θ^m. The subsequent steps in comparing policy implications consist of solving the augmented models, constructing appropriate objects for comparison and computing a metric that quantifies the differences of interest.

3.3 Computing Comparable Policy Implications

Solving the augmented nonlinear structural model defined by Eqs. (5)–(7) involves expressing the expectations of future variables in terms of currently available information. To this end, one needs to define how expectations are formed. Our computational implementation and model archive, Macroeconomic Model Data Base, includes models using several different assumptions. While most of the models are solved under the assumption of rational model-consistent expectations, several models can also be solved under the assumption of adaptive learning in expectations as in Slobodyan and Wouters (2012). Other assumptions regarding expectations formation include the sticky-information model of Mankiw and Reis (2007) with staggered information sets of otherwise rational expectations and VAR-based expectations used in Orphanides (2003) and in a version of the Federal Reserve's FRB-US model.

Here, we proceed under the assumption of rational expectations. The solution step involves checking for existence and uniqueness of equilibrium. For linear models one can use the Blanchard–Kahn conditions. For nonlinear models one may have to rely on search by numerical methods. The solution of the structural model is given by a set of reduced-form equations:

$$z_t = k_z(z_{t-1}, x^m_{t-1}, \eta_t, \epsilon^m_t, \kappa_z) \tag{8}$$

$$x^m_t = k_x(z_{t-1}, x^m_{t-1}, \eta_t, \epsilon^m_t, \kappa_x) \tag{9}$$

If the structural model is nonlinear, the reduced-form equations will also be nonlinear. (κ_z, κ_x) denote the reduced-form parameters. They are complicated functions of the structural parameters, β^m, the policy parameters, γ, and the covariance matrix, Σ^m. Nonlinear models may be solved approximately by means of numerical methods, for example, perturbation-based, projection-based, or two-point-boundary-value algorithms (see Judd, 1998; Fair and Taylor, 1983; Collard and Juillard, 2001). When the model is first linearized around a deterministic steady state, either analytically or numerically, a range of methods are available for computing the solution to the linear system of expectational equations. These methods include the generalized eigenvalue–eigenvector method (see Uhlig, 1995), generalized Schur decomposition (see Klein, 2000), QZ decomposition (see Sims, 2001), the undetermined coefficients method (see Christiano, 2002), or the Anderson–Moore algorithm for solving linear saddle point models (see Anderson and Moore, 1985).

The reduced form solution of the augmented nonlinear model can then be used to obtain particular objects for comparison defined in terms of comparable variables. With regard to policy implications, one object of interest could be the impact of a policy shock and its transmission to key macroeconomic aggregates. This object corresponds to the dynamic response of a particular common variable (an element of z) to a policy shock η_t, conditional on a certain common policy rule, $g(z_t, z_{t+1}, z_{t-1}, \eta_t, \gamma)$. Such impulse response functions describe the isolated effect of a single shock on the dynamic system holding everything else constant. Other objects of interest for comparing policy implications would be the unconditional variances and serial correlation functions. Finally, one may compute suitable metrics for measuring the distance between two or more models. Such metrics could be the absolute difference of the unconditional variances or the absolute difference of the impact effects of policy shocks under different models.

4. PRACTICAL PROBLEMS AND A NEW PLATFORM

Large-scale macroeconomic model comparison exercises have been relatively rare. These exercises are costly because they typically involve multiple meetings of several teams of model developers, with each team analyzing the policy scenarios in its own model. At the same time, the number of policy scenarios studied in these exercises has been limited. In this section, we review some practical problems that have hampered easy and frequent use of model comparison. We also report on the experience with strategies employed in the construction of the Macroeconomic Model Data Base to overcome these problems. At this point, there are 66 models available for easy use by individual researchers and students. It is straightforward to include new models and compare their policy implications to existing benchmarks.

4.1 Replication, Computational Implementation, and Model Archiving

4.1.1 Replication

The first practical problem that arises if a researcher wishes to compare her macroeconomic model to those of others is how to obtain sufficient information about their models to conduct her own analysis. Replicability is a basic scientific principle. The web-course "*Understanding science 101*" at UC Berkeley describes this principle as follows:

> *"Scientists aim for their studies' findings to be replicable - so that, for example, an experiment testing ideas about the attraction between electrons and protons should yield the same results when repeated in different labs. Similarly, two different researchers studying the same dinosaur bone in the same way should come to the same conclusions regarding its measurements and composition. This goal of replicability makes sense. After all, science aims to reconstruct the unchanging rules by which the universe operates, and those same rules apply, 24 hours a day, seven days a week, from Sweden to Saturn, regardless of who is studying them. If a finding can't be replicated, it suggests that our current understanding of the study system or our methods of testing are insufficient."*

Unfortunately, however, there is no general practice guaranteeing replicability of macroeconomic models that are solved and simulated by means of computational methods. This state of the field is somewhat surprising compared to other fields of economics. In economic theory, it is standard that articles in scientific economic journals provide sufficient detail on mathematical derivations and proofs such that academics and advanced students can replicate the analysis. In econometrics, new methods and estimators are fairly quickly implemented in software packages such as EViews, RATS, SAS, GAUSS, and others. Thus, new econometric tools are not only spread to academic researchers and students but widely used by practitioners in many fields of applied economic analysis. In the last two decades, macroeconomic modeling has benefited from a similar development with regard to numerical techniques for solving and estimating models with rational expectations. Initially, individual researchers have made particular toolkits available that have been adopted by many others in their work. Over the years, the software package DYNARE developed by Michel Juillard and collaborators has gained more and more users and contributors such that it has become a widely used tool for macroeconomic model solution and estimation (see Juillard, 2001 and Adjemian et al., 2011). While new techniques for model solution and estimation can now be easily employed by academics, students, and practitioners almost as "black box" systems, this is not true for most of the many new macroeconomic models.

The following problems can arise when one attempts to replicate macroeconomic models presented in economic journals:

1. The published article does not contain all the equations needed to write the model code for replicating the analysis presented in the article. Typically, journals are not willing to devote space to present all the information that is needed. Also, the models are quite complex. Thus, errors may occur in transcribing model equations that were successfully implemented in computer code to the text file for the article.
2. The published article does not contain all the parameter values or steady-state values needed to replicate the model simulations reported in the article.
3. The code for replicating the model is not available from the journal's website. While many journals provide options for online archiving of supplementary materials, only a few have the capacity to insist in every case that authors provide a reliable version of their code.
4. The code is not available from the authors' website and authors are not replying to requests for the code.
5. The code is available but the software needed to simulate is unavailable to individual researchers because its price is high and it is only used at large institutions. An example is the TROLL software used at some policy-making institutions.
6. The code is available but the simulation results it delivers differ from the results published in the article. Such inconsistencies may simply be due to differences in the date of the version used for preparing the results shown in the article and the version made available for replication.
7. The code that is available does not contain sufficient description and explanations such that it is easily understood by users.
8. The software platform for which the code has been written has been updated such that the code cannot be executed successfully anymore on this platform.
9. The researcher attempting replication makes errors in his implementation of the model.
10. The published model cannot be replicated correctly because the derivation of the equations or their implementation in computer code contain errors. Given the complex nature of computational implementation of medium- and large-scale macroeconomic models, such errors are to be expected and can happen to the most meticulous scholars. It is useful to recognize and correct them so as to make it easier for other researchers to build on this work.

These difficulties are not unique to macroeconomic modeling. Replication in reference to computations is more commonly known as *reproducible research* and forms the subject of an expanding literature in computer science, statistics, and related fields of application (see for example Fomel and Claerbout, 2009; Donoho, 2010; Freire et al., 2012; and Sandve et al., 2013). Stanford statistician Donoho (2010) characterizes the central problem in these words:

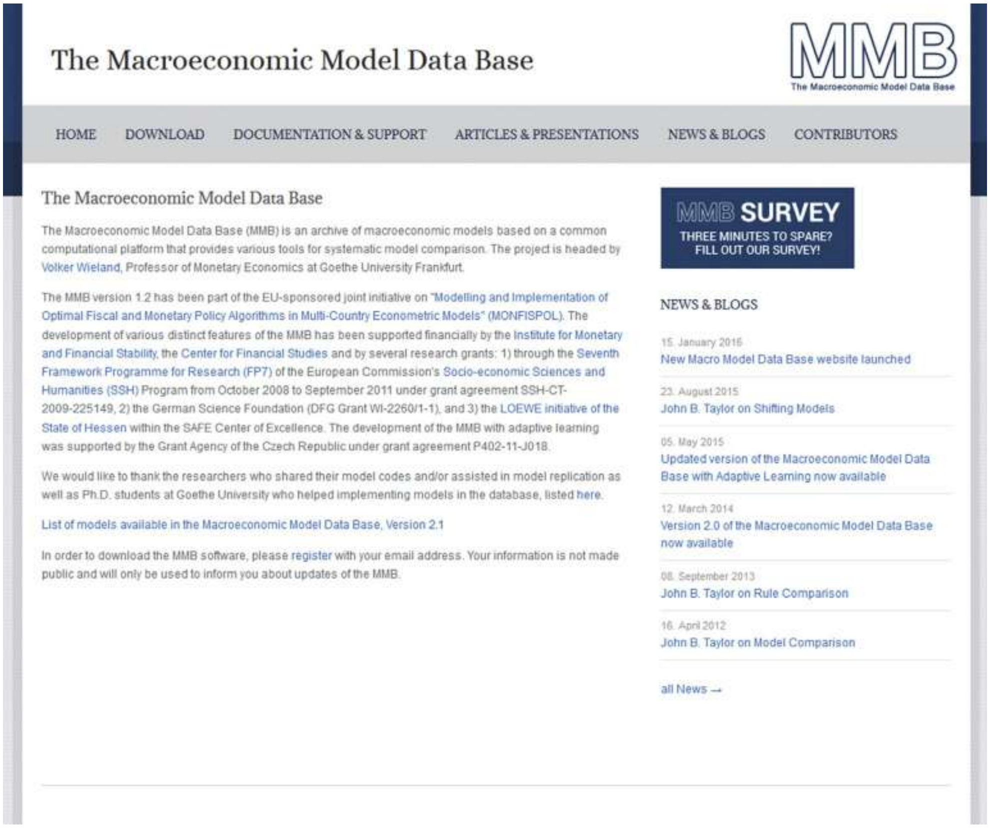
> "an article about computation result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result."[b]

We have pursued the following strategies for replicating models to be included in our model comparison software:

1. The ideal case is that authors or other users of the model provide the code describing the model and integrate it themselves in MMB. Generally, authors can expect wider dissemination, use and citation of their work by other researchers if they make their code available in an easy-to-use format. We have also found that policy-making institutions such as central banks and international organizations have become very open toward making their models available, at least those versions that economists from these institutions have circulated in working papers or used for publications in scientific journals.
2. The next best scenario is when model authors provide the complete code that replicates the findings reported in their article and remain available for answering questions of research assistants in Frankfurt who integrate the model in MMB.
3. Research assistants in the Frankfurt MMB team have replicated a number of models using software made available on journal or author websites.
4. We teach advanced Ph.D. courses that focus on a particular area of new model development. A team of two or three students receives the task of presenting a paper from the literature, replicating the model and integrating the model in MMB. This approach has proved quite successful in terms of training Ph.D. students in model building and getting them to the research frontier, where they can work on extending existing models for new applications. Whether they succeed in replicating the model often depends on whether they receive feedback from model authors on problems or missing items. Students give presentations on the original paper and the technical replication and they also prepare a replication report.
5. Once a model has been replicated, we make the files documenting the replication available for download on the MMB website (http://www.macromodelbase.com) as shown in Fig. 2. The replication package is offered separately from the comparison software itself. It is not augmented for model comparison and remains as close as possible to the authors' original code or article. A readme file and graphics files make reference to the specific original research findings and provide information on how close we came to matching the authors' work.

In total, MMB 2.1 makes available 66 models. Taking into account that some of them are simple variants of one model, MMB 2.1 includes 57 quite distinct models. Out of 57 models, about 12 models were made ready for integration by their respective authors or other researchers, 31 models were implemented by the MMB team in Frankfurt in

[b] Noteworthy, this quote is inspired by Claerbout (1994). Very similar ideas are expressed much earlier by A. Williams (see White, 1978).



The Macroeconomic Model Data Base

The Macroeconomic Model Data Base

The Macroeconomic Model Data Base (MMB) is an archive of macroeconomic models based on a common computational platform that provides various tools for systematic model comparison. The project is headed by Volker Wieland, Professor of Monetary Economics at Goethe University Frankfurt.

The MMB version 1.2 has been part of the EU-sponsored joint initiative on "Modelling and Implementation of Optimal Fiscal and Monetary Policy Algorithms in Multi-Country Econometric Models" (MONFISPOL). The development of various distinct features of the MMB has been supported financially by the Institute for Monetary and Financial Stability, the Center for Financial Studies and by several research grants: 1) through the Seventh Framework Programme for Research (FP7) of the European Commission's Socio-economic Sciences and Humanities (SSH) Program from October 2008 to September 2011 under grant agreement SSH-CT-2009-225149, 2) the German Science Foundation (DFG Grant WI-2260/1-1), and 3) the LOEWE initiative of the State of Hessen within the SAFE Center of Excellence. The development of the MMB with adaptive learning was supported by the Grant Agency of the Czech Republic under grant agreement P402-11-J018.

We would like to thank the researchers who shared their model codes and/or assisted in model replication as well as Ph.D. students at Goethe University who helped implementing models in the database, listed here.

List of models available in the Macroeconomic Model Data Base, Version 2.1

In order to download the MMB software, please register with your email address. Your information is not made public and will only be used to inform you about updates of the MMB.

MMB SURVEY
THREE MINUTES TO SPARE?
FILL OUT OUR SURVEY!

NEWS & BLOGS

15. January 2016
New Macro Model Data Base website launched

23. August 2015
John B. Taylor on Shifting Models

05. May 2015
Updated version of the Macroeconomic Model Data Base with Adaptive Learning now available

12. March 2014
Version 2.0 of the Macroeconomic Model Data Base now available

08. September 2013
John B. Taylor on Rule Comparison

16. April 2012
John B. Taylor on Model Comparison

all News →

Fig. 2 MMB Website.

cooperation with the authors, and the remainder were integrated on the basis of course work by Ph.D. students.

4.1.2 Computational Implementation

In terms of implementing the model comparison approach outlined in the previous section computationally, there are choices to be made regarding computer language as well as model solution and simulation methods. Furthermore, problems to be dealt with concern the compatibility with earlier or subsequent versions of the respective software solutions and operating systems.

Most academic researchers in the area of macroeconomic model building have adopted MATLAB as their preferred high-level programming language. This choice concerns specifically the recent development of DSGE models in the real business cycle and New Keynesian literatures. MATLAB—the name is derived from MATrix LABoratory—is a commercial software product of MATHWORKS Inc. It is fairly widely used in engineering, physics, economics, and other fields applying computational methods. This software product is expensive but there are discounts for student licences. Also, there exists a freeware software GNU OCTAVE that is largely compatible with the proprietary MATLAB software. Thus, executables that run on MATLAB can presumably be executed on OCTAVE without needing major modifications. Competing software packages such as GAUSS or MATHEMATICA are not as popular in macroeconomic modeling but offer advantages in econometric or symbolic methods, respectively.

Developers of numerical solution methods for macroeconomic models with rational expectations have written routines that are MATLAB executables for a long time. Over recent years, the free software package DYNARE has been adopted by many researchers in academia, central banks, and international organizations that are working in the field of macroeconomic modeling (see www.dynare.org). DYNARE runs on MATLAB but can also be used with OCTAVE. There is a growing community of researchers that is contributing freely available routines for solution, estimation, and optimization to the DYNARE environment. Some central banks and international organizations also employ the software system TROLL for simulating models used in policy formulation. TROLL is a commercial software with features that make it easy to manage large data sets.

MMB has been developed as free software to be used with DYNARE and MATLAB. Models are defined in the syntax needed for DYNARE. It should also be possible to use the first version of MMB 1.2 and DYNARE with the free software OCTAVE. Yet, so far we have not had the resources to ensure that MMB 2.0 and 2.1 are OCTAVE compatible. MMB 2.0 has been extended with graphical user interfaces (GUI) to improve user friendliness. At this point, GUI facilities are apparently not yet available on OCTAVE, thus restricting MMB 2.0 to MATLAB environments. A Mac OS compatible version of MMB 1.2 is available for download thanks to the contribution of Raymond Hawkins from the University of Arizona.

4.2 User Friendliness and a MATLAB-Based Platform for Comparative Analysis

4.2.1 User Friendliness

The first version of MMB 1.2 was intended for researchers that work on building macroeconomic models. MMB 2.0 and updates are meant to be accessible to a wider group of interested professional economists in the public and private sector and to students of macroeconomics. Thus, we have built graphical user interfaces that make it easier to simulate a wide variety of scenarios with any of the models included in the archive.

First, the user can choose among different applications, such as the comparison of different models under a common policy rule (*One policy rule, many models*) or an in-depth analysis of one specific model under different policy rules (*One model, many policy rules*).

Then she is offered a menu of choices concerning models, policy rules, simulation scenarios, and output formats.

For example, the menu for *One policy rule, many models* is shown in Fig. 3. This menu offers options to conduct comparisons across models under the assumption that the central bank in each model implements the same interest rate rule. It gives access to a software implementation of the mathematical representation of model comparison in Section 3.

On the left-side of the menu the user can choose multiple models by checking the respective boxes. Models are grouped under different categories, such as calibrated New Keynesian models, estimated models of the U.S. economy, estimated models of the euro area economy, models of other economies such as Canada, Chile, Brazil, or Hong Kong and finally several multicountry models. A button on the bottom right of the menu titled "Models description" leads to a PDF file with further information on the models included in the archive. On the top right side, there is a section for choosing a common policy rule from a list of rules. Alternatively, the user can enter coefficients for the common rule in a submenu popping up after choosing "User-specified rule". Furthermore, there are various options for generating simulation output such as unconditional variances, autocorrelation functions, and impulse response functions to monetary and fiscal policy shocks.

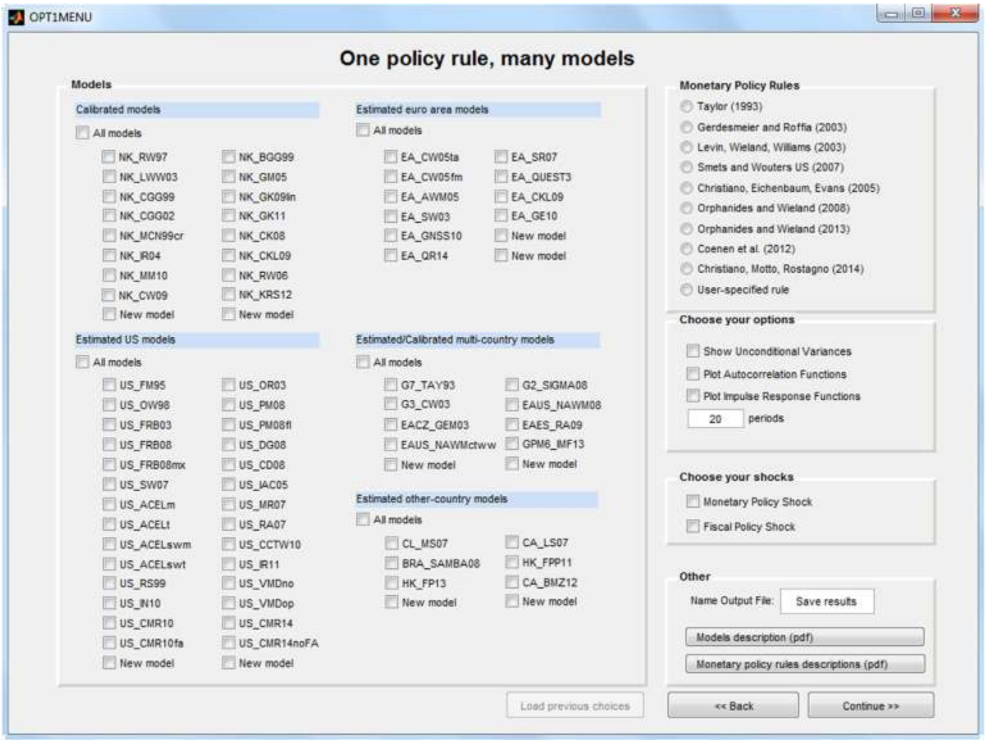
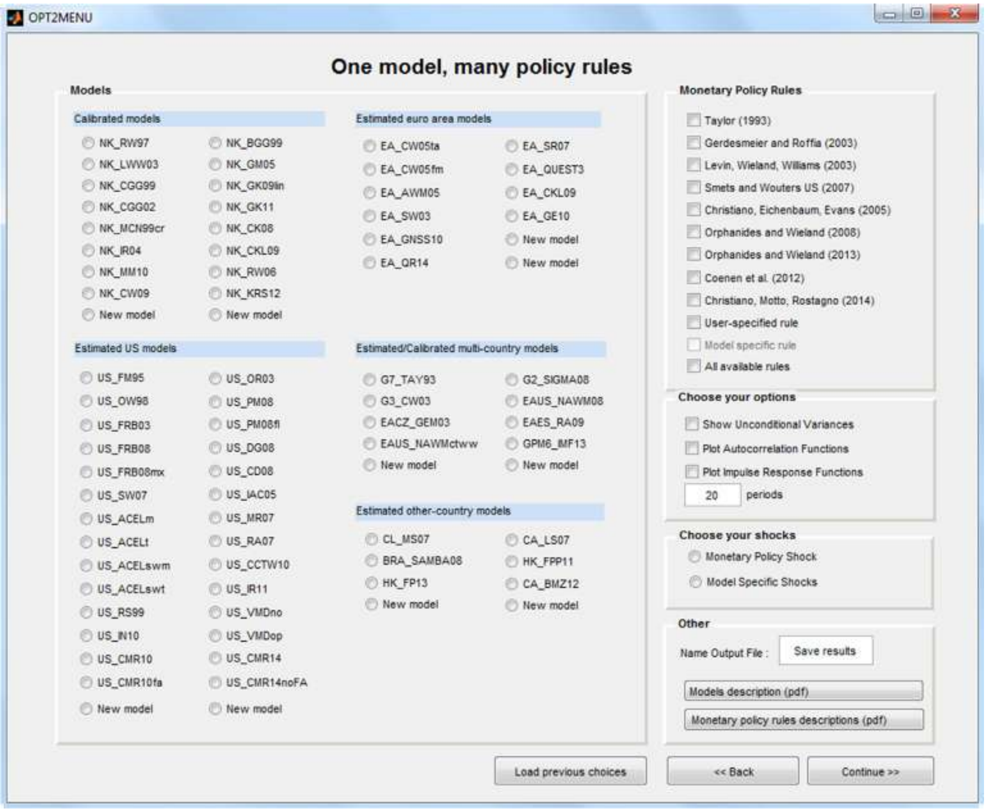


Fig. 3 Modelbase menu: One policy rule, many models.

4.2.2 Common and Model-Specific Policy Rules

The comparison using a common policy rule makes it possible to identify differences in policy implications that are due to differences in model structure and parameter estimates. Yet, there are other interesting questions one might want to ask. For example, it may be of interest to explore the dynamics of one particular model under a variety of different policy rules in more detail. And there are questions that would require simulating each model under the original policy rule estimated or calibrated by the model authors. For example, one would use the model-specific rules if one wants to compare the fit of each model to the data, if one wants to identify the typical empirical response to a particular model-specific shock, or if one wants to compare forecasts obtained from different models.

The application *One model, many policy rules*, for which the menu is shown in Fig. 4, allows a thorough investigation of the properties of a single model and can be used to compare the implications of a variety of policy rules in this model. The user can only



Models

Calibrated models

○ NK_RW97
○ NK_LWW03
○ NK_CGG99
○ NK_CGG02
○ NK_MCN99cr
○ NK_IR04
○ NK_MM10
○ NK_CW09
○ New model
○ NK_BGG99
○ NK_GM05
○ NK_GK09lin
○ NK_GK11
○ NK_CK08
○ NK_CKL09
○ NK_RW06
○ NK_KRS12
○ New model

Estimated US models

○ US_FM95
○ US_OW98
○ US_FRB03
○ US_FRB08
○ US_FRB08mx
○ US_SW07
○ US_ACELm
○ US_ACELt
○ US_ACELswm
○ US_ACELswt
○ US_RS99
○ US_IN10
○ US_CMR10
○ US_CMR10fa
○ New model
○ US_OR03
○ US_PM08
○ US_PM08fl
○ US_DG08
○ US_CD08
○ US_IAC05
○ US_MR07
○ US_RA07
○ US_CCTW10
○ US_IR11
○ US_VMDno
○ US_VMDop
○ US_CMR14
○ US_CMR14noFA
○ New model

Estimated euro area models

○ EA_CW05ta
○ EA_CW05fm
○ EA_AWM05
○ EA_SW03
○ EA_GNSS10
○ EA_QR14
○ EA_SR07
○ EA_QUEST3
○ EA_CKL09
○ EA_GE10
○ New model
○ New model

Estimated/Calibrated multi-country models

○ G7_TAY93
○ G3_CW03
○ EACZ_GEM03
○ EAUS_NAWMctww
○ New model
○ G2_SIGMA08
○ EAUS_NAWM08
○ EAES_RA09
○ GPM6_IMF13
○ New model

Estimated other-country models

○ CL_MS07
○ BRA_SAMBA08
○ HK_FP13
○ New model
○ CA_LS07
○ HK_FPP11
○ CA_BMZ12
○ New model

Monetary Policy Rules

□ Taylor (1993)
□ Gerdesmeier and Roffia (2003)
□ Levin, Wieland, Williams (2003)
□ Smets and Wouters US (2007)
□ Christiano, Eichenbaum, Evans (2005)
□ Orphanides and Wieland (2008)
□ Orphanides and Wieland (2013)
□ Coenen et al. (2012)
□ Christiano, Motto, Rostagno (2014)
□ User-specified rule
□ Model specific rule
□ All available rules

Choose your options

□ Show Unconditional Variances
□ Plot Autocorrelation Functions
□ Plot Impulse Response Functions
20 periods

Choose your shocks

○ Monetary Policy Shock
○ Model Specific Shocks

Other

Name Output File : Save results

Models description (pdf)

Monetary policy rules descriptions (pdf)

Fig. 4 Modelbase menu: One model, many policy rules.

choose one model at a time, but multiple policy rules. It is possible to list the structural shocks in each model and simulate impulse responses for some or all of them under the different rules. In addition to the list of rules and the user-specified rule, the rules menu also includes the model-specific rule estimated or calibrated by the original model's authors as long as the model-specific rule can be written in terms of MMB common variables.

4.2.3 How to Include Your Own Model in MMB

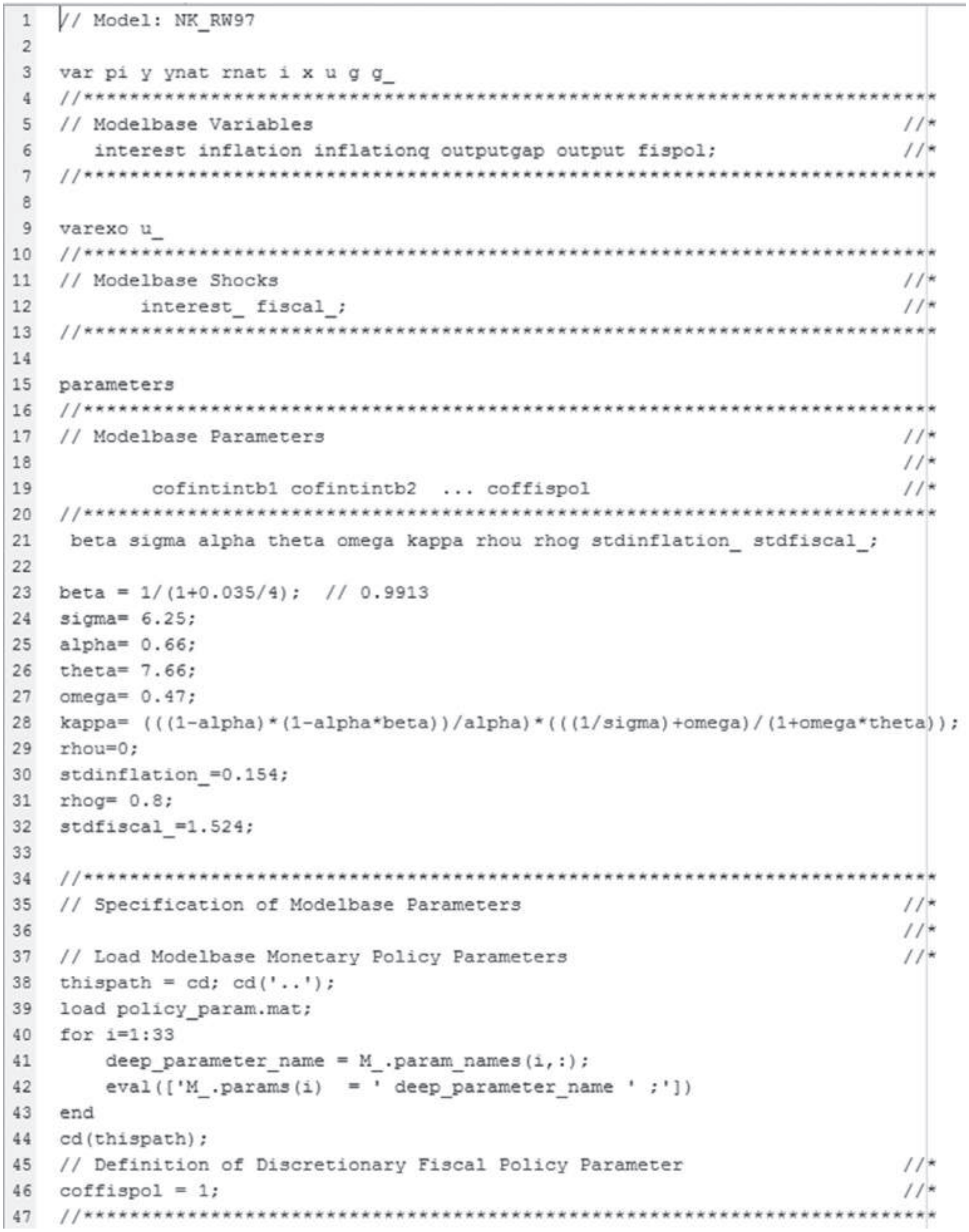
It is fairly straightforward to include additional models in the archive. A detailed description of the necessary steps is provided in the MMB User Guide that can be downloaded along with the MMB software. Thus users can easily integrate their own model for comparison with the models in the archive. The new model can be assigned a button in the graphical user interface. If users send their model file to the model base team in Frankfurt, it can also be included in the publicly available archive.

The complete process of augmenting a model has been described formally in Section 3. If modelers have already simulated their model using DYNARE, they only need to make a few adjustments and additions to the DYNARE model in order to integrate their model in the MMB software. To illustrate this process, Figs. 5 and 6 present the central elements of the DYNARE model file with the New Keynesian model by Rotemberg and Woodford (1997) (NK_RW97) in MMB. A typical model file is comprised of the preamble block, in which variables and parameters are initialized, and the model block.[c]

With regard to the preamble of their model file, contributors simply need to copy and paste in the common variables, common policy shocks, and common policy parameters from another MMB model file. The lines of code that need to be pasted in are shown between starred lines in the preamble section in Fig. 5. They are the same for all MMB model files.

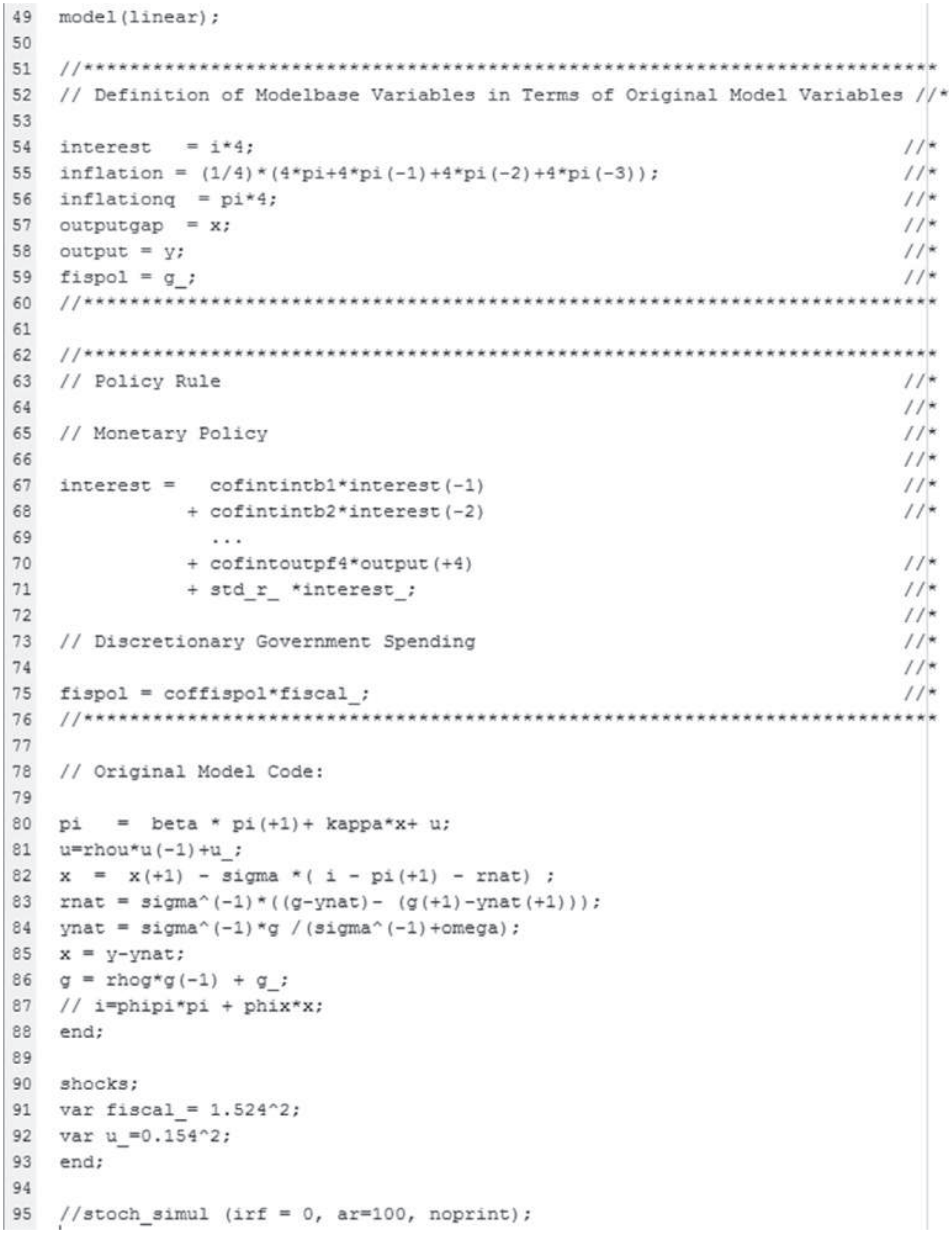
The augmented model block consists of three parts: (i) the common policy rules ($g(.)$ in Eq. (5)); (ii) the definitional equations ($h_m(.)$ in Eq. (6)); (iii) the original model equations ($f_m(.)$ in Eq. (7)). Including the common policy rules is simply another "copy and paste" operation (see lines 63–75 in Fig. 6). Of course, the model-specific monetary policy rule then needs to be commented out (see line 87). The only step that requires more knowledge of the original model concerns adding the definitions of the common variables in terms of model-specific variables to the code. Table 4 describes the relevant common variables. The resulting definitional equations in the case of the NK_RW97 model can be found in lines 54–59 of Fig. 6.

[c] For more detailed explanations, please refer to section 1.4 in the MMB User Guide available online at www.macromodelbase.com.



```
// Model: NK_RW97

var pi y ynat rnat i x u g g_
//***************************************************************************
// Modelbase Variables                                                    //*
   interest inflation inflationq outputgap output fispol;                 //*
//***************************************************************************

varexo u_
//***************************************************************************
// Modelbase Shocks                                                       //*
       interest_ fiscal_;                                                 //*
//***************************************************************************

parameters
//***************************************************************************
// Modelbase Parameters                                                   //*
                                                                          //*
        cofintintb1 cofintintb2  ... coffispol                            //*
//***************************************************************************
 beta sigma alpha theta omega kappa rhou rhog stdinflation_ stdfiscal_;

beta = 1/(1+0.035/4);  // 0.9913
sigma= 6.25;
alpha= 0.66;
theta= 7.66;
omega= 0.47;
kappa= (((1-alpha)*(1-alpha*beta))/alpha)*(((1/sigma)+omega)/(1+omega*theta));
rhou=0;
stdinflation_=0.154;
rhog= 0.8;
stdfiscal_=1.524;

//***************************************************************************
// Specification of Modelbase Parameters                                  //*
                                                                          //*
// Load Modelbase Monetary Policy Parameters                              //*
thispath = cd; cd('..');
load policy_param.mat;
for i=1:33
    deep_parameter_name = M_.param_names(i,:);
    eval(['M_.params(i)  = ' deep_parameter_name ' ;'])
end
cd(thispath);
// Definition of Discretionary Fiscal Policy Parameter                    //*
coffispol = 1;                                                            //*
//***************************************************************************
```

Fig. 5 Structure of the model file for Rotemberg and Woodford (1997) (NK_RW97): The preamble.



```
model(linear);

//***************************************************************************
// Definition of Modelbase Variables in Terms of Original Model Variables //*

interest   = i*4;                                                         //*
inflation = (1/4)*(4*pi+4*pi(-1)+4*pi(-2)+4*pi(-3));                      //*
inflationq  = pi*4;                                                       //*
outputgap   = x;                                                          //*
output = y;                                                               //*
fispol = g_;                                                              //*
//***************************************************************************

//***************************************************************************
// Policy Rule                                                            //*
                                                                          //*
// Monetary Policy                                                        //*
                                                                          //*
interest =    cofintintb1*interest(-1)                                    //*
            + cofintintb2*interest(-2)                                    //*
              ...
            + cofintoutpf4*output(+4)                                     //*
            + std_r_ *interest_;                                          //*
                                                                          //*
// Discretionary Government Spending                                      //*
                                                                          //*
fispol = coffispol*fiscal_;                                               //*
//***************************************************************************

// Original Model Code:

pi   =  beta * pi(+1)+ kappa*x+ u;
u=rhou*u(-1)+u_;
x  =  x(+1) - sigma *( i - pi(+1) - rnat) ;
rnat = sigma^(-1)*((g-ynat)- (g(+1)-ynat(+1)));
ynat = sigma^(-1)*g /(sigma^(-1)+omega);
x = y-ynat;
g = rhog*g(-1) + g_;
// i=phipi*pi + phix*x;
end;

shocks;
var fiscal_= 1.524^2;
var u_=0.154^2;
end;

//stoch_simul (irf = 0, ar=100, noprint);
```

Fig. 6 Structure of the model file for Rotemberg and Woodford (1997) (NK_RW97): The model block.

Table 4 Comparable common variables in MMB

Notation	Variable name	Description
i_t^z	*Interest*	Annualized quarterly money market rate
g_t^z	*Fispol*	Discretionary government purchases (share in GDP)
π_t^z	*Inflation*	Year-on-year rate of inflation
p_t^z	*Inflationq*	Annualized quarter-to-quarter rate of inflation
y_t^z	*Output*	Quarterly real GDP
q_t^z	*Outputgap*	Quarterly output gap (deviation from flex-price level)

5. COMPARING FISCAL AND MONETARY POLICY TRANSMISSION USING THE NEW PLATFORM

The Macroeconomic Model Data Base offers individual users many options for comparing model structures and policy implications and for exploring a particular model in great detail. There is no need for bringing together teams of model builders each analyzing its own model. In the following, we present three exercises that are easy to carry out and serve to showcase the potential usefulness of the MMB technology to economists in academia and at policy institutions.

The first exercise shows how researchers can use it to evaluate the sensitivity of policy implications to key model parameters. Specifically, it reviews the importance of Keynesian consumers and monetary policy accommodation for fiscal stimulus effects in one of the models participating in the Coenen et al. (2012) comparison study. The second exercise extends the study of Taylor and Wieland (2012) on monetary policy transmission across earlier and more recent generations of structural macro models by including the medium-size New Keynesian model with financial frictions and risk shocks that Christiano et al. (2014) have recently estimated for the U.S. economy. Finally, the third exercise shows how to conduct cross-country comparisons and illustrates the use of model-specific rules in order to measure model uncertainty about policy effects.

5.1 Effects of Fiscal Stimulus: Sensitivity to Structural Parameters

The large-scale comparison study of Coenen et al. (2012) has highlighted the importance of monetary policy accommodation for Keynesian fiscal multiplier effects (see Section 2.2). Furthermore, the models participating in this study differed in terms of a relevant structural feature in this regard, namely the relative importance of Keynesian consumers that make decisions based on current income and Friedman–Modigliani permanent-income consumers that make forward-looking decisions based on lifetime income. Here, we show how MMB users can evaluate the sensitivity of fiscal policy effects to the parameters governing household consumption choices and central bank reactions. To this end, we consider one of the models participating in the Coenen et al. (2012) study: the US_CCTW10 model of Cogan et al. (2010).

In terms of fiscal shock, we look at the effects of a surprise increase in government purchases that fades out gradually according to an autoregressive process. The shock is implemented as a common policy shock in MMB, that is, an element of the common shock vector η_t introduced in Section 3.[d] As a consequence, government purchases increase on impact by 1% of GDP and then return slowly toward the original level.

5.1.1 Parameter Sensitivity Analysis: Share of Rule-of-Thumb Consumers

The US_CCTW10 model extends the Smets and Wouters (2007) model with Keynesian-style rule-of-thumb households. These households simply consume all current disposable income. Using the same data as Smets and Wouters (2007), Cogan et al. (2010) estimate the share of Keynesian rule-of-thumb consumers in the population jointly with the other structural parameters of the model. For the Bayesian estimation, the prior mean is assumed to be 50%. The resulting posterior mean is 27% with a standard deviation of 6%. Meanwhile, the other models used in Coenen et al. (2012) calibrate or estimate the population share of financially constrained households to values between 20% and 50%.

Fig. 7 reports on the effects of the fiscal policy shock for three different values of the share of rule-of-thumb consumers that is denoted by ω in the US_CCTW10 model. There are six panels displaying simulation outcomes for GDP, inflation, the nominal interest rate, consumption, investment, and government purchases. Each panel contains three lines indicating outcomes with a share of rule-of-thumb households of 0% ($\omega = 0$), 26.5% ($\omega = 0.265$, US_CCTW10), and 50% ($\omega = 0.5$).[e] For each simulation, the other parameters are kept unchanged at the posterior means estimated by Cogan et al. (2010). Noteworthy, setting $\omega = 0$ or $\omega = 0.5$ implies a deviation from the point estimate delivering the optimal fit of the model to the data. The simulation outcomes are best understood as a sensitivity exercise with respect to the single parameter ω.

With $\omega = 0$ there are no rule-of-thumb households. All consumers are forward-looking and base their decision on expected lifetime income as in Smets and Wouters (2007). By contrast, the value of 50% can be considered an upper limit of estimates for the share of rule-of-thumb consumers found in the literature on the U.S. economy. In all three simulations, government spending increases on impact by 1% of GDP and gradually returns to the steady-state ratio of government spending to output (lower right panel).

[d] At this point, the autoregressive parameter remains model-specific as an element of the parameter vector β_m. Yet, in other exercises we show how to consider common autoregressive parameters.

[e] Technically, users can easily change this structural parameter by editing the model file *US_CCTW10.mod* in the subdirectory */MODELS/US_CCTW10* of the archive. The parameter is found under "*// fixed parameters*" and denoted by "*omega = 0.2651; // share of rule-of-thumb consumers*" as in the published article. Then, the user simply needs to run the fiscal shock simulation in the menu "*One model, many policy rules*," repeating it every time he has edited the model file. Three sets of results can be saved in Excel files and then displayed in graphs.

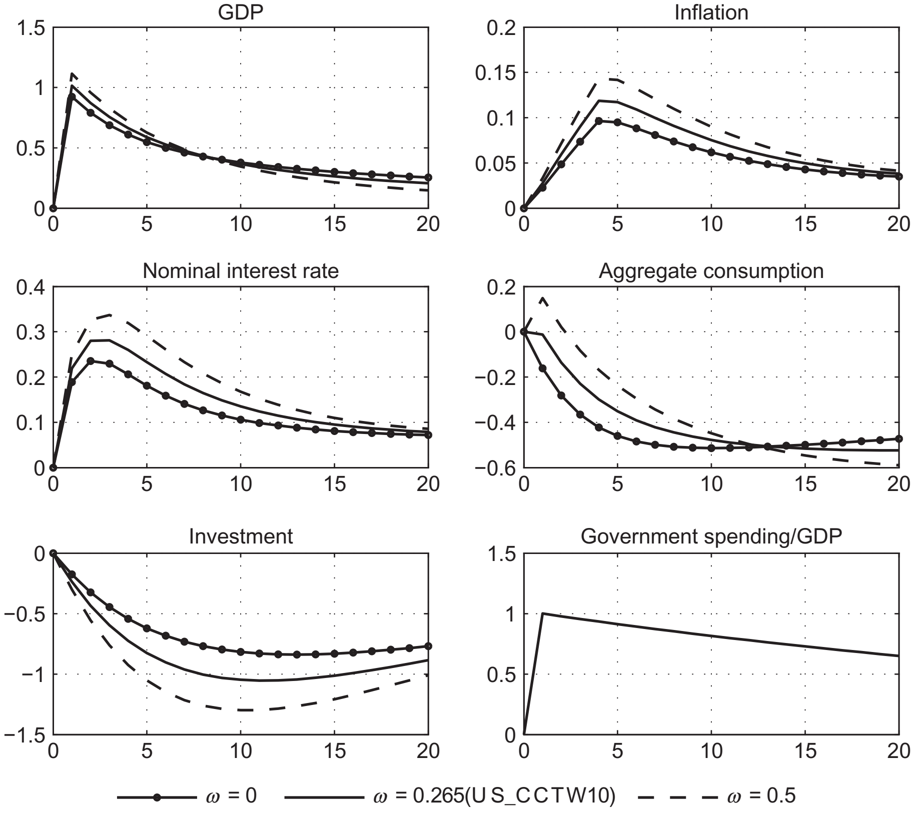


Fig. 7 Impulse responses to an expansionary fiscal policy shock in the US_CCTW10 model for alternative shares of the rule-of-thumb households. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values except government spending. Government spending is expressed as a share of GDP in percentage point deviations from the respective steady-state ratio. Inflation is the rate of inflation over the previous four quarters. The nominal interest rate is annualized. Other variables are expressed in quarterly terms. ω refers to the population share of rule-of-thumb households. The value of $\omega = 0.265$ (US_CCTW10) corresponds to the posterior mean estimate of Cogan et al. (2010). The simulation is carried out under the monetary policy rule of Cogan et al. (2010).

The effect on GDP increases with the population share of rule-of-thumb consumers. Yet, the quantitative differences in the GDP impact of the fiscal shock are not very large. The reason is that a crowding-out effect becomes more pronounced with a larger share of rule-of-thumb consumers. On the one side, aggregate consumption is higher with higher values of ω. With 50% of rule-of-thumb consumers, aggregate consumption even increases a bit in the first quarter consistent with the Keynesian multiplier effect. However, in response to higher GDP and higher inflation, the central bank raises the nominal interest rate. As prices adjust sluggishly due to nominal rigidities, the real interest rate (not shown) rises as well and by a larger amount for higher values of ω. Higher real rates reduce demand for investment purposes and incentivize forward-looking households to postpone consumption. Thus, the expansion in government spending crowds out private spending on investment and consumption. The model with rule-of-thumb consumers also accounts for the dynamics of government debt and taxes. First, government debt increases, then lump-sum taxes respond so as to return debt to the initial debt-to-GDP ratio. While rule-of-thumb consumers ignore the reduction of future disposable income, forward-looking consumers respond by reducing current consumption.

5.1.2 Parameter Sensitivity Analysis: Central Bank Reaction Function

Next, the effect of monetary accommodation is easily evaluated by changing the response coefficients in the monetary policy rule. This can either be accomplished by picking different preset rules under the *One model, many policy rules* menu or by entering different coefficients in the submenu for the *User-specified rule*. We compare the outcomes under the model-specific estimated rule from Cogan et al. (2010) (CCTW10 rule) with the model-specific rule from Bernanke et al. (1999) (BGG99 rule). The latter rule will also be used in Section 6 when we compare the small New Keynesian model of Bernanke et al. (1999) (NK_BGG99 model) to more recent macro-financial models. Here, the BGG99 rule is of interest because it responds only to lagged values for inflation and the interest rate and does not react to GDP (see Eq. (12)).[f] Thus, it should be much more accommodative than the CCTW10 rule.

Fig. 8 presents the implications of the government spending shock for inflation, nominal interest rate and output under the two different policy rules. The panels in the right column report selected results from the previous exercise with the CCTW10 rule (compare Fig. 7). Again, we consider the same three values for ω. The panels in the left column display the outcomes simulated under the BGG99 rule.

The increase in government purchases induces much stronger effects on aggregate GDP under the BGG99 rule. Even in the absence of rule-of-thumb consumers ($\omega = 0$), the GDP impact exceeds unity in the first four quarters. The much more accommodative monetary policy regime exhibited by the BGG99 rule allows for a Keynesian multiplier effect. Private consumption rises due to higher government consumption. This crowding-in effect outweighs the negative wealth effect coming from higher anticipated future taxes. The comparison emphasizes the importance of fiscal-monetary interactions for the effects of discretionary fiscal policy.

[f] This rule can be specified using the *User-specified rule* tab in the panel *Monetary Policy Rules* in the MMB graphical user interface. More specifically, the user needs to assign '0.9' for the entry for *interest (t-1)* and '0.11' for the entry for *inflationq (t-1)*.

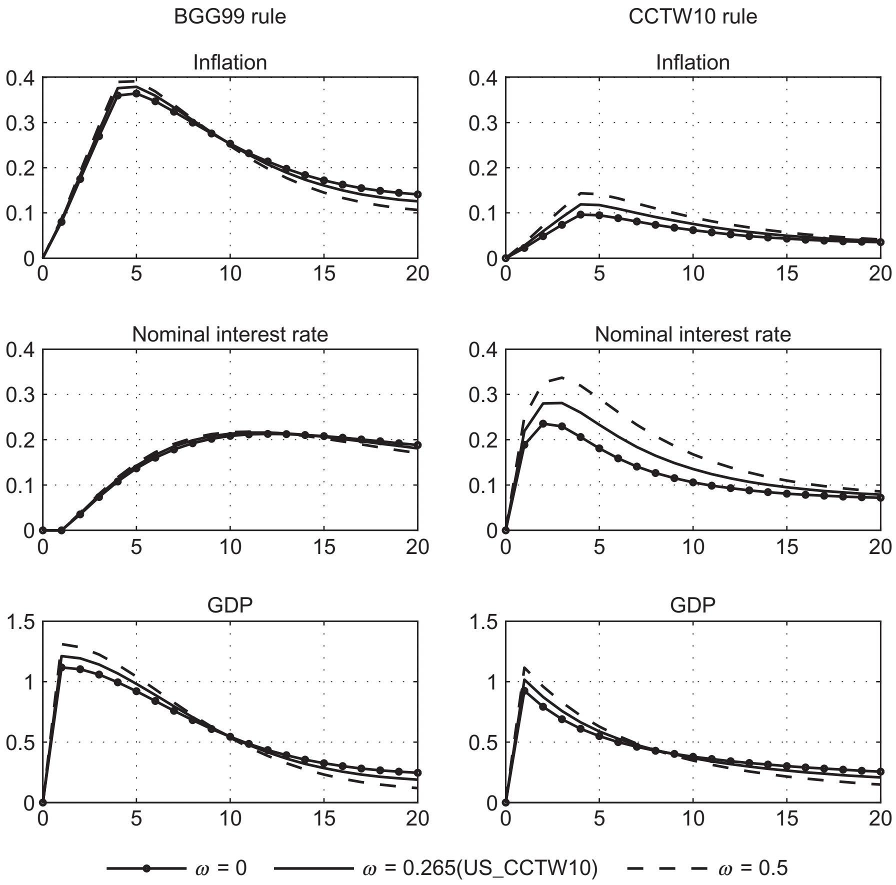


Fig. 8 Impulse responses to an expansionary fiscal policy shock in the US_CCTW10 model with alternative monetary policy rules. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Inflation is the rate of inflation over the previous four quarters. Nominal interest rate is annualized. GDP is expressed in quarterly terms. BGG99 rule refers to the monetary policy rule of the NK_BGG99 model by Bernanke et al. (1999). CCTW10 rule refers to the estimated rule of the US_CCTW10 model by Cogan et al. (2010).

5.2 Monetary Policy Transmission: Comparing Generations of Models

The Macroeconomic Model Data Base serves as an archive of models and contains models developed at different times and based on different theories about how the economy functions. Thus, it offers the possibility to compare different generations of models and their policy implications. One would expect that policy implications change substantially over time, either because new theories offer new insights in macroeconomic interdependencies or because new estimation methods and new data induce different estimates of key parameters. But there may also be surprising similarities.

Table 5 Three model generations

Notation	Description
G7_TAY93	Taylor (1993b): 1st generation New Keynesian model with rational expectations, wage, and price rigidities
US_ACELm	Christiano et al. (2005): 2nd generation New Keynesian medium-size Dynamic Stochastic General Equilibrium model
US_SW07	Smets and Wouters (2007): 2nd generation NK-DSGE model
US_DG08	De Graeve (2008): 3rd generation NK-DSGE model with financial frictions
US_CMR14	Christiano et al. (2014): 3rd generation NK-DSGE model, financial frictions

For example, Taylor and Wieland (2012) compare four different models of the U.S. economy that were developed and estimated at different times with different data, and find very similar estimates of the transmission of a monetary shock to GDP. This holds at least when a common central bank reaction function is used. Here, we extend this comparison to a fifth model that was estimated very recently. The models are listed in Table 5. MMB users can easily replicate and extend this comparison further with the MMB graphical user interface *One policy rule, many models*.

5.2.1 The Models

The G7_TAY93 model, which is a multicountry model of the G7 economies built more than 20 years ago, has been used extensively in the model comparison projects of the late 1980s and 1990s (see Section 2.1). It has New Keynesian properties such as nominal wage and price rigidities, rational expectations, and policy rules. However, it does not yet incorporate the complete set of microeconomic foundations developed in the real and monetary business cycle literature. We refer to it as a first-generation New Keynesian model.

US_ACELm[g] and US_SW07 are the best-known representatives of the second generation of empirically estimated New Keynesian models with additional microeconomic foundations, often referred to as New Keynesian DSGE models. Although they differ from G7_TAY93 also in terms of the estimation approach, data, and sample span, they exhibit almost identical GDP effects of an unexpected change in the federal funds rate. Following the global financial crisis, New Keynesian DSGE models have been fitted out with more detailed financial sectors and financial frictions that serve to amplify financial and economic shocks. Taylor and Wieland (2012) showed that one of these third-generation New Keynesian models, the US_DG08 model, also indicated similar monetary policy effects.

Here, we extend the comparison exercise by bringing one more model with financial frictions into the picture. Noteworthy, the US_CMR14 model is the only one among

[g] The impulse response functions for the monetary policy shock in Altig et al. (2005) are almost identical to those of Christiano et al. (2005). Altig et al. (2005), however, incorporate two additional shocks (a neutral and investment-specific technology shock). The Macroeconomic Model Data Base thus includes the model of Altig et al. (2005).

these five models estimated on data that covers the Great Recession (the sample spans 1985: Q1–2010:Q2) and includes financial time series such as credit to nonfinancial firms, the slope of the term structure, credit spreads on corporate bonds, and an index of stock prices.

> ***US_CMR14 Model Description***: *Christiano et al. (2014) introduce the financial accelerator mechanism of Bernanke et al. (1999) into an otherwise standard New Keynesian model, such as the model of Christiano et al. (2005). This mechanism is described in more detail in Section 6. In contrast to earlier models with financial frictions (see, e.g., Christensen and Dib, 2008; De Graeve, 2008), the authors introduce a shock to the variance of idiosyncratic productivity that influences individual entrepreneur's return to capital. It is referred to as a risk shock. With an agency problem between entrepreneurs and banks, a positive risk shock increases the required return on borrowing, that is, the external finance premium. As a consequence, entrepreneurs' borrowing is reduced and investment declines. As capital prices fall, entrepreneurial net worth decreases, which in turn raises the external finance premium further. These amplification effects are propagated to the real economy over time. Importantly, the authors' analysis suggests that the risk shock is a major driving force of the U.S. business cycle.*

5.2.2 Strikingly Similar Impulse Responses to a Monetary Policy Shock

Fig. 9 displays the effects of a one-percentage point unexpected increase in the federal funds rate on output, inflation, and the interest rate itself in all five models under two alternative monetary policy rules. The panels on the left side show the outcomes when the interest rate is set according to the monetary policy rule estimated in Smets and Wouters (2007) (SW rule), while the panels on the right side refer to the outcomes under the monetary policy rule estimated in Christiano et al. (2014) (CMR rule). The SW rule[h] and the CMR rule are given by Eqs. (10) and (11), respectively.

$$i_t^z = 0.81 i_{t-1}^z + 0.39 p_t^z + 0.97 q_t^z - 0.90 q_{t-1}^z + \eta_t^i. \tag{10}$$

$$i_t^z = 0.85 i_{t-1}^z + 0.36 p_t^z + 0.05 y_t^z - 0.05 y_{t-1}^z + \eta_t^i. \tag{11}$$

The superscript z refers to common variables, that are defined consistently and therefore allow quantitative comparisons. The monetary policy instrument is the annualized short-term federal funds rate in quarter t denoted by i_t^z. p_t^z refers to the annualized quarter-to-quarter rate of inflation, y_t^z is the deviation of quarterly real GDP from its long-run potential, while q_t^z refers to the output gap defined as the difference between actual GDP and the level of GDP that would be realized if prices and wages were flexible. All variables are expressed in percentage deviations from steady-state values. η_t^i refers to the common monetary policy shock.

Under the SW rule, US_SW07, US_ACELm, and G7_TAY93 indicate almost identical GDP responses and quite similar inflation responses to the interest rate shock. GDP declines by 25–30 basis points within 3–4 quarters and then returns to its steady-state level

[h] The monetary policy shock in the estimation of Smets and Wouters (2007) exhibits weak serial correlation with a correlation coefficient of 0.15. In MMB the policy shocks are iid.

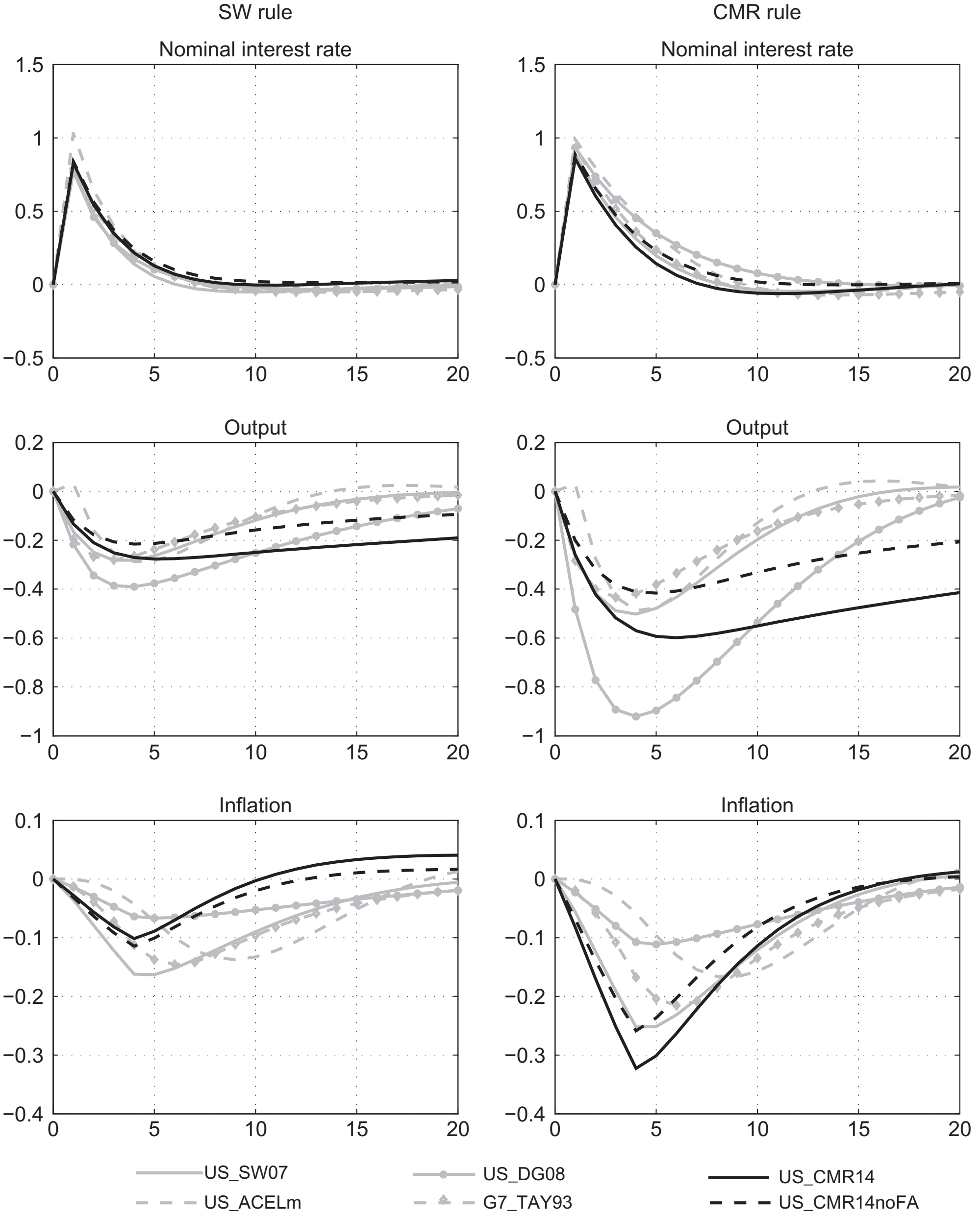


Fig. 9 Impulse responses to a contractionary monetary policy shock in selected models of the U.S. economy under alternative policy rules. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Inflation is the rate of inflation over the previous four quarters. Nominal interest rate is annualized. Output is expressed in quarterly terms. SW rule and CMR rule refer to the monetary policy rules estimated in Smets and Wouters (2007) and Christiano et al. (2014), respectively. US_SW07 is the model of Smets and Wouters (2007), US_ACELm replicates the model of Christiano et al. (2005); G7_TAY93 is the model of Taylor (1993b), US_CMR14 is the model of Christiano et al. (2014); and US_CMR14noFA is the modified version of US_CMR14 without the financial friction.

again. The effects are only slightly larger in the US_DG08 model that was also considered by Taylor and Wieland (2012).

Interestingly, the maximum GDP effect in the US_CMR14 model is again of the same magnitude, about 30 basis points within four quarters under the SW rule. Yet, it is much more persistent. GDP returns to steady state very slowly. It barely moves back over the first 20 quarters. It seems monetary policy has become more powerful in terms of inducing lasting consequences for the real side of the economy. Clearly, this finding requires further study.

When using the CMR rule, we obtain greater effects of the policy shock on GDP and inflation in all five models. The reason is that the CMR rule is more accommodative. Its reaction coefficients concerning real GDP are smaller. Yet again, US_SW07, US_ACELm, and G7_TAY93 imply very similar GDP effects, on the scale of a reduction of 45 to 50 basis points within three to four quarters. With this rule, the differences in the third generation of New Keynesian models with financial frictions come out more clearly. In US_DG08, the impact on GDP is quite a bit stronger reaching −90 basis points, while it is again much longer lasting in US_CMR14.

As a further check on the source of the stronger, more lasting effect of monetary policy on real GDP, we modify the US_CMR14 in order to shut down the financial accelerator mechanism. We label this modified version of the model US_CMR14noFA. All other parameters are kept at the values in the original model specification. Noteworthy, the US_CMR14noFA model is structurally very close to the US_ACELm model, the main difference being the presence of the cost channel in US_ACELm. We find that the GDP response is less pronounced in the version without the financial accelerator, yet it remains somewhat more persistent than in the other models.

5.2.3 Unusually Persistent Real Effects of Monetary Shocks in the Model of Christiano et al. (2014)

To investigate the possible origin of the unusually long-lasting real effects of monetary policy in the US_CMR14 model, we repeat the same exercise with four different versions of the model, and show impulse responses over 40 quarters after the shock. In doing so, we always use the model-specific rule, ie, the CMR rule. In addition to US_CMR14noFA, which shuts down the financial accelerator, we consider a version that shuts down nominal wage rigidities (US_CMR14noNW), and a version without both, wage rigidities and financial frictions (US_CMR14noFA&NW). Fig. 10 presents the resulting impulse responses.

The persistent response of GDP in the baseline model (US_CMR14) is reflected in both investment and consumption. Investment falls for eight quarters and then returns very slowly to the steady state. Consumption falls and stays far below the steady state for about thirty quarters and then starts returning to the steady state. Such a long-lasting effect of a monetary shock on household consumption in real terms appears rather unrealistic. In particular, as inflation, nominal and real interest rates return to steady state in ten quarters.

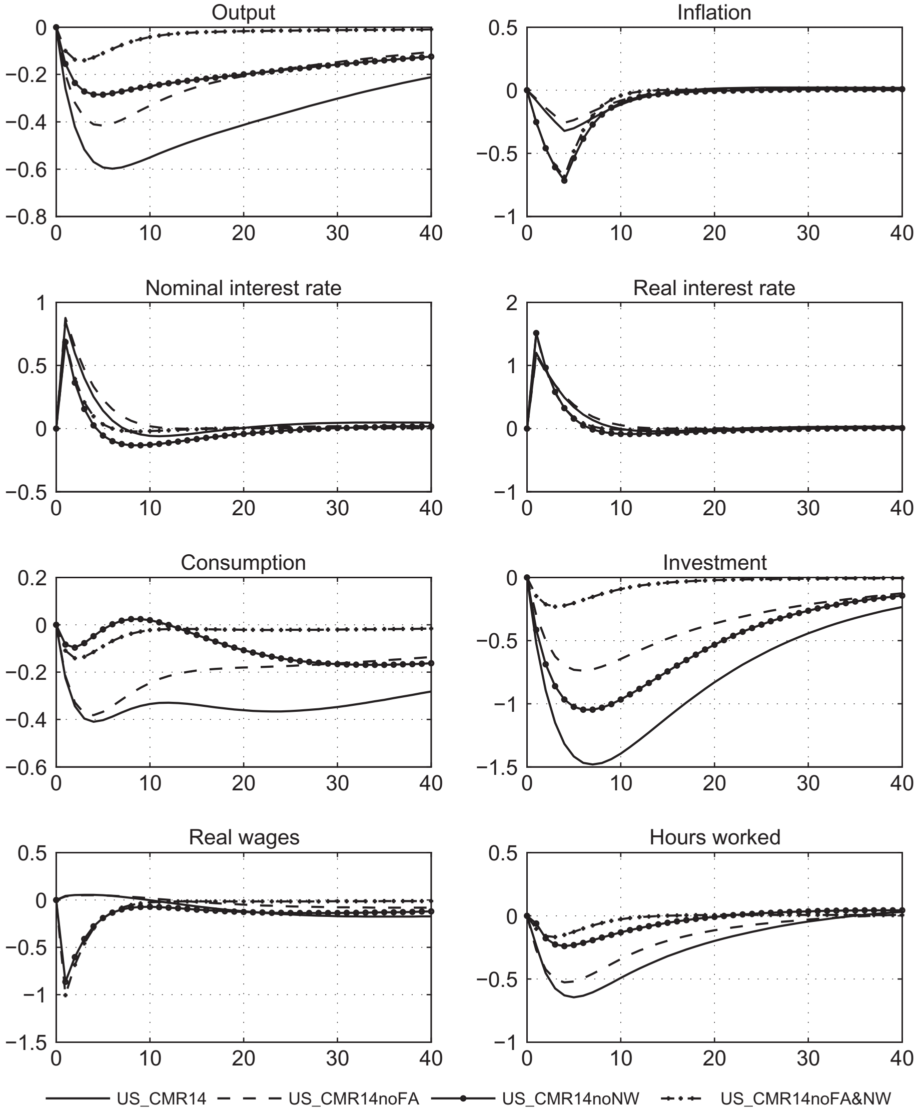


Fig. 10 Impulse responses to a contractionary monetary policy shock in modified versions of the Christiano et al. (2014) model with CMR Rule. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Inflation is the rate of inflation over the previous four quarters. Nominal interest rate is annualized. Other variables are expressed in quarterly terms. US_CMR14 is the model of Christiano et al. (2014); US_CMR14noFA is the version of US_CMR14 without the financial friction; US_CMR14noNW is the version of US_CMR14 without nominal wage rigidities; and US_CMR14noFA&NW is the version of US_CMR14 without both the financial friction and nominal wage rigidities.

While the magnitude of the effect is reduced in the model without the financial accelerator (US_CMR14noFA), consumption, and investment remain highly persistent. In the model without nominal wage rigidities (US_CMR14noNW), real wages fall sharply, while hours worked decline somewhat. The reason is that intermediate good producing firms respond to the contractionary policy shock by adjusting prices rather than quantities. It leads to smaller effects on output and larger effects on inflation. However, the dynamics of output, consumption, and investment are still very persistent. For instance, consumption first returns to the steady state ten quarters after the shock but then continues to decrease even more than during the initial ten quarters.

Finally, in the model without wage rigidities and financial friction (US_CMR14noFA&NW), the real effects of a policy shock are much reduced. The magnitude of the maximum output effect is about one-third of that in the baseline model. More importantly, the strikingly persistent dynamic responses of real variables disappear. Output, investment, consumption, and hours worked return close to steady state in 10 to 15 quarters.

Our comparative exercise shows that the US_CMR14 model implies highly persistent output effects of monetary policy shocks relative to other estimated models of the U.S. economy. Given that the policy debate after the global financial crisis and the Great Recession has been dominated by the fear that monetary policy has become less effective, this result is surprising. The comparisons with modified versions of the model suggests that this feature of the US_CMR14 model is rooted in the parameter estimates that govern the importance of wage rigidities and of the financial accelerator. Yet, it would be important to explore further whether this effect depends on unusual combinations of parameter estimates and whether the extreme persistence disappears if the model is estimated over part of the data sample.

5.3 Predicted Effects of Identified Policy Shocks: United States vs Euro Area

5.3.1 When a Model Comparison Should Make Use of Model-Specific Policy Rules

In the preceding exercise, we have considered the consequences of monetary policy shock across models when the central bank in each model applies the same common policy rule. The idea of these simulations is to examine model differences stemming from model structure, while eliminating the differences stemming from model-specific monetary policy rules. They correspond exactly to the approach laid out in Section 3. Assuming a common rule serves its purpose in making a clean comparison of the policy implications of different model structures.

However, there are other questions that can be answered with model comparisons that employ model-specific rules. For example, if one wants to compare the forecasting performance of different models, the model should be used as fitted to the data. Using a different policy rule would reduce the fit of the model to the data it was estimated on, and its forecasting performance would presumably deteriorate. Thus, for such a comparison each model should be used with the model-specific policy rules that were estimated along with the rest of the model.

Hence any question that involves comparing model fit would make use of model-specific rules. The question to be considered here concerns the empirical degree of model uncertainty about the consequences of identified monetary policy shocks in the United States vs the euro area. Specifically, we aim to assess the range of predicted effects across models. Conditional on the structural assumptions of a model and the sample the model was estimated on, the impulse response under the model-specific rule represents the most likely data-driven reaction of the economy to the monetary policy shock. The comparison exercise then provides a measure of the degree of model uncertainty about monetary policy transmission.

5.3.2 Models with Different Structural Features Estimated with United States and Euro Area Data

Specifically, we choose models from the MMB archive for which all equations are jointly estimated and the model-specific monetary policy rule is formulated for the nominal short-term interest rate. This selection includes twelve U.S. models and eight euro area models (see Table 6). Although all models share certain New Keynesian features, there is a lot of heterogeneity in terms of structural assumptions, observables, and estimation techniques.[i]

All the U.S. models and most euro area models are closed economies. Exceptions include EA_SR07 and EA_QUEST3, which are small open economies, and the two-country models, EAES_RA09 and EA_QR14. Most models only consider forward-looking permanent-income households. EA_QUEST3 and US_CCTW10, however, also include rule-of-thumb households. Models with housing finance, such as US_IAC05, US_IN10, EA_GNSS10, and EA_QR14, feature two types of households that behave as borrowers and savers, respectively. The difference in decision making arises from differences in their discount factors. Savers are more patient than borrowers. Impatient agents face a borrowing constraint and use housing as collateral for borrowing.

Another financial friction that influences credit demand—the financial accelerator mechanism of Bernanke et al. (1999)—is incorporated in the US_CD08, US_DG08, US_CMR14, and EA_GE10 models. Frictions in credit supply are considered in the EA_GNSS10 model, which includes a more detailed banking sector. US_PM08fl, the IMF's small projection model for the U.S. economy, also includes a macro-financial linkage in form of a behavioral relation between bank lending conditions and the real

[i] A brief description of each model is included in the MMB software package and can be downloaded from the MMB website.

Table 6 Estimated models used in the comparison across economies

Estimated U.S. models		Estimated euro area models	
US_ACELm	Christiano et al. (2005)	EA_SW03	Smets and Wouters (2003)
US_IAC05	Iacoviello (2005)	EA_SR07	Adolfson et al. (2007)
US_MR07	Mankiw and Reis (2007)	EA_QUEST3	Ratto et al. (2009)
US_RA07	Rabanal (2007)	EAES_RA09	Rabanal (2009)
US_SW07	Smets and Wouters (2007)	EA_CKL09	Christoffel et al. (2009)
US_CD08	Christensen and Dib (2008)	EA_GE10	Gelain (2010)
US_DG08	De Graeve (2008)	EA_GNSS10	Gerali et al. (2010)
US_PM08fl	Carabenciov et al. (2008)	EA_QR14	Quint and Rabanal (2014)
US_IN10	Iacoviello and Neri (2010)		
US_CCTW10	Cogan et al. (2010)		
US_IR11	Ireland (2011)		
US_CMR14	Christiano et al. (2014)		

Note: The first and third columns contain the model name in the MMB for the respective paper.

economy. The propagation mechanisms generated by financial frictions are to be studied more thoroughly in Section 6.

The models considered in this exercise also incorporate different labor market structures. Some models (US_IAC05, US_CD08, US_IR11, EA_QR14) assume competitive labor markets, but a majority of the models accounts for monopolistic competition in labor supply and Calvo-style rigidity in nominal wages. EA_CKL09 additionally introduce Mortensen and Pissarides (1994) type of matching frictions in the labor market.

Furthermore, US_MR07 differs from all other models due to the assumption of sticky information. In this model, only a fraction of agents (consumers, workers, and firms) updates their information regularly when making decisions. The other agents are inattentive. This feature gives rise to sluggish macroeconomic adjustment.

With regard to model-specific interest rate rules, most models feature interest rate smoothing as well as a reaction to inflation and a real variable (typically, the output gap or output growth). The exceptions are US_CD08 and EA_SR07, where the monetary policy rule also includes reactions to money growth and the real exchange rate, respectively.

Finally, there are also important differences in terms of the time series employed in estimating the models. At a minimum, these include real GDP, inflation, and the short-term nominal interest rate. Most of the models, however, are estimated on a larger set of observables. For example, Smets and Wouters (2007), De Graeve (2008), Smets and Wouters (2003), and Gelain (2010) use seven macroeconomic time series: real GDP, inflation, consumption, investment, real wages, employment, and the short-term nominal interest rate. Adolfson et al. (2007) employ fifteen macroeconomic time series to estimate the euro area model of Sveriges Riksbank (EA_SR07). Iacoviello and Neri (2010) use ten observables, including measures of housing construction and prices. In terms of sample period, the U.S. models are typically estimated on longer samples than the euro area models. Most models are estimated with Bayesian techniques. However, US_CD08 and US_IR11 are estimated with maximum likelihood techniques, while US_ACELm and US_IAC05 are estimated by minimizing the distance between VAR-based and model-implied impulse responses.

5.3.3 Dynamic Responses of Output, Inflation, and Interest Rates: United States vs Euro Area

Fig. 11 reports the outcomes for a one-percentage point contractionary shock to the nominal interest rate under model-specific rules.[j] The panels in the left column display the results for twelve estimated models of the U.S. economy, while the panels in the right column show the results for eight euro area models.

In every case, the unexpected increase in the nominal interest rate leads to a decline in output and inflation. Due to sticky prices the real interest rate rises, which depresses aggregate demand. Lower demand curbs production. As a fraction of price setters adjust to lower demand, inflation falls.

At first glance, there appears to be considerable variation in the magnitude and dynamic patterns of effects. Yet, this impression results from a few outliers. Outliers with regard to output are US_IAC05, US_RA07 and EA_SW03, while US_MR07, US_RA07 and EA_SW03 are outliers with regard to inflation dynamics. Except for US_IAC05, the strong reactions to the policy shock are largely due to a coefficient near unity on the lagged interest rate in the policy rule.[k] The anticipation of a longer period of higher interest rates induces a larger and longer lasting effect on output and inflation, because households and firms take into account expectations of future interest rates in their decision making. In the case of US_IAC05, the lack of important real rigidities, such as habit formation in consumption and investment adjustment cost, coupled with the presence of collateral constraints gives rise to a large initial impact of the monetary policy shock on output.

5.3.4 A Few Summary Statistics

Table 7 provides some summary statistics. In the U.S. models, the trough of output following a contractionary policy shock is reached within one to six quarters, and on average

[j] We obtain simulation results in two ways. When a model-specific policy rule is nested in the generalized rule in MMB, it is available for simulation for the model in question using the options menu *One model, many policy rules*. If this is not the case, we use the replication files for the original models that are provided together with the MMB comparison software.

[k] The coefficients on the lagged interest rate in the policy rule of the models with the strongest responses are as follows: 0.94 for US_RA07, 0.92 for US_MR07, 0.96 for EA_SW03. Noteworthy, the model-specific policy rule of Mankiw and Reis (2007) does not explicitly include a lagged interest rate but the policy shock is modeled as an AR(1)-process with the persistence coefficient of 0.92.

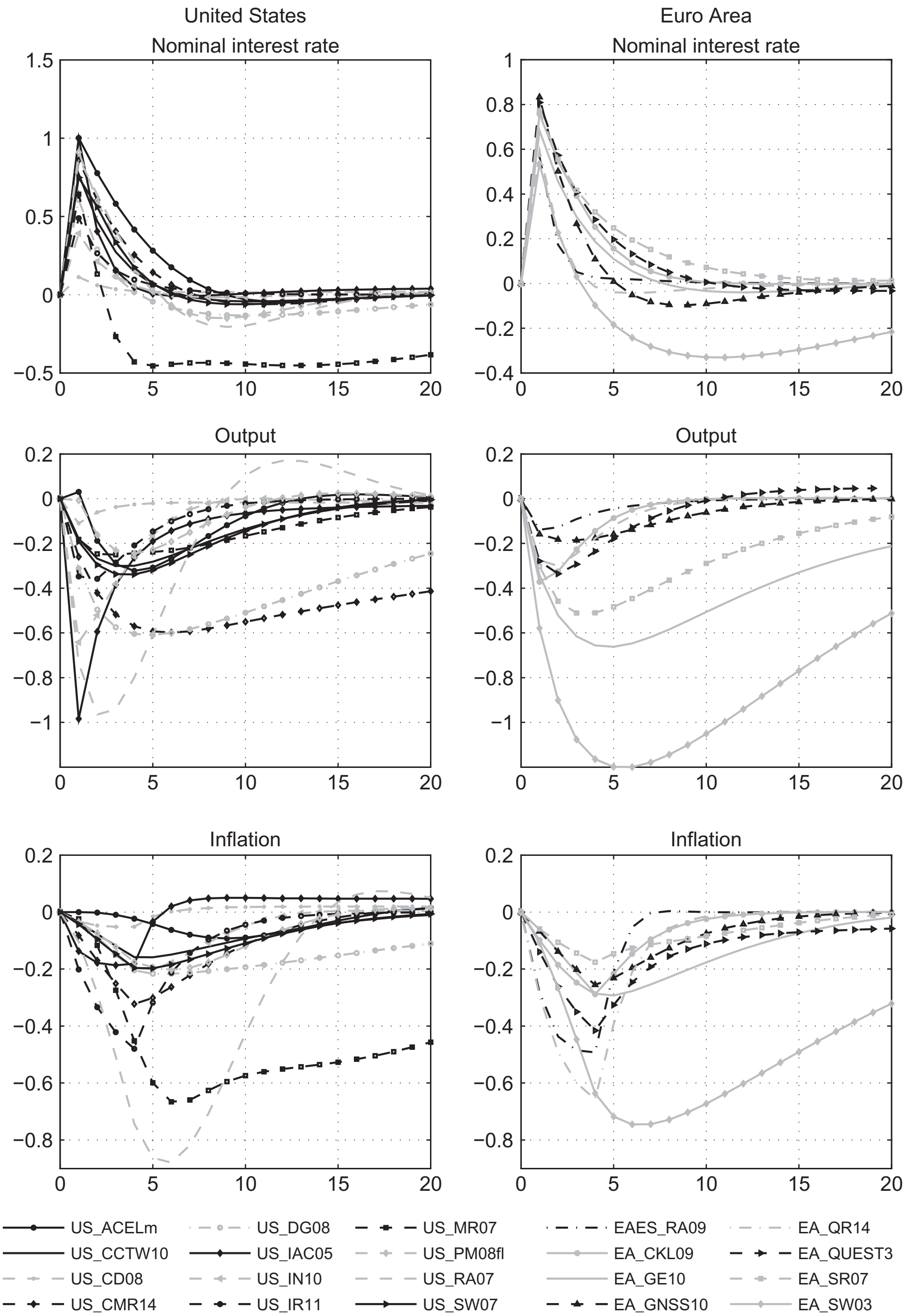


Fig. 11 Impulse responses to a contractionary monetary policy shock in various models with model-specific rules. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Nominal interest rate is annualized. Inflation is the rate of inflation over the previous four quarters. Output is expressed in quarterly terms.

Table 7 Effects of a one percentage point unexpected increase in the policy rate on output and inflation in the United States models and the Euro area models

	Output		Inflation	
	Timing	**Magnitude**	**Timing**	**Magnitude**
(a) Estimated U.S. models				
US_ACELm	4	−0.32%	9	−0.09%
US_IAC05	1	−0.98%	3	−0.19%
US_MR07	3	−0.25%	6	−0.67%
US_RA07	2	−0.96%	6	−0.88%
US_SW07	4	−0.34%	5	−0.20%
US_CD08	1	−0.11%	3	−0.05%
US_DG08	5	−0.61%	6	−0.22%
US_PM08fl	4	−0.25%	6	−0.20%
US_IN10	1	−0.64%	5	−0.20%
US_CCTW10	3	−0.30%	5	−0.16%
US_IR11	2	−0.36%	4	−0.48%
US_CMR14	6	−0.60%	4	−0.32%
Model averages	3.0	−0.48%	5.2	−0.30%
Standard deviations	1.7	0.28%p	1.6	0.25%p
(b) Estimated euro area models				
EA_SW03	6	−1.20%	6	−0.75%
EA_SR07	3	−0.51%	4	−0.18%
EA_QUEST3	2	−0.34%	4	−0.42%
EAES_RA09	1	−0.14%	4	−0.49%
EA_CKL09	1	−0.37%	4	−0.29%
EA_GE10	5	−0.66%	5	−0.29%
EA_GNSS10	3	−0.19%	4	−0.26%
EA_QR14	2	−0.30%	4	−0.66%
Model averages	2.9	−0.46%	4.4	−0.42%
Standard deviations	1.8	0.34%p	0.7	0.20%p

Note: Timing refers to the quarter after the shock, when the trough or the deepest point in the response of the respective variable is reached.

in the third quarter. The average magnitude of the drop in output is 0.48% with a standard deviation of 0.28%p. Interestingly, the timing of the trough and the magnitude of the output drop in the euro area models is very similar. In the euro area, output also reaches the trough within 1–6 quarters, on average in the third quarter. The average output decline at the trough corresponds to −0.46%. Thus, it is very close to its U.S. counterpart, albeit the standard deviation of 0.34%p is a bit larger.

As for inflation, the U.S. models imply that the deepest point in the inflation response occurs within 3–9 quarters, on average, in the fifth quarter. In the euro area models, the span of this range is more narrow at 4–6 quarters, with an average of 4.4 quarters. The average decline in inflation at the trough corresponds to −0.30% for the U.S. models and −0.42% for the euro area models. The respective standard deviations are very similar: 0.25%p for the U.S. and 0.20%p for the euro area.

Thus, the above comparison exercise serves to show that model averages of the predicted impact of identified policy shocks on output and inflation are very similar for the United States and the euro area in terms of timing and magnitude of the resulting contraction.

6. COMPARING IMPLICATIONS OF NEW MACRO-FINANCIAL MODELS

6.1 Key Characteristics: Investment Finance, Housing Finance, and Banking Capital

The global financial crisis has drawn attention to the need for improving the characterization of the financial sector in macroeconomic models used for business cycle and policy analysis. Many new contributions have included financial market imperfections in New Keynesian DSGE models, in particular in three areas: the financing of new investment in firms' capital for production purposes, the financing of housing investment, and the role of banks and bank capital in financial intermediation. These financial frictions help explain how the consequences of economic shocks for macroeconomic aggregates can be amplified via the financial sector, and how financial sector stress and financial crises can spill over into the real economy.

6.1.1 Corporate Investment Financing and the Financial Accelerator

Fortunately, research on integrating financial frictions in macroeconomic models for policy analysis do not need to start from scratch. A prominent starting point is the so-called financial accelerator model of Bernanke et al. (1999) (BGG99). Here, the accelerator term refers to the amplification of economic fluctuations via the financial sector. Long before the global financial crisis, they already provided a tractable approach for including information asymmetries, which are central to the relationship between borrowers and lenders, in dynamic New Keynesian models.

Lending institutions and financial contracts aim to reduce the costs of collecting information and to mitigate principal-agent problems in credit markets. By contrast, economic shocks may increase the cost of extending credit and reduce the efficiency of matching borrowers and lenders. Hence, the credit market imperfections may amplify the effects of shocks from the financial sector as well as other sectors of the economy. BGG99 focus on the financing of investment in firms' capital for production purposes. Their model includes risk-averse households, risk-neutral entrepreneurs, and retailers. Entrepreneurs use capital and labor to produce wholesale goods. These are sold to the retailers. The retail market is characterized by monopolistic competition and price rigidities. Entrepreneurs borrow funds from households via a financial intermediary. These funds serve to pay for part of the new capital, which becomes productive in the next period. The agency problem arises because the return to capital is subject to idiosyncratic risk and can only be observed by the financial intermediary after paying some auditing cost. As a result, the entrepreneurs' net worth becomes a key factor determining their borrowing costs. Entrepreneurs with high net worth need less external funding for a given capital investment and pay lower premia. To the extent that net worth rises and falls with the business cycle, the premium to be paid for external borrowing varies counter-cyclically. Thus, it increases fluctuations in borrowing, investment, spending, and production.

A version of the BGG99 model is included in MMB. The implementation differs somewhat from the handbook article because it omits entrepreneurial consumption. Its short-hand reference in MMB is NK_BGG99. The model archive also contains several more recent contributions of empirically estimated models that extend the financial accelerator mechanism of BGG99. For example, Christensen and Dib (2008) (US_CD08) extend the dynamic New Keynesian model of Ireland (2003) (see US_IR04) with a financial accelerator and estimate the model on U.S. data. In their model, debt contracts are written in nominal terms in contrast to BGG99. De Graeve (2008) (US_DG08) includes the financial accelerator in the medium-scale New Keynesian model of Smets and Wouters (2007) (US_SW07) and estimates the extended model with Bayesian methods using U.S. data on the same nonfinancial macroeconomic time series as Smets and Wouters (2007). In addition, he documents a reasonably close match between the model-implied external finance premium and lower-grade corporate bond spreads. Similarly, Christiano et al. (2014) (US_CMR14) incorporate financial frictions à la BGG99 into the version of the model by Christiano et al. (2005) (US_ACEL). Unlike De Graeve (2008), they also employ financial data including the credit spread in the estimation. Furthermore, they allow the volatility of idiosyncratic productivity to vary over time. Table 8 summarizes the key features of the financial accelerator models relative to the comparison benchmark, US_SW07.

6.1.2 Housing Finance

Real estate booms and busts have played a central role in triggering the global financial crisis. These include not only the subprime mortgage boom and bust in the United States but also the credit-driven housing booms in a number of European countries such as Spain and Ireland. Thus, models with a more detailed housing sector that recognize the relevant financing constraints are of great interest to policy makers.

The underlying rationale of housing finance is the limited enforceability of debt contracts, as borrowers may choose to default. To overcome this limited commitment problem, lenders require collateral, typically housing and land, and provide funds only below the value of the collateral. Thus, the borrowing capacity, and hence the size of the loan is tied to the housing value. A starting point for modeling borrowing and lending under such a collateral constraint in macroeconomic models is to introduce an incentive for

Table 8 Comparison of key modeling features: Financial accelerator models and the US_SW07 benchmark

	US_SW07	NK_BGG99	US_CD08	US_DG08 / US_CMR14
Model structure				
Key agents	Representative household –	Representative household Risk-neutral entrepreneurs	Representative household Risk-neutral entrepreneurs	Representative household Risk-neutral entrepreneurs
Production sector	One-sector Cobb-Douglas (CD) technology	One-sector CD technology	One-sector CD technology	One-sector CD technology
Real and nominal rigidities				
Consumption habit formation	Yes	No	No	Yes
Expenditure adjustment cost	Investment adjustment cost	Capital adjustment cost	Capital adjustment cost	Investment adjustment cost
Capital utilization	Yes	No	No	Yes
Consumer prices	Calvo pricing, partial indexation	Calvo pricing, full indexation to steady-state inflation	Calvo pricing, full indexation to steady-state inflation	Calvo pricing, partial indexation
Nominal wages	Calvo pricing, partial indexation	Flexible	Flexible	Calvo pricing, partial indexation
Financial frictions				
Debt contract	–	Standard risky debt, real terms	Standard risky debt, nominal terms	Standard risky debt, real/nominal terms
Model Parameters				
Estimation/ Calibration	Bayesian estimation, U.S. data: 1966Q1–2004Q4	Calibration, U.S. data	ML estimation, U.S. data: 1979Q3–2004Q3	Bayesian estimation, U.S. data: 1954Q1–2004Q4 (US_DG08), 1985Q1–2010Q2 (US_CMR14)
Reference paper				
	Smets and Wouters (2007)	Bernanke et al. (1999)	Christensen and Dib (2008)	De Graeve (2008)/ Christiano et al. (2014)

economic agents to act as lenders or borrowers. Technically, it is assumed that the agents differ in their discount factors: some are more patient than others. In equilibrium, the more patient ones become savers while the impatient ones become borrowers.

The collateral constraint has the following consequences: suppose an aggregate shock shifts housing demand upwards such that house prices increase. As a result, borrowing capacity expands. On this basis, the impatient agents increase expenditure on nonhousing and housing goods, which in turn puts additional upward pressure on house prices. Thus, the effect of the initial shock is amplified over time due to the presence of the collateral constraint.

Kiyotaki and Moore (1997) develop a simple dynamic model with patient (and unproductive) entrepreneurs and impatient (and productive) entrepreneurs to show that the collateral channel can generate large and persistent business cycles. Iacoviello (2005) then incorporated such collateral constraints together with nominal debt in a dynamic New Keynesian model. In his model, impatient households and entrepreneurs face collateral constraints, when borrowing funds from patient households. Both household types obtain utility from housing services, while entrepreneurs use housing for the production of nonhousing (consumption) goods.[1] The model is estimated with U.S. data and referred to as US_IAC05 in the MMB model archive.

MMB includes two other U.S. models with housing finance. The model of Iacoviello and Neri (2010) (US_IN10) features a two-sector production structure with housing and nonhousing goods and imposes a collateral constraint only on impatient households. They consider various real and nominal rigidities similar to medium-scale New Keynesian models such as Christiano et al. (2005) and Smets and Wouters (2007). The US_IN10 model is estimated on U.S. macroeconomic and housing data. The model by Kannan et al. (2012) (NK_KRS12) is a simplified version of Iacoviello and Neri (2010). Key elements of the model are the presence of financial intermediaries and the determination of the spread between the lending rate and the deposit rate. The functional form for the determination of the spread is assumed rather than derived from a microfounded optimization problem. Financial intermediaries take deposits from patient households and lend to impatient households charging a spread that varies inversely with the net worth of borrowers. Thus, the financial accelerator mechanism operates in housing finance. Table 9 provides further information concerning key features of the three models with housing finance.

6.1.3 Financial Intermediation and Bank Capital

Banks' illiquidity, insolvency as well as counter-party risks played a prominent role during the global financial crisis, impairing credit supply by banks and thereby deepening the negative impact from excessive leverage of borrowers on the real economy. In contrast

[1] Aggregate housing supply is assumed to be fixed.

Table 9 Comparison of key modeling features: Models with housing

	US_IAC05	US_IN10	NK_KRS12
Model structure			
Key agents	Patient households Impatient households Impatient entrepreneurs	Patient households Impatient households	Patient households Impatient households
Production sector	One-sector CD technology Fixed housing supply	Two-sector CD technologies (Nonhousing & housing)	Two-sector constant returns to scale (CRS) technologies No capital
Real and nominal rigidities			
Consumption habit formation	No	Yes	Yes
Expenditure adjustment cost	Capital adjustment cost	No	Housing investment adjustment cost
Capital utilization	No	Yes	No
Consumer prices	Calvo pricing, no indexation	Calvo pricing, partial indexation	Calvo pricing, full indexation to past inflation
Nominal wages	Flexible	Calvo pricing, partial indexation	Flexible
House prices	Flexible	Flexible	Calvo pricing, full indexation to past housing price growth
Financial frictions			
Collateral constraints	Kiyotaki and Moore (1997) type, nominal terms Constant loan-to-value ratio	Kiyotaki and Moore (1997) type, nominal terms Constant loan-to-value ratio	In the spirit of Bernanke et al. (1999), real terms Variable loan-to-value ratio
Model Parameters			
Estimation/ Calibration	Estimation by minimizing a measure of the distance between the VAR impulse responses and model responses, U.S. data: 1974Q1–2003Q2	Bayesian estimation, U.S. data: 1965Q1–2006Q4	Calibration, U.S. data
Reference paper			
	Iacoviello (2005)	Iacoviello and Neri (2010)	Kannan et al. (2012)

with financial accelerator and housing sector models, which focus on frictions stemming from the demand side of financial intermediation, banking sector models deal with frictions on the supply side. In these models, the balance sheet and decision processes of banks are treated explicitly. Thus, shocks originating in the banking sector can have significant spillover effects on the macroeconomy and standard nonfinancial shocks can operate via new transmission channels, when macro-financial linkages are taken into account. In what follows, we focus on three quantitative monetary DSGE models in which banking capital plays a key role.

In the model of Gertler and Karadi (2011) (NK_GK11), banks obtain short-term funds from households and lend them to nonfinancial firms by purchasing the firms' long-term securities. There is no financial friction between banks and nonfinancial firms. Instead, the possibility that the banker can divert part of the bank' assets creates a moral hazard problem between the bank and households. In order to induce households to provide funds, the bank has to satisfy an incentive constraint: the pecuniary benefit from diverting funds must be at least as small as the gain from staying in business. This condition serves as an endogenous constraint on the bank's leverage. Such financial intermediaries are imbedded into an otherwise standard medium-scale New Keynesian model such as Christiano et al. (2005).

Meh and Moran (2010) (NK_MM10) use the double moral hazard framework of Holmstrom and Tirole (1997) and introduce banking decisions via an optimal financial contract. The first moral hazard problem is between a representative household and a representative bank. As the bank's monitoring technology is not directly observed by the investor, the latter requires the bank to participate in the project with its own net worth to mitigate this information asymmetry. Therefore, the ability of the bank to attract loanable funds depends on its capital position. The second moral hazard problem is between the bank and the entrepreneur, because entrepreneurial effort is private information. The bank requires entrepreneurs to participate financially, ie, "to put some skin in the game." The double moral hazard problem is then incorporated within a standard New Keynesian framework.

In Gerali et al. (2010) (EA_GNSS10), banks channel funds from patient households to entrepreneurs and impatient households. Meanwhile, the bank faces a leverage constraint as a form of paying a pecuniary cost whenever its net worth to asset ratio moves away from an exogenously given target. The bank's optimal decision implies that credit supply depends positively on bank net worth. In addition, banks have monopolistic power to set deposit and loan rates. These rates exhibit stickiness due to adjustment costs. The banking sector is included in a model with collateral constraint à la Iacoviello (2005). While the preceding two models are calibrated, the EA_GNSS10 model is estimated on the euro area macroeconomic data. Table 10 summarizes the key features of the models with bank capital.

Table 10 Comparison of key modeling features: Bank capital models

	NK_GK11	NK_MM10	EA_GNSS10
Model structure			
Key agents	Representative household	Representative household Risk-neutral entrepreneurs Risk-neutral bankers	Patient and impatient households Utility-maximizing entrepreneurs Monopolistic competitive banks
Production sector	One-sector CD technology	One-sector CD technology	One-sector CD technology
Real and nominal rigidities			
Consumption habit formation	Yes	Yes	Yes
Expenditure adjustment cost	Investment adjustment cost	No	Investment adjustment cost
Capital utilization	Yes	Yes	Yes
Consumer prices	Calvo pricing, partial indexation	Calvo pricing, partial indexation	Rotemberg pricing, partial indexation
Nominal wages	Flexible	Calvo pricing, partial indexation	Rotemberg pricing, partial indexation
Housing prices	–	–	Flexible
Financial frictions			
The role of the bank	Moral hazard problem between depositors and financial intermediaries	Holmstrom and Tirole (1997) double moral hazard: first between depositors and banks and second between banks and entrepreneurs	Adjustment cost of the bank capital to asset ratio, stickiness in deposit and lending rates
Collateral constraints	–	–	Kiyotaki and Moore (1997) type, nominal terms
Model Parameters			
Estimation/ Calibration	Calibration, U.S. data	Calibration, U.S. data	Bayesian estimation, euro area data: 1998Q1–2009Q1
Reference paper			
	Gertler and Karadi (2011)	Meh and Moran (2010)	Gerali et al. (2010)

6.1.4 Exploring How the Financial Sector Propagates and Amplifies Disturbances

In the following, we use MMB to explore and compare the dynamics of the above-mentioned macro-finance models. For models with financial accelerator on corporate investment and models with housing finance, we compare impulse response functions to a monetary policy shock, a general technology shock, and shocks that are more akin to aggregate demand shocks. Here, we extend the model comparison approach outlined in Section 3 by utilizing some economic shocks as common shocks in the models considered. The Smets and Wouters (2007) model (US_SW07) serves as a benchmark for comparison. Furthermore, we use the monetary policy rule estimated by Smets and Wouters (2007) as the common policy rule for all models. In this manner, we can isolate differences due to structural assumptions of each model from differences due to different assumptions on monetary policy. The SW rule is given in Eq. (10). For models with a role of bank capital, we simulate the original model to investigate the effects of an unexpected reduction in bank capital.

6.2 Propagation Mechanisms: Investment Financing and the Financial Accelerator

6.2.1 Striking Differences in Amplification of the Effect of Monetary Policy

To begin, we compare the transmission of the monetary policy shock in the four models with financial accelerator effects due to information asymmetries in the financing of corporate investment, (NK_BGG99, US_CD08, US_DG08, and US_CMR14), relative to the benchmark (US_SW07). Fig. 12 displays the effects of an unanticipated increase in the nominal interest rate of one percentage point for the commonly defined macroeconomic aggregates. In all four models, the nominal interest rate increases while output and inflation decline. The standard channel of monetary transmission is reflected in higher real interest rates that lead households to reduce consumption today and firms to refrain from investment.

The financial accelerator mechanism is at work in all four models that contain financial frictions. As can be seen from Fig. 13, firms' net worth falls due to a reduction in the price and return of capital.[m] Borrowing needs and leverage[n] of entrepreneurs increase, and the external finance premium (EFP) rises, depressing investment. The US_CD08 model, where the financial contract is in nominal terms, also exhibits a debt-deflation mechanism.

Yet, the magnitude, timing, and dynamic pattern of responses differ substantially across the models. It is particularly striking that the smaller New Keynesian models NK_BGG99 and US_CD08 display much stronger responses of output and inflation

[m] Note that the financial variables have not been redefined as common variables. Thus, the differences can only be interpreted qualitatively. Yet, the impact on GDP is directly comparable.

[n] Leverage is defined as the ratio of the value of capital, Q_tK_t, to the entrepreneur's net worth.

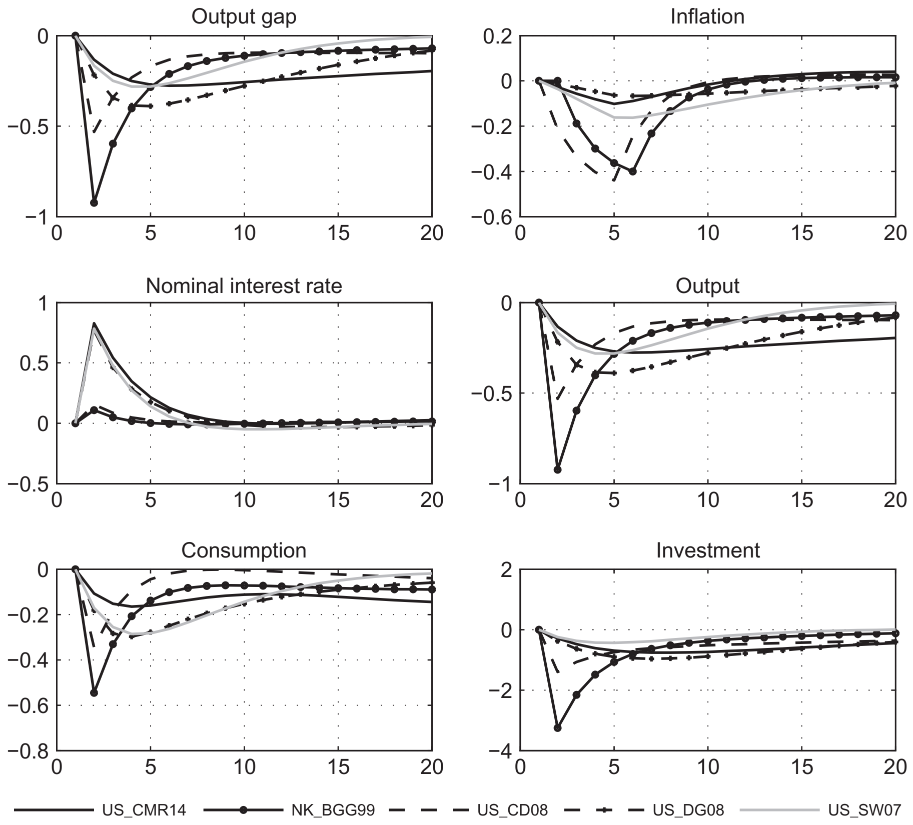


Fig. 12 Impulse responses to a contractionary monetary policy shock under SW rule: Macro variables. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Inflation is the rate of inflation over the previous four quarters. Nominal interest rate is annualized. Other variables are expressed in quarterly terms.

and a much smaller response of the nominal interest rate than the medium-size DSGE models US_SW07, US_DG08, and US_CMR14. This diversity of responses to a monetary policy shock stands in contrast to the findings of Taylor and Wieland (2012). The estimated medium-size DSGE models with financial accelerator US_DG08 and US_CMR14 still remain close to the other medium-size models, although the response of output in US_CMR14 is substantially more persistent as discussed in Section 5.2.

In US_DG08, investment responds more strongly to the unexpected policy tightening than in US_SW07 due to the financial accelerator effect.[o] The effect on consumption remains very similar. In sum, the impact on GDP is magnified a bit. GDP declines by about 40 basis points relative to 30 basis points in US_SW07. There is no similar

[o] Noteworthy, estimates of the curvature of the investment adjustment costs function are almost identical in US_SW07 and US_DG08.

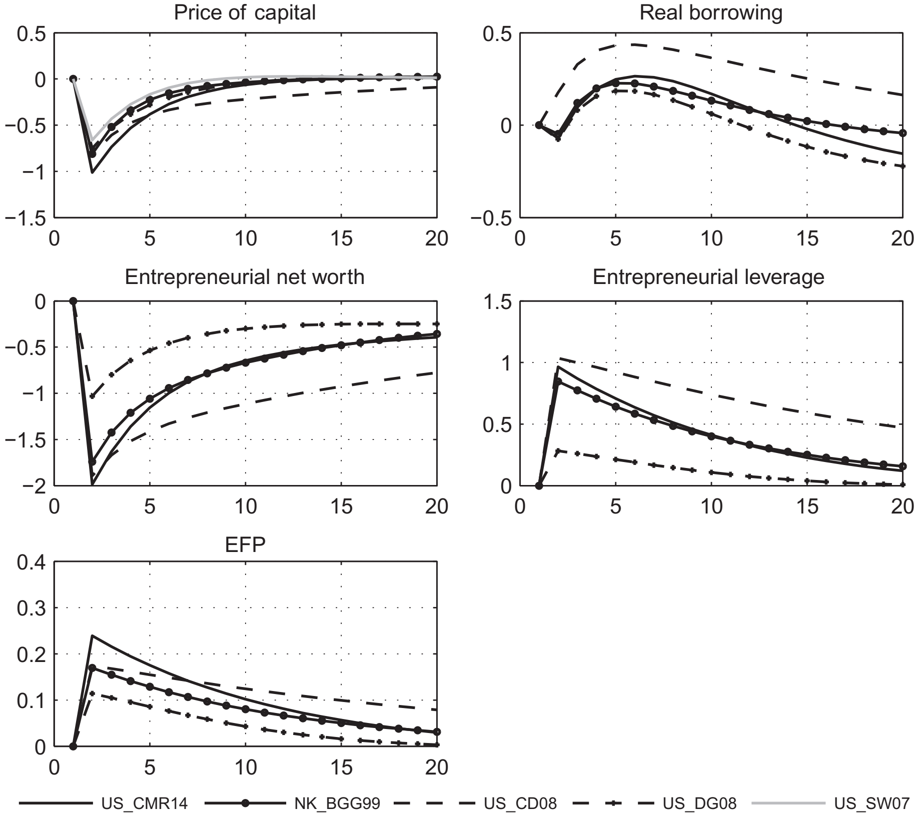


Fig. 13 Impulse responses to a contractionary monetary policy shock under SW rule: Financial variables. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. EFP (external finance premium) is annualized. Other variables are expressed in quarterly terms.

magnification effect, when output responses in US_SW07 and US_CMR14 are compared. The reasons are a somewhat smaller consumption response and a weaker financial accelerator effect to investment in US_CMR14. In particular, the response of investment in US_CMR14 is less pronounced than in US_DG08 due to a higher curvature of the investment adjustment cost function.

6.2.2 Investment Adjustment Costs Attenuate Sharp Responses of the External Finance Premium

Where does the big difference in GDP effects between medium-size DSGE models and smaller models with financial accelerator come from? The reason is the different working of the financial accelerator effect on investment in the two smaller models. The sharp increase in the external finance premium translates directly into a sharp reduction in investment in the two smaller financial accelerator models. In US_DG08 and

US_CMR14 instead the response of investment is hump-shaped and persistent, reaching a substantially lower peak effect than in NK_BGG99 and US_CD08. This is due to different specifications of adjustment costs across models: US_DG08 and US_CMR14 assume investment adjustment costs (as in Christiano et al., 2005), whereas NK_BGG99 and US_CD08 assume capital adjustment costs. As in US_DG08 and US_CMR14 it is costly to adjust the flow of investment, forward-looking agents adjust investment already today in expectation of an increase in the external finance premium. Accordingly, fluctuations in the premium have a smaller effect on the economy under investment adjustment costs than under capital adjustment costs ceteris paribus (see De Graeve, 2008). One might also ask why the largest impact on GDP occurs in NK_BGG99, rather than in US_CD08 where the financial accelerator is reinforced by a debt-deflation mechanism. This has to do with the calibration of capital adjustment costs. It is less costly to adjust capital in NK_BGG99 than in US_CD08.

Given the importance of the capital vs investment adjustment cost assumption and the striking differences it implies for output responses, one might ask which of the assumptions is supported by the data. We compare impulse responses of output from the models with the empirical impulse responses stemming from a VAR. To this end, we estimate a VAR, using the same observables and recursive identification as in Christiano et al. (2005) on the sample 1965Q3–2007Q3.[p] Fig. 14 presents impulse

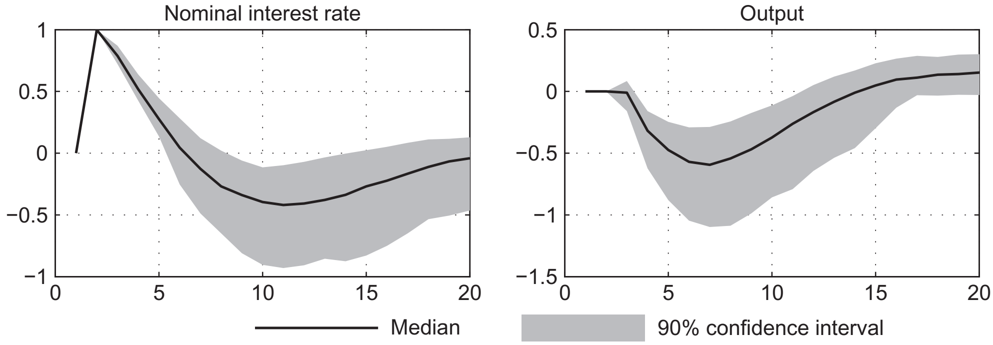


Fig. 14 Impulse responses to a one percentage point increase in the federal funds rate in a structural VAR. *Notes*: The variables and recursive identification are consistent with Christiano et al. (2005) on the sample of 1965Q3–2007Q3. The horizontal axis represents quarters after the shock. The solid lines refer to the median impulse responses. Shaded areas represent the 90% confidence intervals obtained by bootstrapping.

[p] The order of variables in the vector of observables is as follows: real GDP, real consumption, the GDP deflator, real investment, real wage, labor productivity, federal funds rate, the change in real money stock (M2), and real profits. The lag length is set to two quarters based on the Akaike Information Criterion. The VAR model also includes an intercept and a linear trend. The confidence bands are obtained by bootstrapping with 500 draws.

responses of the federal funds rate and real GDP to a one percentage point increase in the monetary policy rate. The federal funds rate increases on impact by one percentage point and then gradually declines. Real GDP exhibits a hump-shaped response, reaching a trough six quarters after the shock. This dynamic pattern of GDP response is consistent with investment adjustment costs assumption. Noteworthy, the VAR-based median response of output in the trough period is quantitatively close to the model-average of the trough output effect in the U.S. estimated models reported in Table 7.

6.2.3 Sharp GDP Responses Trigger Strong Contemporaneous Policy Feedback

Another difference between the medium-size models and the smaller models concerns the behavior of the nominal interest rate (see Fig. 12). In US_DG08, US_CMR14, and US_SW07 the nominal interest rate increases by about one percentage point in response to the policy shock as one might have expected. By contrast, the interest rate rises by less than 20 basis points in NK_BGG99 and US_CD08. In these two models, monetary policy has a strong contemporaneous effect on GDP growth that feeds back to the interest rate via the contemporaneous response to GDP growth in the SW rule. At first sight, this finding appears odd, particularly in light of the simulation of monetary policy shocks reported in Bernanke et al. (1999) which indicates a much stronger within-quarter effect of the policy shock on the interest rate. However, it turns out that the model dynamics are quite different under the original monetary policy rule. To illustrate this effect, we simulate all the other models under the original policy rule from Bernanke et al. (1999) (BGG99 rule).[q] The rule is given by:

$$i_t^z = 0.9 i_{t-1}^z + 0.11 p_{t-1}^z + \eta_t^i, \tag{12}$$

where i_t^z refers to the annualized short-term interest rate; p_t^z is the annualized quarter-to-quarter rate of inflation and η_t^i refers to the common monetary policy shock. As shown in Fig. 15, the strong contemporaneous feedback to the nominal interest rate disappears when simulating this rule with lagged inflation. Since this rule implies no reaction to the current state of the economy, the resulting impact of the policy shock on output and inflation is much greater.

The sensitivity of interest rate dynamics to the timing assumption of the policy rule in the two smaller models suggests that the specification of dynamics in these models is not rich enough to be used to assess the transmission of monetary policy in a quantitative manner for policy purposes. It indicates the usefulness of building and estimating medium-size DSGE models for this purpose. Interestingly, the five medium-size models considered here continue to indicate fairly similar GDP impact of policy shocks under the

[q] Simulations are carried out in the MMB menu *One policy rule, many models* by assigning the policy rule coefficients under the tab "User-specified rule."

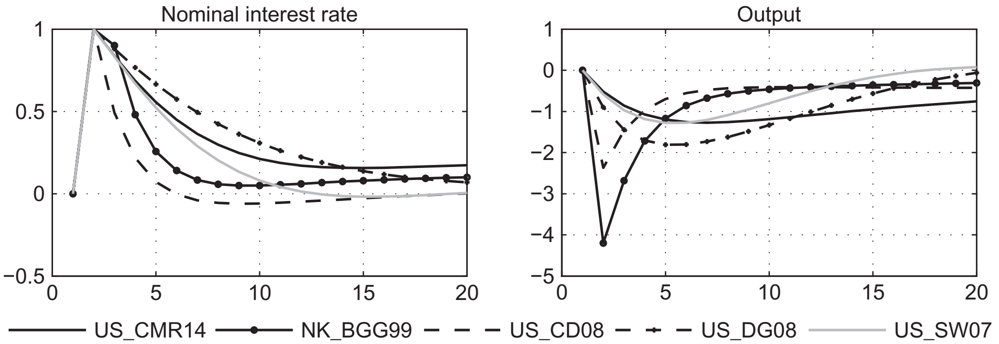


Fig. 15 Impulse responses to a contractionary monetary policy shock under BGG99 rule: Nominal interest rate and output. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Nominal interest rate is annualized. Output is expressed in quarterly terms.

rule from Bernanke et al. (1999) (BGG) (US_SW07, US_DG08, and US_CMR14 are shown in Fig. 15, but not G7_TAY93 and US_ACEL).

6.2.4 Financial Accelerator or Decelerator of Productivity Disturbances?

Figs. 16 and 17 report on the impact of a positive 1% technology shock.[r] The degree of exogenous persistence of this shock is assumed to be identical in the models considered. In particular, we set the common persistence parameter of the AR(1)-technology process to 0.9. Again, the common monetary policy rule corresponds to the estimated interest rate rule in US_SW07.

In all four models output increases in response to such technological progress. This increase is also visible in investment and consumption. Due to the rigidity of price adjustment, and in the case of the US_SW07 and US_DG08 models also nominal wage adjustment, actual output increases less than the output level that would be realized under flexible prices. For some time, a gap opens up between actual output and this measure of potential output.[s] The negative output gap leads to a decline in inflation. The SW rule then calls for monetary easing and the nominal interest rate declines.

With regard to the financial accelerator effect, the price of capital, firms' net worth and real borrowing increase in response to the technology shock. As leverage first declines and then rises, so does the external finance premium. Magnitudes and dynamic

[r] For comparison, the size of the shock in each model is scaled such that it would increase output on impact by 1% in the absence of endogenous responses of other variables.

[s] De Graeve (2008) defines potential output as the level of output under flexible prices and *in absence* of financial frictions. For direct comparability with the other financial accelerator models US_CD08 and NK_BGG99, we employ a common definition of potential output—under flexible prices and *in the presence* of financial frictions—also for US_DG08. The results are, however, not sensitive to the definition of potential output in this case.

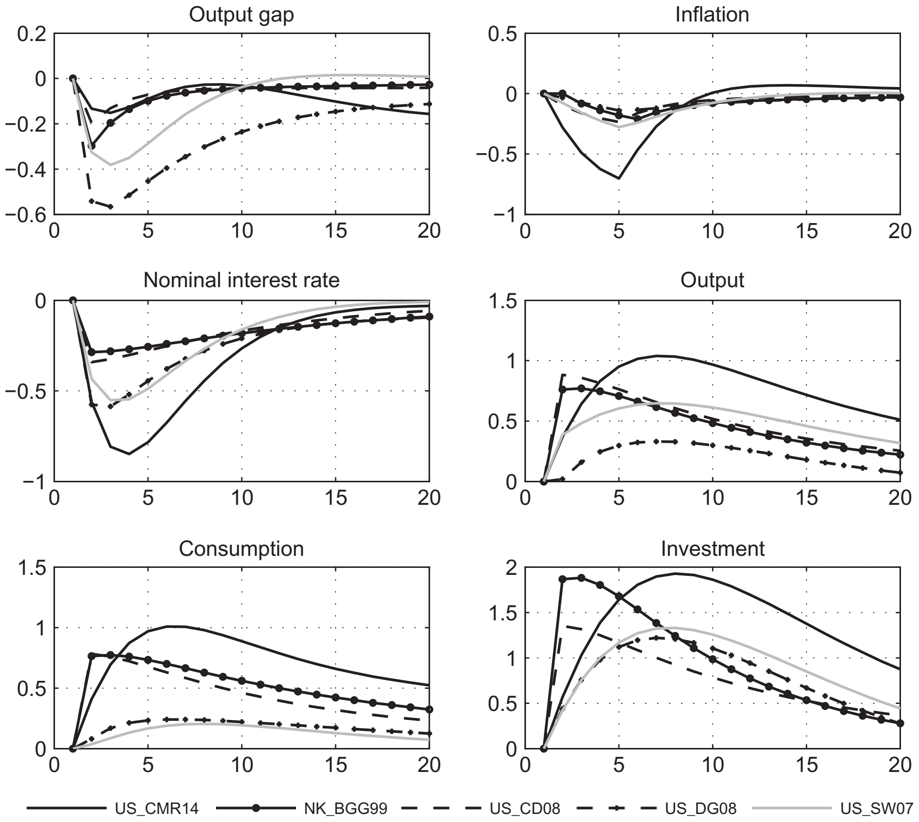


Fig. 16 Impulse responses to a positive technology shock under SW rule: Macro variables. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Inflation is the rate of inflation over the previous four quarters. Nominal interest rate is annualized. Other variables are expressed in quarterly terms.

patterns differ. Again, the NK_BGG99 and US_CD08 indicate a sharp positive impact of the change in financial variables on firms' investment. Investment and output dynamics in US_SW07, US_DG08, and US_CMR14 follow a hump-shaped pattern departing from and returning to steady state more slowly than in the other two models. The presence of investment adjustment costs in the medium-size models explains the more sluggish responses than in the NK_BGG99 and US_CD08 models that assume capital adjustment costs. Bernanke et al. (1999) showed that the financial accelerator amplified the effect of technology shocks on investment and GDP relative to the benchmark without the financial friction. The model of De Graeve (2008) delivers the opposite result. Relative to the model without the financial friction, the financial accelerator mechanism added by De Graeve (2008) actually dampens the investment and GDP response to a technology shock. As the demand for and price of capital increase, investment stays high for some

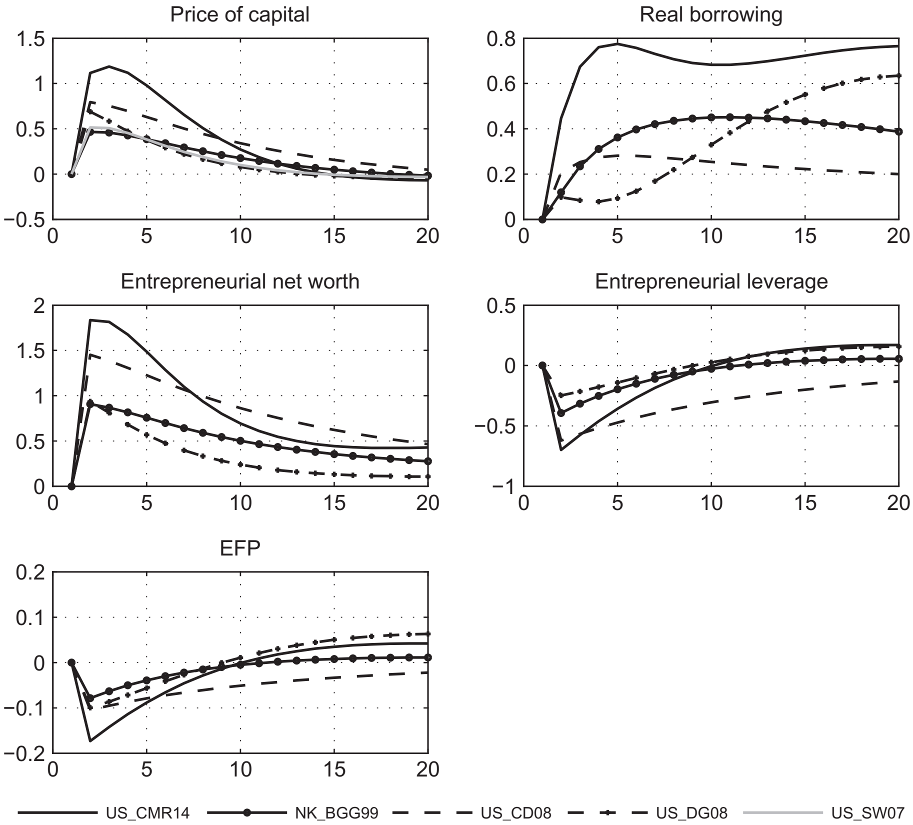


Fig. 17 Impulse responses to a positive technology shock under SW rule: Financial variables. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. EFP (external finance premium) is annualized. Other variables are expressed in quarterly terms.

time. The value of the capital stock then outgrows net worth and increases borrowing needs for quite some time. Accordingly, the external finance premium rises. As De Graeve (2008) notes, because long-lasting positive investment will be costly due to a high future premium for external finance, investment will be lower in all periods than otherwise. Indeed, the investment response in US_DG08 is smaller relative to US_SW07, which also features investment adjustment costs but no financial friction. However, this is not the case for US_CMR14, that is structurally very close to US_DG08 and yet exhibits larger responses of consumption, investment and hence output relative to US_SW07.

6.2.5 Estimated Parameters of Price Stickiness Make a Difference for the Accelerator Effect

A key parameter for investment dynamics is the curvature of the investment adjustment costs. In US_SW07 and US_DG08, this parameter is almost identical (5.76 and 5.77, respectively), whereas in US_CMR14 it is estimated to be substantially higher (10.78). Ceteris paribus, a higher curvature makes the adjustment process costlier, dampening the investment response. Yet, the investment response in US_CMR14 is stronger than in the other models.

Another important parameter is the degree of price stickiness. In US_DG08, the probability that firms will not be able to reset the price, is estimated at 0.92, whereas the corresponding estimate in US_SW07 is 0.65 and 0.74 in US_CMR14. In other words, prices are more flexible in US_SW07 and US_CMR14. The degree of price flexibility determines the strength of response of consumption and investment variables to a technology shock, because it determines the dynamics of the real interest rate. Conditional on the monetary policy rule, which is common for all models in this exercise, more flexible prices imply that inflation will ceteris paribus fall by more in response to a technology shock, causing the central bank to loosen the nominal rate by more. As a result, the real interest rate is lower for more flexible prices. Thus, consumption increases more for more flexible prices. Also investment rises more substantially. In a model without financial friction such as US_SW07, equilibrium requires the real rate to move together with the aggregate return on capital. Therefore, in an economy with more flexible prices, capital increases by more ceteris paribus, causing a higher investment response. With the financial accelerator, a lower real rate translates into a lower external finance premium, strengthening entrepreneurial net worth and thereby also supporting the investment boom.

To sum up, greater price stickiness dampens the responses of consumption and investment to technology shocks. This is a further reason why the responses in US_DG08 are smaller relative to the model without the financial friction—US_SW07. The other important reason is the presence of the 'decelerator' effect, as described in De Graeve (2008). The differences in price stickiness also explain why consumption and investment responses are stronger in US_CMR14 when compared to US_DG08.

With regard to the earlier findings of Bernanke et al. (1999) it is noteworthy to point out the sensitivity to the assumption for the monetary policy rule and the persistence of the technology process. They use a random walk process for technology. In this case, a shock has very large and persistent effects on output. Consequently, actual output exceeds potential output and inflation goes up.

6.2.6 Investment-Specific Shocks

We have also simulated and compared the impact of investment-specific shocks in the US_SW07, US_DG08, and US_CD08 models. De Graeve (2008) calls this shock an investment supply shock, since it causes investment to increase and the price of capital to decrease. Smets and Wouters (2007) group it under (aggregate) demand shocks because they lead to an increase in both output and inflation. In this context, it is of interest to note that such investment-specific shocks play an important role in explaining the Great Recession following the global financial crisis when the US_SW07 model is extended to cover this period (see Wieland and Wolters, 2013). Conditional on the model parameterization, the financial friction included in the US_DG08 and US_CD08 models dampens the impact of such investment shocks on investment and GDP.

6.3 Propagation Mechanisms: Housing Finance and Credit Booms

Next, we compare the effects of monetary and technology shocks in the three models with housing finance, US_IAC05, US_IN10, and NK_KRS12, relative to the US_SW07 model as benchmark.[t] In addition, we examine the impact of demand shocks originating in the housing sector on the broad economy.

6.3.1 Monetary Transmission via Housing Finance

Fig. 18 shows the consequences of a contractionary monetary policy shock on macro variables. Qualitatively, the three models with housing finance exhibit the same Keynesian-style features as the benchmark. Due to price rigidities, the contractionary monetary shock induces an increase in the real interest rate, output declines below its flexible price level,[u] and this gap causes lower inflation. Both, consumption and investment decrease. Quantitatively, the impact on real GDP is much sharper and more pronounced in the US_IAC05 and US_IN10 models. The NK_KRS12 model, however, is closer to the US_SW07 benchmark. The latter two models exhibit more muted and hump-shaped responses of GDP and its components, consumption and investment.

Fig. 19 displays the transmission of the monetary shock via housing finance. The collateral constraints on nominal borrowing in the US_IAC05 and US_IN10 models magnify the effect of unanticipated policy tightening. As inflation falls and real house prices decrease, the debt capacity of borrowers is reduced. In the US_IAC05 model impatient households and entrepreneurs are both borrowing constrained. Accordingly, the impatient households cut back further on consumption, while the entrepreneurs reduce nonresidential investment along with consumption. Likewise, in the US_IN10 model impatient households curtail consumption by more. Moreover, residential investment declines significantly, because sticky wages in combination with flexible house prices intensify the effect of a monetary shock on output in the residential sector. Meanwhile, output shows no hump-shaped responses in these two models. The reasons is that the

[t] The components of aggregate consumption and investment differ in the models with housing due to different assumptions on the production sector and housing market. In the US_IN10 and NK_KRS12 models, aggregate consumption consists of the consumption of patient and impatient households, and investment is defined as the sum of nonresidential and residential investment. Meanwhile, in the US_IAC05 model, aggregate consumption includes the consumption of entrepreneurs additionally to two types of households' consumption and investment comprises only nonresidential investment.

[u] In the NK_KRS12 model, the potential output is defined as the level of output that could be realized without nominal *and* financial frictions.

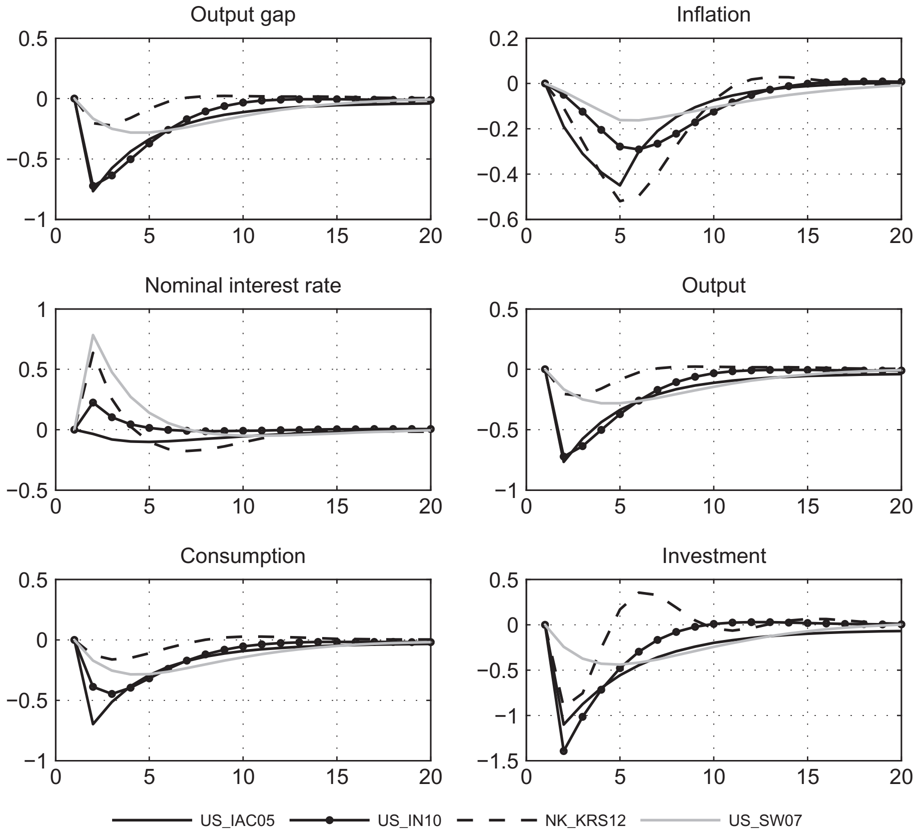


Fig. 18 Impulse responses to a contractionary monetary policy shock under SW rule: Macro variables. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Inflation is the rate of inflation over the previous four quarters. Nominal interest rate is annualized. Other variables are expressed in quarterly terms.

US_IAC05 model exhibits no habit formation in consumption and only small capital adjustment costs, while the US_IN10 model features no capital adjustment costs.

The NK_KRS12 model exhibits a more flexible collateral constraint. This generates less amplification than the standard collateral constraints used in the other models with housing. A higher loan-to-value ratio in this model is accompanied by a rise in lending rates. By contrast, the amount of borrowing is restricted to a certain fraction of collateral in case of the standard collateral constraint. Accordingly, a fall in the collateral value leads directly to the reduction of borrowing. In the NK_KRS12 model, impatient households still take out more loans even with higher interest rate in response to a contractionary monetary shock. This dampens the responses of consumption and residential investment. Furthermore, since there is no capital in this model, aggregate demand does not include nonresidential investment which is an interest-sensitive component of GDP. Overall, the

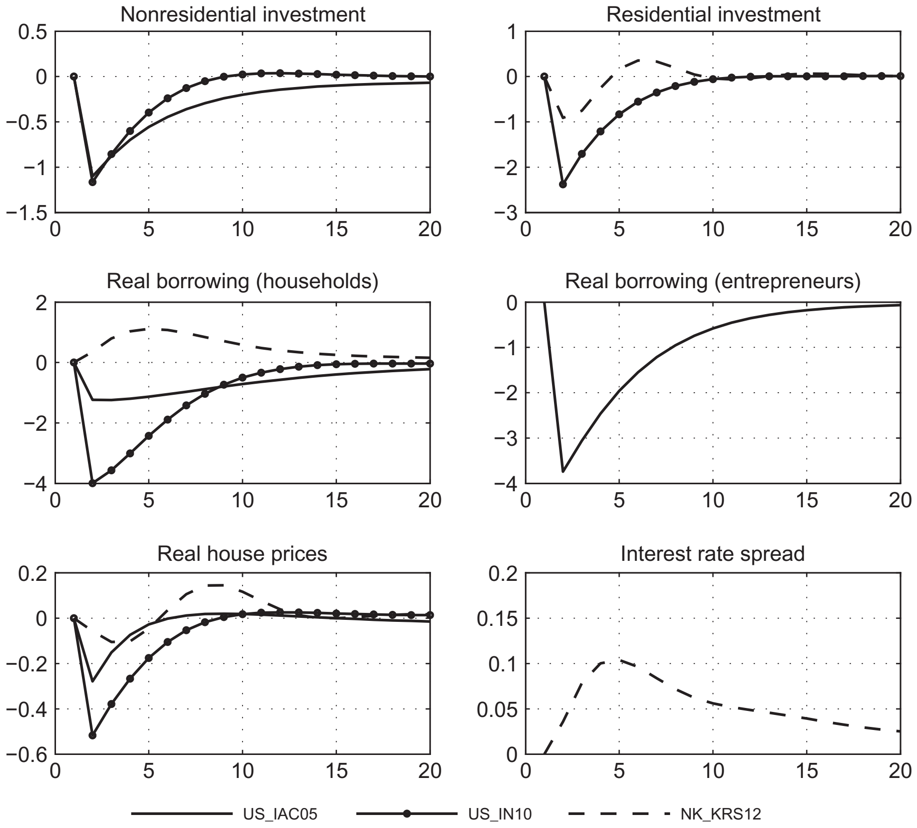


Fig. 19 Impulse responses to a contractionary monetary policy shock under SW rule: Investment & financial variables. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Interest rate spread is annualized. Other variables are expressed in quarterly terms.

impact of the monetary shock on output is smaller in the NK_KRS12 model than in the other models.

As in the case of the NK_BGG99 and US_CD08 models in Section 6.2.3, we find that due to insufficient real rigidities, the US_IAC05 and US_IN10 models exhibit a sharp contemporaneous response of output that strongly feeds back via the SW rule to the contemporaneous nominal interest rate. For the US_IAC05 model, the positive monetary policy shock implies a slight decline in the nominal interest rate. Similarly to the financial accelerator models analyzed earlier, this strong contemporaneous effect disappears when the models are simulated under the policy rule of Bernanke et al. (1999) (see Fig. 20).

The sensitivity of interest rate dynamics to the timing assumption of the policy rule suggests that the dynamics in these models are not rich enough to be used to assess the transmission of monetary policy in a quantitative manner as in the case of medium-size

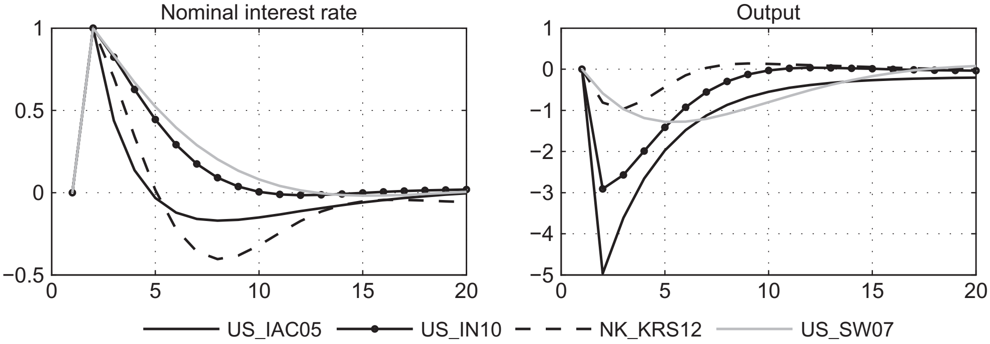


Fig. 20 Impulse responses to a contractionary monetary policy shock under BGG99 rule: Nominal interest rate and GDP. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Nominal interest rate is annualized. Output is expressed in quarterly terms.

DSGE models with more sources of endogenous persistence. Thus, the comparison of monetary transmission mechanisms in the two groups of macro-financial models supports including habit formation in consumption and investment adjustment costs in models for quantitative monetary policy analysis.

6.3.2 General Technological Progress and Housing Finance

We also examine effects of a common technology shock in the housing finance models. The shock has a common autocorrelation coefficient of 0.9. In the US_IN10 and NK_KRS12 models, which contain two production sectors, the shock increases the total factor productivity in the nonresidential (consumption goods) sector. Figs. 21 and 22 present the impact of a 1% increase in such a shock.

As in the US_SW07 model, GDP rises and inflation declines in response to a positive technology shock in the models with housing. It leads to a housing boom without inflation, which is amplified by collateral constraints.

The persistent but temporary increase of productivity in the nonresidential sector is followed by a lower real interest rate so that aggregate demand is equated to the expanded aggregate supply. The reduction of the real rate causes real house prices to rise, which in turn increases the borrowing capacity of collaterally constrained agents. This allows borrowers to obtain more funds, which are either consumed or invested. The amplifying effect of the collateral channel is most apparent in the responses of consumption. Consumption increases two or four times more in the housing finance models than in the US_SW07 model. Though the decline in inflation reduces collateral values, the collateral channel outweighs the debt deflation channel.

Surprisingly, the output gap in the US_IN10 model increases, whereas inflation declines. When house prices rise, the combination of flexible house prices and sticky wages in the residential sector increases new housing construction by more than in

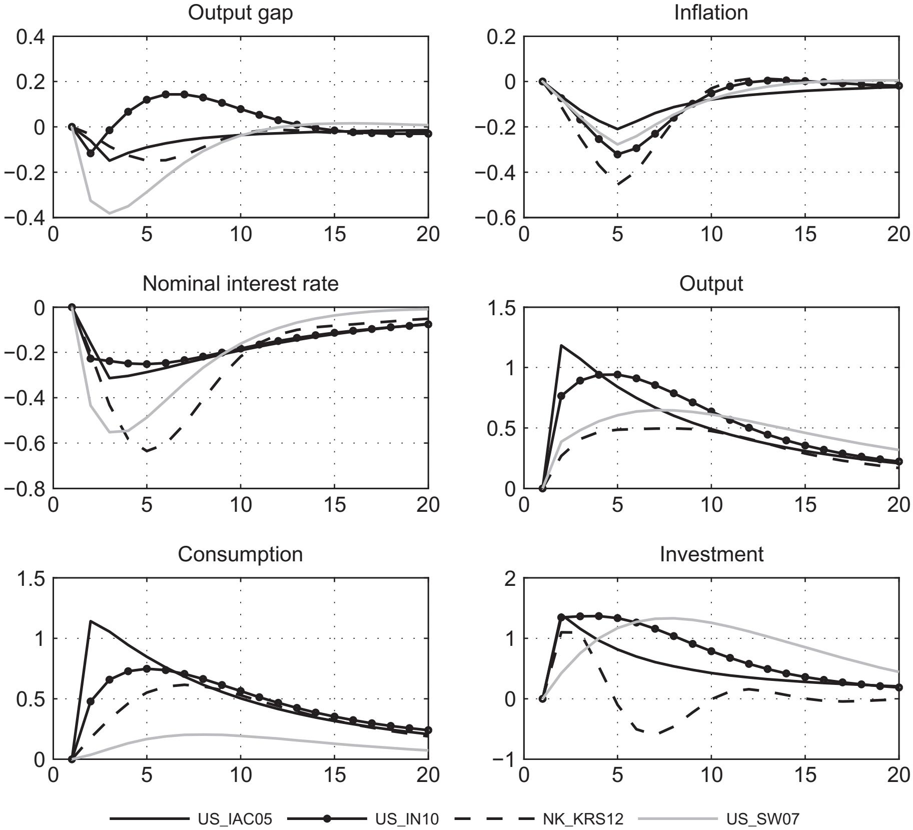


Fig. 21 Impulse responses to a positive technology shock under SW rule: Macro variables. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Inflation is the rate of inflation over the previous four quarters. Nominal interest rate is annualized. Other variables are expressed in quarterly terms.

the case of flexible wages. As a result, total output, the sum of the value added of the two sectors, increases beyond the level of output that would be realized if prices and wages were flexible. Yet, with the two-sector production structure, a positive output gap does not necessarily lead to an increase in inflation in the consumption sector. Though positive spillover effects from the residential to the nonresidential sector put upward pressure on inflation, the positive technology shock also lowers the marginal cost of intermediate goods. The latter effect dominates and inflation declines.

6.3.3 Housing Demand Shocks Driving a Housing Boom

The models with a housing sector include new types of shocks emanating from this sector that have potentially major macroeconomic consequences. In the following, we consider a housing demand shock. It could also be called a housing preference shock, since it is

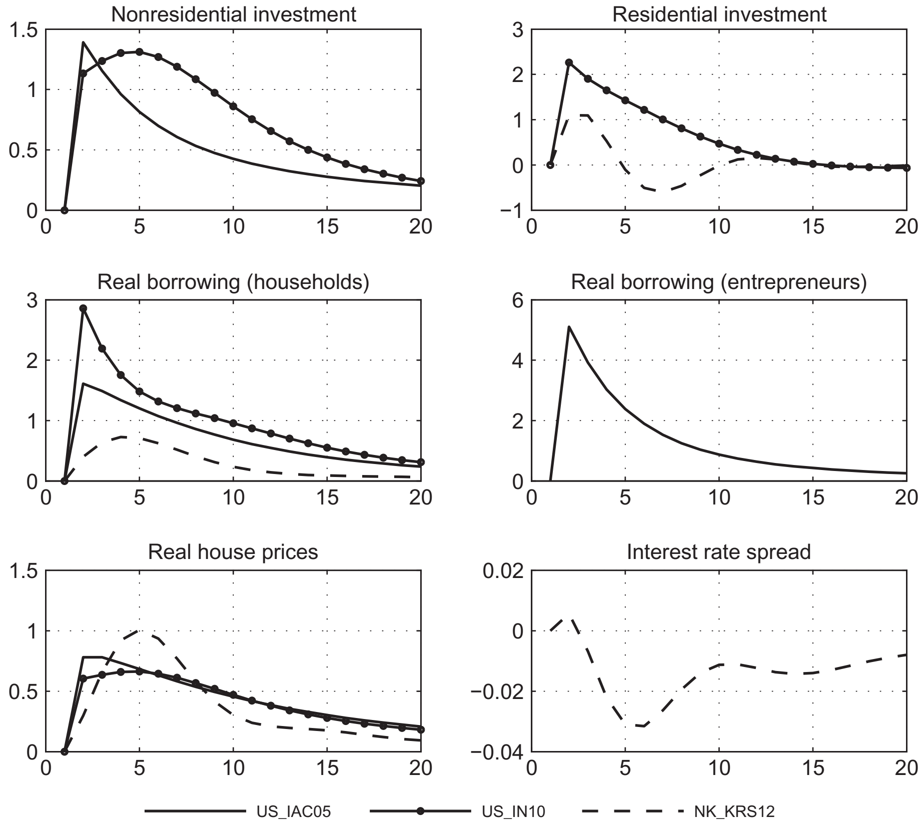


Fig. 22 Impulse responses to a positive technology shock under SW rule: Investment & financial variables. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Interest rate spread is annualized. Other variables are expressed in quarterly terms.

modeled as random disturbance to utility from housing services. For comparison, the size of the shock is adjusted across the models such that it increases the real house prices on impact by 1%. Yet, we ask a slightly different question than previously with the technology shock, namely, what the consequences of such a housing demand shock would be when the degree of exogenous persistence remains model-specific.[v] Under this scenario, GDP increases in all the models. The housing boom leads to an economic boom.

However, the responses of other macroeconomic and financial variables are quite different across the models as shown in Figs. 23 and 24. The heterogenous dynamics reflect the different model structure and assumptions with regard to the housing market.

[v] The AR(1) coefficients of a housing demand shock for each model are the following: 0.85 (US_IAC05), 0.96 (US_IN10), 0.95 (NK_KRS12)

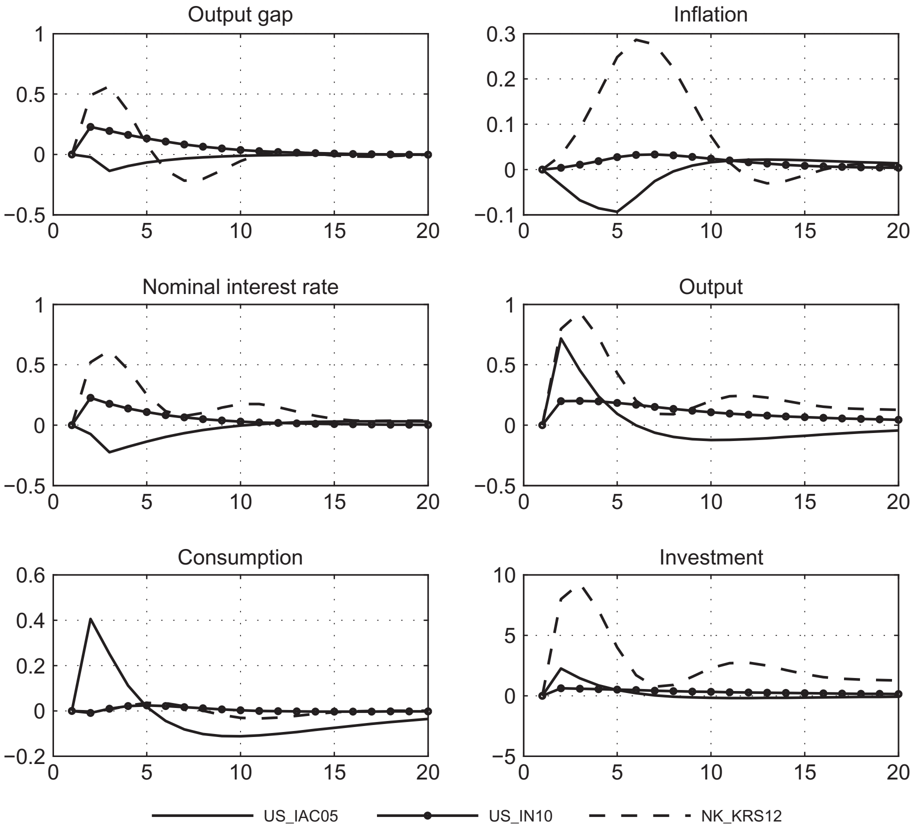


Fig. 23 Impulse responses to a positive housing demand shock under SW rule: Macro variables. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Inflation is the rate of inflation over the previous four quarters. Nominal interest rate is annualized. Other variables are expressed in quarterly terms.

Unlike the US_IAC05 and US_IN10 models, house prices are subject to Calvo-type nominal frictions in the NK_KRS12 model. Thus, house prices continue to rise for more than one year after the initial shock.[w] The effects of higher house prices on investment and GDP are amplified by the financial accelerator mechanism. As shown in Fig. 24, the surge in residential investment and housing prices dominates the increase in households' borrowing. As households' leverage decreases, financial intermediaries charge a lower spread of the lending rate over the deposit rate. The reduced spread results in a further increase of borrowers' housing demand, which in turn leads to another increase of house prices. Actual GDP rises more than it would under flexible prices, hence a gap opens up and inflation goes up.

[w] The Calvo pricing parameter in the housing sector is 0.75.

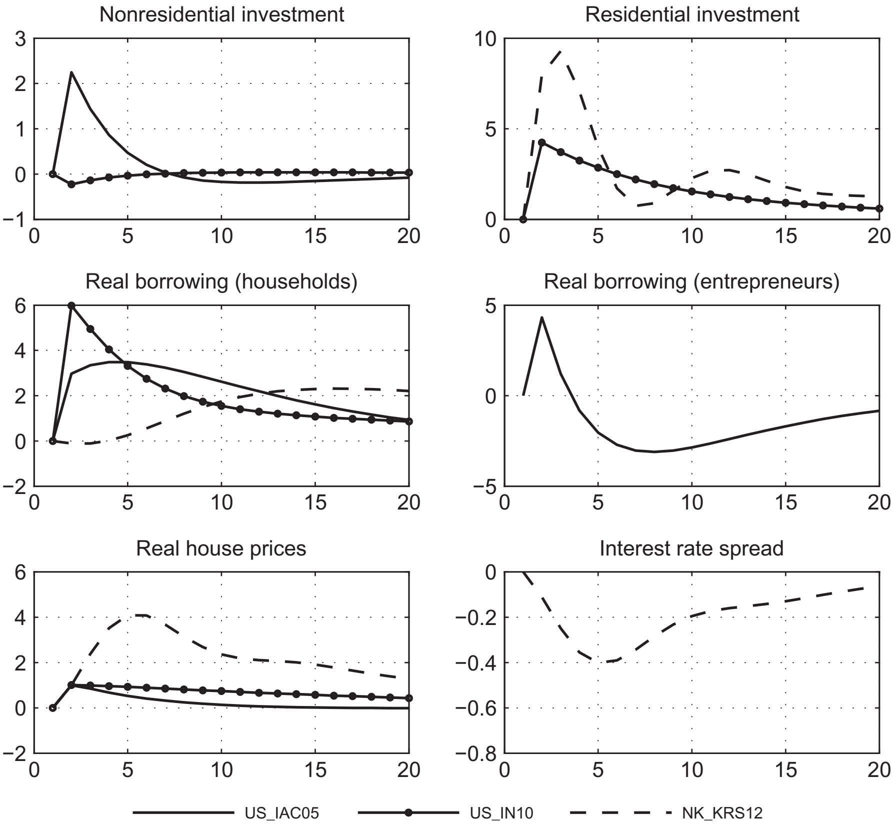


Fig. 24 Impulse responses to a positive housing demand shock under SW rule: Investment & financial variables. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Interest rate spread is annualized. Other variables are expressed in quarterly terms.

In the US_IAC05 model, the housing demand shock sharply pushes up consumption, investment, and GDP. The increase in households' demand for housing drives up house prices. As a consequence, the collateral value of borrowers rises and the borrowing capacity is expanded. It leads impatient households to increase consumption and entrepreneurs to invest more. The model does not exhibit hump-shaped dynamics since it assumes no habit formation in consumption and only a small adjustment cost in nonresidential investment. Contrary to the other two models, flexible price output rises more than actual output. Accordingly, inflation declines. The reason is that the increased physical capital and housing stock raise factor productivities, thereby shifting the aggregate supply curve outwards. With borrowing-constrained entrepreneurs, the housing preference shock acts like an aggregate supply shock, which causes output and inflation to move in opposite directions.

The response of output is smallest in the US_IN10 model. The housing demand shock expands the borrowing capacity of impatient households, so that they increase consumption and housing investment. The role of the collateral channel is illustrated by the responses of residential investment and real borrowing of the households. However, patient households decrease consumption and investment in response to the increase in interest rates. Overall, GDP increases less than in the other two models.

6.4 Propagation Mechanisms: Financial Intermediation and Bank Capital

Finally, we explore macroeconomic consequences of shocks emanating from the banking sector. To this end, we make use of the three macro-financial models with a detailed representation of the banking sector: the model of Gertler and Karadi (2011) (NK_GK11), the model of Meh and Moran (2010) (NK_MM10), and the model of Gerali et al. (2010) (EA_GNSS10). Specifically, we evaluate the impact of an unanticipated reduction in bank capital on macroeconomic and financial variables. This shock can be interpreted as a sudden reduction in bank capital due to bank loan losses and asset writedowns.

The technical definition of the shock, however, differs across the three models. In the NK_GK11 model, the shock is modeled as a one-time wealth transfer from banks to households, whereas in the NK_MM10 model the shock is defined as sudden accelerated depreciation of bank net worth. In the EA_GNSS10 model, the shock implies an unexpected deadweight loss to bank net worth. In the NK_MM10 and EA_GNSS10 models, the shock follows a first-order autoregressive process, whereas in NK_GK11 the shock is assumed to have no persistence at all.[x] The size of the shock is normalized such that bank capital declines by 5% on impact in all models.

The question to be answered with this comparison exercise differs from the previous comparative analysis. Rather than investigating the consequences of bank capital shocks under a common monetary policy and a common shock process, we ask what consequences would be predicted by the different models. Thus, the scenario assumes model-specific policy rules and model-specific bank capital shock processes.

Fig. 25 displays simulation outcomes of a shock that reduces bank net worth.[y] In all models, the decrease in bank net worth in the presence of a constraint on bank leverage leads to a decline in lending, which in turn reduces investment and output. However, in the NK_GK11 model, investment and output recover relatively quickly from the decrease, whereas in the EA_GNSS10 and NK_MM10 models, they decline for some time.

[x] The autocorrelation coefficients are as follows: 0.95 (EA_GNSS10), 0.9 (NK_MM10), and 0 (NK_GK11). Gerali et al. (2010) set the parameter to 0.95 for this simulation exercise, although the median of the posterior distribution for this parameter is 0.81.

[y] To perform these simulations, we use replication files for each model.

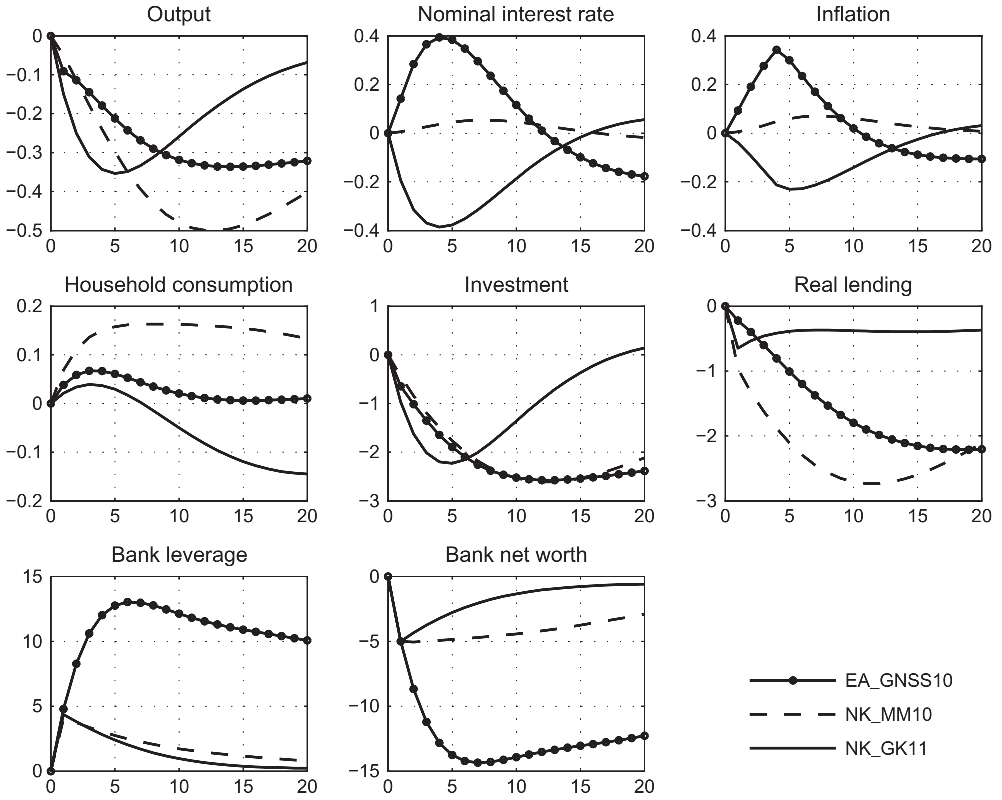


Fig. 25 Impulse responses to a negative shock to bank net worth. *Notes*: Horizontal axis represents quarters after the shock. Units of the vertical axis are percentage deviations from steady-state values. Inflation is the rate of inflation over the previous four quarters. Nominal interest rate is annualized. Other variables are expressed in quarterly terms.

The transmission and propagation channels differ across models. In NK_GK11, the financial accelerator mechanism applies to the bank. Since bank net worth declines, financing conditions get tighter. Correspondingly, bank lending goes down, external finance premia rise sharply, and aggregate investment declines. As to the speed of return to steady-state conditions, the main reason for a faster rebound of investment is the absence of serial correlation in the bank capital shock. Household consumption increases somewhat following the one-time redistribution from the bank but declines steadily afterwards.

In EA_GNSS10, banks reduce credit supply and increase the lending rates in order to repair their balance sheets after a shortfall in bank net worth. It also depresses demand for loans via the collateral channel. As a result, investment declines. Since bank interest rates adjust only in a sticky fashion, tight financing conditions persist for several periods, depressing investment further. The decline in bank net worth is persistent. This is due to the endogenous decline of bank retained earnings as well as the exogenous persistence of the shock process. Meanwhile, household consumption slightly increases mainly due to higher wages.

In NK_MM10, the financial contract imposes a solvency condition on banks that determines banks' ability to attract funds for lending. Therefore, in response to an unanticipated fall in bank net worth, banks' ability to attract funds deteriorates and they reduce lending. The decline in loan supply depresses investment, which lowers the retained earnings of banks and therefore bank net worth, reinforcing the initial shock endogenously. However, household consumption increases. The reason is that capital prices (not shown) increase in response to the shock, which in turn leads households to consume more as consumption goods become cheaper relative to capital goods.

With regard to output and inflation, the bank net worth shock appears to act as a negative demand shock in NK_GK11, where output and inflation decrease and call for monetary easing. By contrast, the shock acts as a negative supply shock in NK_MM10 and EA_GNSS10, where contraction in output is accompanied by modest inflationary pressures calling for some monetary tightening.

7. HOW TO ASSESS POLICY ROBUSTNESS: AN ILLUSTRATIVE EXAMPLE

Finally, we close the series of comparative exercises with an example of how one can evaluate the robustness of policy rules under model uncertainty. The idea is simple: A policy rule is more robust than another one, if it performs better, on average, across a range of models. The search for robustness has been a central objective for many contributions to the literature on model comparison (see Section 2). Here, we just illustrate how the MMB software can be employed to this end.

The global financial crisis has been preceded by a massive credit-driven real-estate boom in the United States and other economies. If central banks had responded earlier by raising interest rates, they might have been able to avoid excessive credit growth and housing price inflation. The Taylor rule, for example, would have recommended higher interest rates ahead of the crisis. The models with financial frictions include mechanisms that can explain such credit-driven booms. Thus, it is of interest to evaluate what rules perform better and whether it might be advantageous to "lean against the wind," that is, to include an explicit reaction to credit growth into the policy rule.

7.1 Participating Models and Rules

In this exercise, policy performance is evaluated across four different models with financial frictions that have been estimated for the U.S. economy and have appeared in the preceding sections: the US_DG08, US_CMR14, US_IAC05 and US_IN10 models. We consider eight simple monetary policy rules (see Table 11). These include the four model-specific rules that were estimated along with the respective macro-financial

Table 11 Eight interest rate rules

Model-specific rules	
DG08 rule	$i_t^z = 0.90 i_{t-1}^z + 0.23 p_t^z - 0.08 p_{t-1}^z + 1.14 q_t^z - 1.10 q_{t-1}^z$
IAC05 rule	$i_t^z = 0.73 i_{t-1}^z + 0.34 p_{t-1}^z + 0.14 y_{t-1}^z$
IN10 rule	$i_t^z = 0.60 i_{t-1}^z + 0.56 p_t^z + 0.82 y_t^z - 0.82 y_{t-1}^z$
CMR14 rule	$i_t^z = 0.85 i_{t-1}^z + 0.36 p_t^z + 0.05 y_t^z - 0.05 y_{t-1}^z$
Other simple rules	
Taylor rule	$i_t^z = 1.5 \pi_t^z + 0.50 q_t^z$
SW rule	$i_t^z = 0.81 i_{t-1}^z + 0.39 p_t^z + 0.97 q_t^z - 0.90 q_{t-1}^z$
OW08 rule	$i_t^z = 2.34 E_t \pi_{t+3}^z + 0.765 E_t q_{t+3}^z$
DIF rule	$i_t^z = i_{t-1}^z + 0.5 \pi_t^z + 0.5 (q_t^z - q_{t-4}^z)$

Note: The superscript z refers to common variables. i_t^z is the annualized short-term federal funds rate in quarter t. p_t^z refers to the annualized quarter-to-quarter rate of inflation, π_t^z is the year-on-year inflation rate, y_t^z is the deviation of quarterly real GDP from its long-run potential, while q_t^z refers to the output gap defined as the difference between actual GDP and the level of GDP that would be realized if prices and wages were flexible. All variables are expressed in percentage deviations from steady-state values.

model, that is, the DG08, CMR14, IAC05, and IN10 rules. They will be compared to four other simple rules: the well-known Taylor rule (see Taylor, 1993a); the SW rule; a forecast-based rule that was estimated to fit FOMC decisions in response to FOMC forecasts with real-time data by Orphanides and Wieland (2008) (OW08 rule); and a simple difference rule that performed very well in the studies of policy robustness by Levin et al. (2003) and Orphanides and Wieland (2013). The latter is referred to as the DIF rule.

7.1.1 Stabilization Performance and Robustness

A simple central bank loss function serves as a measure of performance. It is the sum of the unconditional variance of inflation deviations from the central bank's target and the unconditional variance of the output gap. Both variances are standard simulation output in MMB.

$$L = Var(\pi_t^z) + Var(q_t^z) \tag{13}$$

The resulting losses are reported in Table 12. The first row shows the losses under the model-specific rules in the respective models. The second row indicates the performance of one of them, the CMR14 rule, in all four models. Performance deteriorates in US_DG08 and US_IAC05 while it improves in US_IN10 relative to the model-specific estimated rule. The Taylor rule delivers much more stable outcomes than any of the rules estimated with the four macro-financial models. Average loss is much lower than under the CMR14 rule. The SW rule performs a little worse than Taylor's rule in US_DG08 but improves outcomes in the other three models even further.

Table 12 Stabilization performance and robustness

	US_DG08	US_CMR14	US_IAC05	US_IN10	Average loss
Model-specific rule	5.8	47.6	12.3	6.9	–
CMR14 rule	9.1	47.6	20.4	3.0	20.0
Taylor rule	5.3	34.5	6.2	4.3	12.5
SW rule	5.7	19.6	5.1	3.3	8.3
OW08 rule	4.6	29.3	∞	3.0	∞
DIF rule	2.7	5.5	3.3	2.6	3.6

Notes: The loss function is the sum of the unconditional variances of inflation and output gap. ∞ indicates indeterminacy.

The rule estimated on FOMC forecasts lacks robustness (OW08 rule). While it further improves performance in the US_DG08 model and US_IN10 model, losses in US_CMR14 deteriorate and it causes instability and multiple equilibria in US_IAC05. Finally, the DIF rule performs best in each of the four models and, thus, also on average.

7.1.2 Leaning Against Credit Growth

Next, we investigate whether it helps to add an explicit reaction to credit growth to the rules. The reaction coefficient on the quarterly growth rate of credit in real terms[z] takes on one of two values: 0.1 and 0.3.

Table 13 reports on the effectiveness of such a leaning-against-the-wind policy. Indeed, it helps adding a direct reaction to credit growth to the model-specific estimated rules in US_DG08, US_CMR14, US_IAC05. With a reaction coefficient of 0.1 they outperform the original model-specific rules. In US_IN10, however, leaning against credit growth does not improve the stabilization performance of the baseline rule. In US_CMR14 and US_IAC05, a stronger reaction to credit growth (0.3) further reduces loss, whereas in US_DG08 and US_IN10 the loss increases again.

For the SW rule, at least some leaning against credit growth is beneficial in the US_DG08, US_CMR14, and US_IN10 models, but not in US_IAC05. In case of the DIF rule, which is already very effective in stabilizing output and inflation in any of the four models, leaning against credit growth is destabilizing, except with the US_CMR14 model.

These results suggest that some degree of leaning against credit growth can help reduce output and inflation variability. Yet, the possibility of improvement depends on the baseline rule and the particular model. If the baseline rule without credit growth is already fairly robust, "leaning against the wind" is more likely to hurt performance.

As a next step, it would be of great interest to employ techniques for model-averaging and worst-case analysis to search for robust rules within a larger set of macro-financial

[z] As the model-specific rule in US_IAC05 reacts to lagged outcomes, a lagged credit growth rate is considered for this rule. In other cases, contemporaneous credit growth is used.

Table 13 Stabilization performance of policy rules with leaning-against-the-wind (credit growth)

	US_DG08	US_CMR14	US_IAC05	US_IN10
Model-specific rule				
Baseline	5.8	47.6	12.3	6.9
Leaning (0.1)	5.3	28.8	11.4	7.0
Leaning (0.3)	6.1	19.8	11.3	7.8
SW rule				
Baseline	5.7	19.4	5.1	3.3
Leaning (0.1)	4.9	13.1	5.3	3.1
Leaning (0.3)	4.7	8.4	6.7	3.7
DIF rule				
Baseline	2.7	5.5	3.3	2.6
Leaning (0.1)	2.8	4.7	3.9	2.7
Leaning (0.3)	3.5	5.2	5.0	3.3

Notes: the loss function includes the variance of inflation and the variance of the output gap.

models. Such a search could make use of optimization procedures from earlier work on policy robustness under model uncertainty (see Section 2).

8. CRITICAL ASSESSMENT AND OUTLOOK

While there is a limited set of macroeconomic time series and country experiences, there is a multitude of macroeconomic models, and their number is growing rapidly. This is as much due to economists' creativity as to the great challenges faced by policy makers, for which they need advice and guidance based on more adequate models. There are many urgent policy questions. For example, economists need to obtain a better understanding of the macroeconomic consequences and interactions of banking regulation, private and public debt, fiscal consolidation, macroprudential policies and structural reforms. Furthermore, globalization and growth have created a demand for economic modeling expertise in many countries around the world. While academic research in macroeconomics largely focuses on the United States, Europe, and Japan, central banks and government institutions in many other countries need models that are more appropriate for analyzing their economies.

At the same time, much effort is invested in building models that will never be used by anyone else. Rather than building directly on work by others, researchers usually start from scratch. Practices that would ensure easy reproducibility are not widespread. There is little systematic comparison of existing models.

This need not be the case. There has been tremendous progress with regard to model design, model solution techniques, econometric estimation procedures and software solutions. Many researchers are using those same techniques. While model comparison was extremely cumbersome in the past, a task reserved for meetings of teams of modelers from policy institutions, it can now be accomplished fairly easily by individual researchers. As this chapter aimed to show, comparative model analysis helps critically assessing available models, identifying similarities and differences as well as empirical inconsistencies that require more research.

The potential for comparative work has barely been tapped. Going forward, key areas for more methodological work that could rapidly bear fruit are the following: comparisons of the role of expectations formation, learning and heterogeneity; model validation and real-time estimation of competing macroeconomic models; combining statistical nowcasting techniques with model-based forecasting for the medium term; and implementation of nonlinear solution techniques for occasionally binding constraints.

Another important aspect concerns openness to competing modeling paradigms. Despite many critical assessments of the DSGE approach and its microeconomic foundations in the aftermath of the global financial crisis, DSGE modeling remains by far the most productive branch of macroeconomic modeling at this time. It takes on board elements from behavioral economics and other fields. Model comparison techniques help create standards that make it possible to compare models based on different paradigms. Thus, they support a more pluralistic yet rigorous approach to research in macroeconomics.

ACKNOWLEDGMENTS

The chapter has benefited from helpful discussions with and comments by Thorsten Drautzburg, Simon Gilchrist, Johannes Pfeifer, John B. Taylor, Harald Uhlig, as well as participants at the Handbook of Macroeconomics Conference in Chicago, the Macrofinancial Modeling Winter Meeting 2015 of the Becker Friedman Institute in New York, MACFINROBODS Annual Conference in Paris, the 10th Workshop on Macroeconomics and the Business Cycle at the Ifo Institute in Dresden, and AEA Meetings in San Francisco. E.A. gratefully acknowledges support from the FP7 Research and Innovation Funding Program (grant FP7-SSH-2013-2). J.Y. and M.K. gratefully acknowledge support from the Research Center SAFE, funded by the State of Hessen initiative for research LOEWE. J.Y. has since re-joined Bank of Korea. Of course, none of the views expressed in the chapter should be attributed to the Bank of Korea.

REFERENCES

Adjemian, S., Bastani, H., Juillard, M., Karame, F., Mihoubi, F., Perendia, G., Pfeifer, J., Ratto, M., Villemot, S., 2011. Dynare: reference manual, version 4. Working Papers 1, Dynare.

Adolfson, M., Laseen, S., Linde, J., Villani, M., 2007. Bayesian estimation of an open economy DSGE model with incomplete pass-through. J. Int. Econ. 72, 481–511.

Altig, D.E., Christiano, L.J., Eichenbaum, M., Linde, J., 2005. Firm-specific capital, nominal rigidities and the business cycle. Discussion Paper 4858, CEPR.

Anderson, G.S., Moore, G., 1985. A linear algebraic procedure for solving linear perfect foresight models. Econ. Lett. 17 (3), 247–252.

Ball, L., 1999. Policy rules for open economies. In: Taylor, J.B. (Ed.), Monetary Policy Rules. University of Chicago Press, Chicago, IL.

Bernanke, B., Gertler, M., Gilchrist, S., 1999. The financial accelerator in a quantitative business cycles framework. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics. In: vol. 1C. Elsevier Science, Amsterdam.

Bryant, R., Henderson, D.W., Holtham, G., Hooper, P., Symansky, S.A., 1988. Empirical Macroeconomics for Interdependent Economies. The Brookings Institution, Washington, DC.

Bryant, R., Currie, D., Frenkel, J., Masson, P., Portes, R., 1989. Macroeconomic Policies in an Interdependent World. The Brookings Institution, Washington, DC.

Bryant, R., Hooper, P., Mann, C., 1993. Evaluating Policy Regimes: New Research in Empirical Macroeconomics. The Brookings Institution, Washington, DC.

Carabenciov, I., Ermolaev, I., Freedman, C., Juillard, M., Kamenik, O., Korshunov, D., Laxton, D., 2008. A small quarterly projection model of the US economy. Tech. Rep. 08/278, IMF.

Christensen, I., Dib, A., 2008. The financial accelerator in an estimated new Keynesian model. Rev. Econ. Dyn. 11, 155–178.

Christiano, L.J., 2002. Solving dynamic equilibrium models by a method of undetermined coefficients. Comput. Econ. 20 (1-2), 21–55.

Christiano, L.J., Eichenbaum, M., Evans, C.L., 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. J. Polit. Econ. 113 (1), 1–45.

Christiano, L., Eichenbaum, M., Rebelo, S., 2011. When is the government spending multiplier large? J. Polit. Econ. 119, 78–121.

Christiano, L.J., Motto, R., Rostagno, M., 2014. Risk shocks. Am. Econ. Rev. 104 (1), 27–65.

Christoffel, K., Kuester, K., Linzert, T., 2009. The role of labor markets for euro area monetary policy. Eur. Econ. Rev. 53, 908–936.

Claerbout, J., 1994. Hypertext documents about reproducible research. http://sepwww.stanford.edu/sep/jon/nrc.html.

Coenen, G., Erceg, C.J., Freedman, C., Furceri, D., Kumhof, M., Lalonde, R., Laxton, D., Linde, J., Mourougane, A., Muir, D., Mursula, S., de Resende, C., Roberts, J., Roeger, W., Snudden, S., Trabandt, M., in't Veld, J., 2012. Effects of fiscal stimulus in structural models. Am. Econ. J. Macroecon. 4, 22–68.

Cogan, J., Cwik, T., Taylor, J., Wieland, V., 2010. New Keynesian versus old Keynesian government spending multipliers. J. Econ. Dyn. Control 34, 281–295.

Cogan, J., Taylor, J., Wieland, V., Wolters, M., 2013. Fiscal consolidation strategy. J. Econ. Dyn. Control 37, 404–421.

Collard, F., Juillard, M., 2001. Accuracy of stochastic perturbation methods: the case of asset pricing models. J. Econ. Dyn. Control. 25, 979–999.

Cwik, T., Wieland, V., 2011. Keynesian government spending multipliers and spillovers in the Euro area. Econ. Policy 26, 493–549.

De Graeve, F., 2008. The external finance premium and the macroeconomy: US post-WWII evidence. J. Econ. Dyn. Control. 32, 3415–3440.

Donoho, D., 2010. An invitation to reproducible computational research. Biostatistics 11 (3), 385–388.

Drautzburg, T., Uhlig, H., 2015. Fiscal stimulus and discretionary taxation. Rev. Econ. Dyn. 18 (4), 894–920.

Fair, R.C., Taylor, J.B., 1983. Solution and maximum likelihood estimation of dynamic nonlinear rational expectations models. Econometrica 51, 1169–1185.

Fomel, S., Claerbout, J.F., 2009. Reproducible research. Comput. Sci. Eng. 2009 (1), 5–7.

Frankel, J., Rockett, K., 1988. International macroeconomic policy coordination when policymakers do not agree on the true model. Am. Econ. Rev. 78, 318–340.

Freire, J., Bonnet, P., Shasha, D., 2012. Computational reproducibility: state-of-the-art, challenges, and database research opportunities. In: Proceedings of SIGMOD, pp. 593–596.

Fuhrer, J.C., 1997. Inflation/output variance trade-offs and optimal monetary policy. J. Money Credit Bank. 29 (2), 214–234.

Gelain, P., 2010. The external finance premium in the euro area: a dynamic stochastic general equilibrium analysis. N. Am. J. Econ. Finance 21, 49–71.

Gerali, A., Neri, S., Sessa, L., Signoretti, F.M., 2010. Credit and banking in a DSGE model of the euro area. J. Money Credit Bank. 42 (s1), 107–141.

Gerke, R., Jonsson, M., Kliem, M., Kolasa, M., Lafourcade, P., Locarno, A., Makarski, K., McAdam, P., 2013. Assessing macro-financial linkages: a model comparison exercise. Econ. Model. 31 (C), 253–264.

Gertler, M., Karadi, P., 2011. A model of unconventional monetary policy. J. Monet. Econ. 58 (1), 17–34.

Guerrieri, L., Iacoviello, M., Covas, F.B., Driscoll, J.C., Kiley, M.T., Jahan-Parwvar, M., Olive, A.Q., Sim, J.W., 2015. Macroeconomic effects of banking sector losses across structural models. Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series 2015-044.

Holmstrom, B., Tirole, J., 1997. Financial intermediation, loanable funds, and the real sector. Q. J. Econ. 112 (3), 663–691.

Hughes-Hallett, A., 1989. Empirical macroeconomics for interdependent economies: book review. J. Int. Econ. 26 (1-2), 189–194.

Hughes-Hallett, A., Wallis, K.F., 2004. EMU Macroeconomic Model Comparison excercise for the Euro-conference 7-8 June 2002. vol. 21, p. 5.

Iacoviello, M., 2005. House prices, borrowing constraints, and monetary policy in the business cycle. Am. Econ. Rev. 95 (3), 739–764.

Iacoviello, M., Neri, S., 2010. Housing market spillovers: evidence from an estimated DSGE model. Am. Econ. J. Macroecon. 2 (2), 125–164.

Ireland, P., 2003. Endogenous money or sticky prices? J. Monet. Econ. 50, 1623–1648.

Ireland, P., 2011. A new Keynesian perspective on the great recession. J. Money Credit Bank. 43 (1), 31–54.

Judd, K., 1998. Numerical Methods in Economics. MIT Press, Cambridge, MA.

Juillard, M., 2001. Dynare: a program for the simulation of rational expectation models. Computing in Economics and Finance 213.

Kannan, P., Rabanal, P., Scott, A.M., 2012. Monetary and macroprudential policy rules in a model with house price booms. B.E. J. Macroecon. 12 (1), 16.

Kilponen, J., Pisani, M., Schmidt, S., Corbo, V., Hledik, T., Hollmayr, J., Hurtado, S., Julio, P., Lozej, M., Landfall, H., Maria, J.R., Micallef, B., Papageorgiou, D., Rysanek, J., Sideris, D., Thomas, C., De Walque, G., 2015. Comparing fiscal multipliers across models and countries in Europe. Working Paper Series 1760, European Central Bank.

Kiyotaki, N., Moore, J., 1997. Credit cycles. J. Polit. Econ. 105 (2), 211–248.

Klein, L., 1991. Comparative Performance of U.S. Econometric Models. Oxford University Press, Oxford, UK.

Klein, P., 2000. Using the generalized Schur form to solve a multivariate linear rational expectations model. J. Econ. Dyn. Control 24 (10), 1405–1423.

Kuester, K., Wieland, V., 2010. Insurance policies for monetary policy in the Euro area. J. Eur. Econ. Assoc. 8 (4), 872–912.

Levin, A., Wieland, V., Williams, J.C., 1999. Robustness of simple monetary policy rules under model uncertainty. In: Taylor, J.B. (Ed.), Monetary Policy Rules. University of Chicago Press, Chicago, IL.

Levin, A., Wieland, V., Williams, J.C., 2003. The performance of forecast-based monetary policy rules under model uncertainty. Am. Econ. Rev. 93 (3), 622–645.

Mankiw, N.G., Reis, R., 2007. Sticky information in general equilibrium. J. Eur. Econ. Assoc. 5 (2-3), 603–613.

McCallum, B., Nelson, E., 1999. Performance of operational policy rules in an estimated semi-classical structural model. In: Taylor, J.B. (Ed.), Monetary Policy Rules. University of Chicago Press, Chicago, IL.

Meh, C.A., Moran, K., 2010. The role of bank capital in the propagation of shocks. J. Econ. Dyn. Control 34 (3), 555–576.

Mortensen, D., Pissarides, C., 1994. Job creation and job desctruction in the theory of unemployment. Rev. Econ. Stud. 61 (3), 397–415.

Orphanides, A., 2003. The quest for prosperity without inflation. J. Monet. Econ. 50, 633–663.

Orphanides, A., Wieland, V., 2008. Economic projections and rules of thumb for monetary policy. Fed. Reserve Bank St. Louis Rev. 90 (4), 307–324.

Orphanides, A., Wieland, V., 2013. Complexity and monetary policy. J. Int. Cent. Bank. 9 (1), 167–204.

Quint, D., Rabanal, P., 2014. Monetary and macroprudential policy in an estimated DSGE model of the euro area. Int. J. Cent. Bank. 10 (2), 169–236.

Rabanal, P., 2007. Does inflation increase after a monetary policy tightening? answers based on a estimated DSGE model. J. Econ. Dyn. Control 31, 906–937.

Rabanal, P., 2009. Inflation differentials between Spain and the EMU: a DSGE perspective. J. Money Credit Bank. 41 (6), 1141–1166.

Ratto, M., Roeger, W., in't Veld, J., 2009. QUEST III: an estimated open-economy DSGE model of the euro area with fiscal and monetary policy. Econ. Model. 26 (1), 222–233.

Romer, C., Bernstein, J., 2009. The job impact of the American recovery and reinvestment plan.

Rotemberg, J.J., Woodford, M., 1997. An optimization-based econometric framework for the evaluation of monetary policy. NBER Macroecon. Annu. 12, 297–346.

Rudebusch, G.D., Svensson, L.E.O., 1999. Policy rules for inflation targeting. In: Taylor, J.B. (Ed.), Monetary Policy Rules. University of Chicago Press, Chicago, IL.

Sandve, G.K., Nekrutenko, A., Taylor, J., Hovig, E., 2013. Ten simple rules for reproducible computational research. PLoS Comput. Biol. 9 (10), 1–4.

Schmidt, S., Wieland, V., 2013. The new Keynesian approach to dynamic general equilibrium modeling: models, methods and macroeconomic policy evaluation. In: Dixon, P.B., Jorgenson, D.W. (Eds.), Handbook of Computable General Equilibrium Modeling. North Holland, Amsterdam.

Sims, C., 2001. Solving linear rational expectations models. J. Comput. Econ. 20 (1-2), 1–20.

Slobodyan, S., Wouters, R., 2012. Learning in an estimated medium-scale DSGE model. J. Econ. Dyn. Control.

Smets, F., Wouters, R., 2003. An estimated dynamic stochastic general equilibrium model of the Euro area. J. Eur. Econ. Assoc. 1 (5), 1123–1175.

Smets, F., Wouters, R., 2007. Shocks and frictions in US business cycles: a bayesian DSGE approach. Am. Econ. Rev. 97 (3), 586–606.

Taylor, J.B., 1993. Discretion versus policy rules in practice. Carn.-Roch. Conf. Ser. Public Policy 39, 195–214.

Taylor, J.B., 1993. Macroeconomic Policy in a World Economy. W.W. Norton, New York, NY. http://www.stanford.edu/johntayl/MacroPolicyWorld.htm.

Taylor, J.B., 1999. Monetary Policy Rules. The University of Chicago Press, Chicago, IL.

Taylor, J.B., Wieland, V., 2012. Surprising comparative properties of monetary models: results from a new data base. Rev. Econ. Stat. 94 (3), 800–816.

Uhlig, H., 1995. A toolkit for analyzing nonlinear dynamic stochastic models easily. Discussion Paper 97, Tilburg University, Center for Economic Research.

Wallis, K.F., 2004. Comparing empirical models of the euro area economy. Econ. Model. 21 (5), 735–758.

White, W.W., 1978. Computers and mathematical programming. In: Proceedings of the Bicentennial Conference on Mathematical Programming Held at the National Bureau of Standards, Gaithersburg, Maryland, November 29-December 1, 1976, Band 13, U.S. Government Printing Office, Washington.

Wieland, V., Wolters, M., 2013. Forecasting and policy making. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. North Holland, Amsterdam.

Wieland, V., Cwik, T., Mueller, G.J., Schmidt, S., Wolters, M., 2012. A new comparative approach to macroeconomic modeling and policy analysis. J. Econ. Behav. Organ. 83 (3), 523–541.

Volume 2B

SECTION 3

Financial-Real Connections

CHAPTER 16

Wholesale Banking and Bank Runs in Macroeconomic Modeling of Financial Crises

M. Gertler*[,†,‡], N. Kiyotaki*[,†,‡], A. Prestipino*[,†,‡]
*NYU, New York, NY, United States
†Princeton University, Princeton, NJ, United States
‡Federal Reserve Board of Governors, Washington, DC, United States

Contents

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.009

Abstract

There has been considerable progress in developing macroeconomic models of banking crises. However, most of this literature focuses on the retail sector where banks obtain deposits from households. In fact, the recent financial crisis that triggered the Great Recession featured a disruption of wholesale funding markets, where banks lend to one another. Accordingly, to understand the financial crisis as well as to draw policy implications, it is essential to capture the role of wholesale banking. The objective of this chapter is to characterize a model that can be seen as a natural extension of the existing literature, but in which the analysis is focused on wholesale funding markets. The model accounts for both the buildup and collapse of wholesale banking and also sketches out the transmission of the crises to the real sector. We also draw out the implications of possible instability in the wholesale banking sector for lender-of-last resort policy as well as for macroprudential policy.

Keywords

Financial crises, Wholesale banking, Interbank markets, Rollover risk

JEL Classification Code

E44

1. INTRODUCTION

One of the central challenges for contemporary macroeconomics is adapting the core models to account for why the recent financial crisis occurred and for why it then devolved into the worst recession of the postwar period. On the eve of the crisis, the basic workhorse quantitative models used in practice largely abstracted from financial market frictions. These models were thus largely silent on how the crisis broke out and how the vast array of unconventional policy interventions undertaken by the Federal Reserve and Treasury could have worked to mitigate the effects of the financial turmoil. Similarly, these models could not provide guidance for the regulatory adjustments needed to avoid another calamity.[a]

From the start of the crisis there has been an explosion of literature aimed at meeting this challenge. Much of the early wave of this literature builds on the financial accelerator and credit cycle framework developed in Bernanke and Gertler (1989) and Kiyotaki and Moore (1997). This approach stresses the role of balance sheets in constraining borrower spending in a setting with financial market frictions. Procyclical movement in balance sheet strength amplifies spending fluctuations and thus fluctuations in aggregate economic activity. A feedback loop emerges as conditions in the real economy affect the condition of balance sheets and vice-versa. Critical to this mechanism is the role of leverage: The exposure of balance sheets to systemic risk is increasing in the degree of borrower leverage.

[a] For a description of the causes leading to the recent financial crisis see Bernanke (2010).

The new vintage of macroeconomic models with financial frictions makes progress in two directions: First, it adapts the framework to account for the distinctive features of the current crisis. In particular, during the recent crisis, it was highly leveraged financial institutions along with highly leveraged households that were most immediately vulnerable to financial distress.[b] The conventional literature featured balance sheet constraints on nonfinancial firms. Accordingly, a number of recent macroeconomic models have introduced balance sheet constraints on banks, while others have done so for households.[c] The financial accelerator remains operative, but the classes of agents most directly affected by the financial market disruption differ from earlier work.

Another direction has involved improving the way financial crises are modeled. For example, financial crises are inherently nonlinear events, often featuring a simultaneous sudden collapse in asset prices and rise in credit spreads.[d] A sharp collapse in output typically ensues. Then recovery occurs only slowly, as it is impeded by a slow process of deleveraging. A number of papers have captured this nonlinearity by allowing for the possibility that the balance sheet constraints do not always bind.[e] Financial crises are then periods where the constraints bind, causing an abrupt contraction in economic activity. Another approach to handling the nonlinearity is to allow for bank runs.[f] Indeed, runs on the shadow banking system were a salient feature of the crisis, culminating with the collapse in September 2008 of Lehman Brothers, of some major money market funds and ultimately of the entire investment banking sector. Yet another literature captures the nonlinearity inherent in financial crises by modeling network interactions (see, eg, Garleanu et al., 2015).

One area the macroeconomics literature has yet to address adequately is the distinctive role of the wholesale banking sector in the breakdown of the financial system. Our notion of wholesale banks corresponds roughly, though not exactly, to the shadow banking sector on the eve of the 2007–09 financial crisis. Shadow banking includes all financial intermediaries that operated outside the Federal Reserve's regulatory framework. By wholesale banking, we mean the subset that (i) was highly leveraged, often with

[b] To be sure, the financial distress also directly affected the behavior of nonfinancial firms. See Giroud and Mueller (2015) for evidence of firm balance sheet effects on employment during the crisis.

[c] See Gertler and Karadi (2011), Gertler and Kiyotaki (2010), and Curdia and Woodford (2010) for papers that incorporate banking and Iacoviello (2005), Eggertsson and Krugman (2012), Guerrieri and Lorenzoni (2011), and Midrigan and Philippon (2011) for papers that included household debt.

[d] See He and Krishnamurthy (2014) for evidence in support of the nonlinearity of financial crises.

[e] See Brunnermeier and Sannikov (2014), He and Krishnamurthy (2013), He and Krishnamurthy (2014), and Mendoza (2010).

[f] For the seminal contribution on bank runs see Diamond and Dybvig (1983). Some recent examples of macroeconomic models that consider bank runs include Gertler and Kiyotaki (2015), Ferrante (2015a), Robatto (2014), Martin et al (2014), Angeloni and Faia (2013) and Ennis and Keister (2003). See Boissay, Collard, and Smets (2013) for an alternative way to model banking crises that does not involve runs per se. For other related literature see Allen and Gale (2007), Cooper and Ross (1998), Farmer (1999), Holmstrom and Tirole (2011) and the references within.

short-term debt and (ii) relied heavily on borrowing from other financial institutions in "wholesale" markets, as opposed to borrowing from households in "retail" markets for bank credit.

When the crisis hit, the epicenter featured malfunctioning of the wholesale banking sector. Indeed, retail markets remained relatively stable while wholesale funding markets experienced dry-ups and runs. By contrast, much of the macroeconomic modeling of banking features traditional retail banking. In this respect, it misses some important dimensions of both the run-up to the crisis and how exactly the crisis played out. In addition, by omitting wholesale banking, the literature may be missing some important considerations for regulatory design.

In this Handbook chapter, we present a simple canonical macroeconomic model of banking crises that (i) is representative of the existing literature and (ii) extends this literature to feature a role for wholesale banking. The model will provide some insight both into the growth of wholesale banking and into how this growth led to a build-up of financial vulnerabilities that ultimately led to a collapse. Because the model builds on existing literature, our exposition of the framework will permit us to review the progress that is made. However, by turning attention to wholesale banks and wholesale funding markets, we are able to chart a direction we believe the literature should take.

In particular, the model is an extension of the framework developed in Gertler and Kiyotaki (2011), which had a similar twofold objective: first, present a canonical framework to review progress that has been made and, second, chart a new direction. That paper characterized how existing financial accelerator models that featured firm level balance sheet constraints could be extended to banking relationships in order to capture the disruption of banking during the crisis. The model developed there considered only retail banks which funded loans mainly from household deposits. While it allowed for an interbank market for credit among retail banks, it did not feature banks that relied primarily on wholesale funding, as was the case with shadow banks.

For this Handbook chapter, we modify the Gertler and Kiyotaki framework to incorporate wholesale banking alongside retail banking, where the amount credit intermediated via wholesale funding markets arises endogenously. Another important difference is that we allow for the possibility of runs on wholesale banks. We argue that both these modifications improve the ability of macroeconomic models to capture how the crisis evolved. They also provide insight into how the financial vulnerabilities built up in the first place.

As way to motivate our emphasis on wholesale banking, Section 1 presents descriptive evidence on the growth of this sector and the collapse it experienced during the Great Recession. Section 3 presents the baseline macroeconomic model with banking, where a wholesale banking sector arises endogenously. Sector 4 conducts a set of numerical experiments. While the increased size of the wholesale banking improves the efficiency of financial intermediation, it also raises the vulnerability of this sector to runs. Section 5 considers the case where runs in the wholesale sector might be anticipated. It illustrates how the model can capture some of the key phases of the financial collapse, including the slow run period up to Lehman and the ultimate "fast run" collapse. In Section 6, we introduce a second asset in which retail banks have a comparative advantage in intermediating. We then show how a crisis in wholesale banking can spill over and affect retail banking, consistent with what happened during the crisis. Section 7 analyzes government policy to contain financial crises, including both ex-post lender of last resort activity and ex-ante macroprudential regulation. Finally, we conclude in Section 8 with some directions for future research.

2. THE GROWTH AND FRAGILITY OF WHOLESALE BANKING

In this section, we provide some background motivation for the canonical macroeconomic model with wholesale funding markets that we develop in the following section. We do so by presenting a brief description of the growth and ultimate collapse of wholesale funding markets during the Great Recession. We also describe informally how the disruption of these markets contributed to the contraction of the real economy.

Fig. 1 illustrates how we consider the different roles of retail and wholesale financial intermediaries, following the tradition of Gurley and Shaw (1960).[g] The arrows indicate

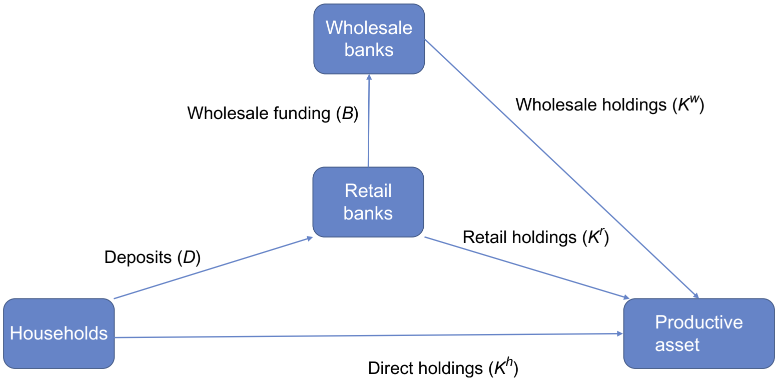


Fig. 1 Modes of financial intermediation.

[g] Gurley and Shaw (1960) consider that there are two ways to transfer funds from ultimate lenders (with surplus funds) to ultimate borrowers (who need external funds to finance expenditure): direct and indirect finance. In direct finance, ultimate borrowers sell their securities directly to ultimate lenders to raise funds. In indirect finance, financial intermediaries sell their own securities to raise funds from ultimate lenders in order to buy securities from ultimate borrowers. By doing so, financial intermediaries transform relatively risky, illiquid, and long maturity securities of ultimate borrowers into relatively safe, liquid, and short maturity securities of intermediaries. Here, we divide financial intermediaries into wholesale and retail financial intermediaries, while both involve asset transformation of risk, liquidity, and maturity. We refer to intermediaries as "banks" and to ultimate lenders as "households" for short.

the direction that credit is flowing. Funds can flow from households (ultimate lenders) to nonfinancial borrowers (ultimate borrowers) through three different paths: they can be lent directly from households to borrowers $\left(K^h\right)$; they can be intermediated by retail banks that raise deposits (D) from households and use them to make loans to nonfinancial borrowers $\left(K^r\right)$; alternatively, lenders' deposits can be further intermediated by specialized financial institutions that raise funds from retail banks in wholesale funding markets (B) and, in turn, make loans to ultimate borrowers $\left(K^w\right)$. In what follows we refer to these specialized financial institutions as wholesale banks. We think of wholesale banks as highly leveraged shadow banks that rely heavily on credit from other financial institutions, particularly short-term credit. We place in this category institutions that financed long-term assets, such as mortgaged back securities, with short-term money market instruments, including commercial paper and repurchase agreements. Examples of these kinds of financial institutions are investment banks, hedge funds, and conduits. We focus attention on institutions that relied heavily on short-term funding in wholesale markets to finance longer term assets because it was primarily these kinds of entities that experienced financial turmoil.

Our retail banking sector, in turn, includes financial institutions that rely mainly on household saving for external funding and provide a significant amount of short-term financing to the wholesale banks. Here, we have in mind commercial banks, money market funds, and mutual funds that raised funds mainly from households and on net provided financing to wholesale banks.

Fig. 1 treats wholesale banking as if it is homogenous. In order to understand how the crisis spread, it is useful to point out that there are different layers within the wholesale banking sector. While the intermediation process was rather complex, conceptually we can reduce the number of layers to three basic ones: (1) origination, (2) securitization, (3)

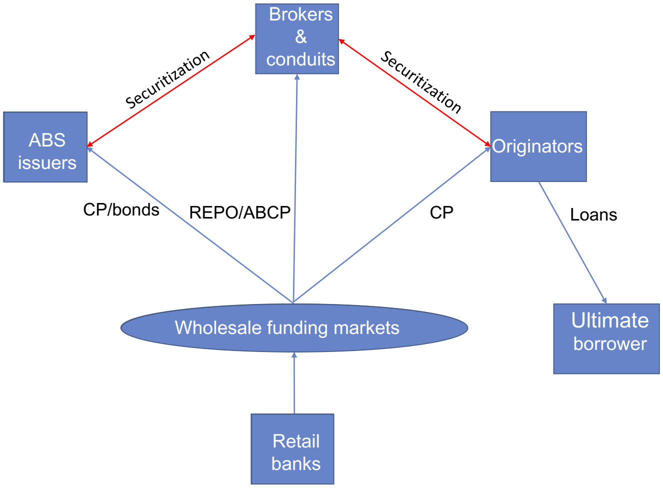


Fig. 2 Wholesale intermediation.

and funding. Fig. 2 illustrates the chain. First there are "loan originators," such as mortgage origination companies and finance companies, that made loans directly to nonfinancial borrowers. At the other end of the chain were shadow banks that held securitized pools of the loans made by originators. In between were brokers and conduits that assisted in the securitization process and provided market liquidity. Dominant in this group were the major investment banks (eg, Goldman Sachs, Morgan Stanley, and Lehman Brothers). Each of these layers relied on short-term funding, including commercial paper, asset-backed commercial paper and repurchase agreements. While there was considerable interbank lending among wholesale banks, retail banks (particularly money market funds) on net provided short-term credit in wholesale credit markets.

We next describe a set of facts about wholesale banking. We emphasize three sets of facts in particular: (1) wholesale banking grew in relative importance over the last four decades, (2) leading up to the crisis wholesale banks were highly exposed to systemic risk because they were highly leveraged and relied heavily on short-term debt, and (3) the subsequent disruption of wholesale funding markets raised credit costs and contracted credit flows, likely contributing in a major way to the Great Recession.

1. *Growth in Wholesale Banking*

We now present measures of the scale of wholesale banking relative to retail banking as well as to household's direct asset holdings. Table 1 describes how we construct measures of assets held by wholesale vs retail banks. In particular, it lists how we categorized the various types of financial intermediaries into wholesale vs retail banking.[h,i] As the table

Table 1 Wholesale and Retail sector in the Flow of Funds

Retail sector	Private depository institutions Money market mutual funds Mutual funds	
Wholesale sector	*Origination*	Finance companies Real estate investment trusts Government sponsored enterprises
	Securitization	Security brokers dealers ABS issuers
	Funding	GSE mortgage pools Funding corporations Holding companies

[h] Appendix D provides details about measurement of the time series shown in this section from Flow of Funds data.

[i] It is important to notice that the measures we report are broadly in line with analogous measures computed for shadow banking. See, eg, Adrian and Ashcraft (2012), for an alternative definition of shadow banking that yields very similar conclusions and Pozsar et al. (2013), for a detailed description of shadow banking.

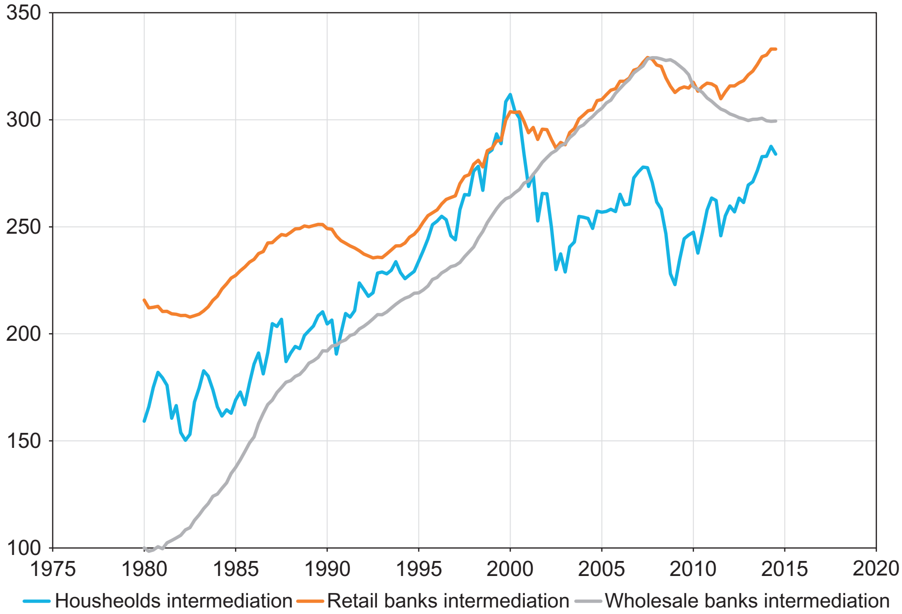


Fig. 3 Intermediation by sector. The graph shows the evolution of credit intermediated by the three different sectors. Nominal data from the Flow of Funds are deflated using the CPI and normalized so that the log of the normalized value of real wholesale intermediation in 1980 is equal to 1. The resulting time series are then multiplied by 100.

indicates, the wholesale banking sector aggregates financial institutions that originate loans, that help securitize them and that ultimately fund them. A common feature of all these institutions, though, is that they relied heavily on short-term credit in wholesale funding markets.

Fig. 3 portrays the log level of credit to nonfinancial sector provided by wholesale banks, by retail banks, and directly by households from the early 1980s until the present.[j] The figure shows the rapid increase in wholesale banking relative to the other means of credit supply to nonfinancial sector. Wholesale banks went from holding under 15% of total credit in the early 1980s to roughly 40% on the eve of the Great Recession, an amount on par with credit provided by retail banks.

Two factors were likely key to the growth of wholesale banking. The first is regulatory arbitrage. Increased capital requirements on commercial banks raised the incentive to transfer asset holding outside the commercial bank system. Second, financial innovation improved the liquidity of wholesale funding markets. The securitization process in particular improved the (perceived) safety of loans by diversifying idiosyncratic risks as

[j] The measure we present also include nonfinancial corporate equities. Excluding equities, households would become negligible but the relative size of wholesale and retail banks would evolve very similarly. See Appendix D for details on how we construct the measures reported.

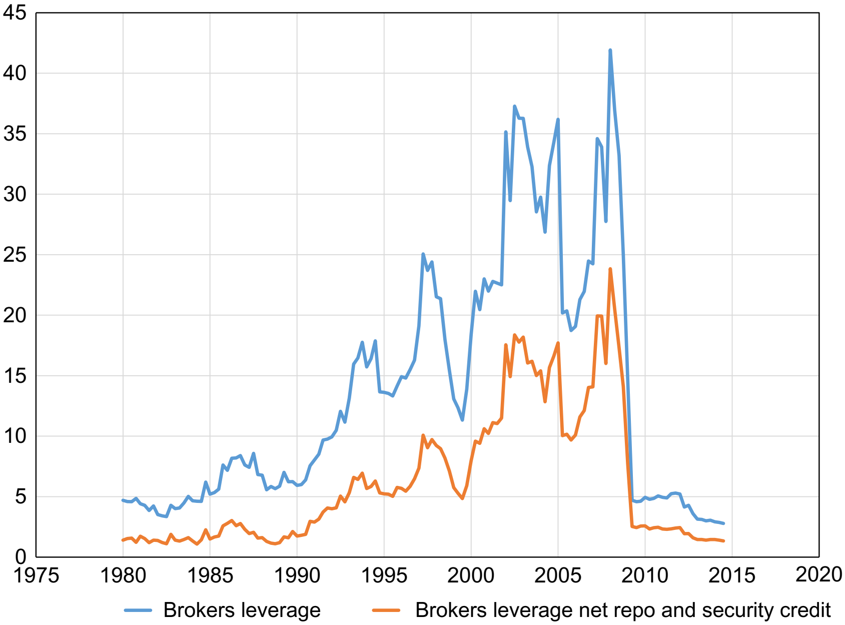


Fig. 4 Brokers leverage. Leverage is given by the ratio of total financial assets over equity. Equity is computed from the Flow of Funds by subtracting total financial liabilities from total financial assets. The net position leverage computes assets by netting out long and short positions in REPO and Security Credit. See the Appendices for details.

well as by enhancing the liquidity of secondary markets for bank assets. The net effect was to raise the borrowing capacity of the overall financial intermediary sector.

2. *Growth in Leverage and Short-Term Debt in Wholesale Banking*

Wholesale banking not only grew rapidly, it also became increasingly vulnerable to systemic disturbances. Fig. 4 presents evidence on the growth in leverage in the investment banking sector. Specifically it plots the aggregate leverage multiple for broker dealers (primarily investment banks) from 1980 to the present. We define the leverage multiple as the ratio of total assets held to equity.[k] The greater is the leverage multiple, the higher is the reliance on debt finance relative to equity. The key takeaway from Fig. 4 is that the leverage multiple grew from under five in the early 1980s to over forty at the beginning of the Great Recession, a nearly tenfold increase.

Arguably, the way securitization contributed to the overall growth of wholesale banking was by facilitating the use of leverage. By constructing assets that appeared safe and liquid, securitization permitted wholesale banks to fund these assets by issuing debt.

[k] The data is from the Flow of Funds and equity is measured by book value. We exclude nonfinancial assets from measurement as they are not reported in the Flow of Funds.

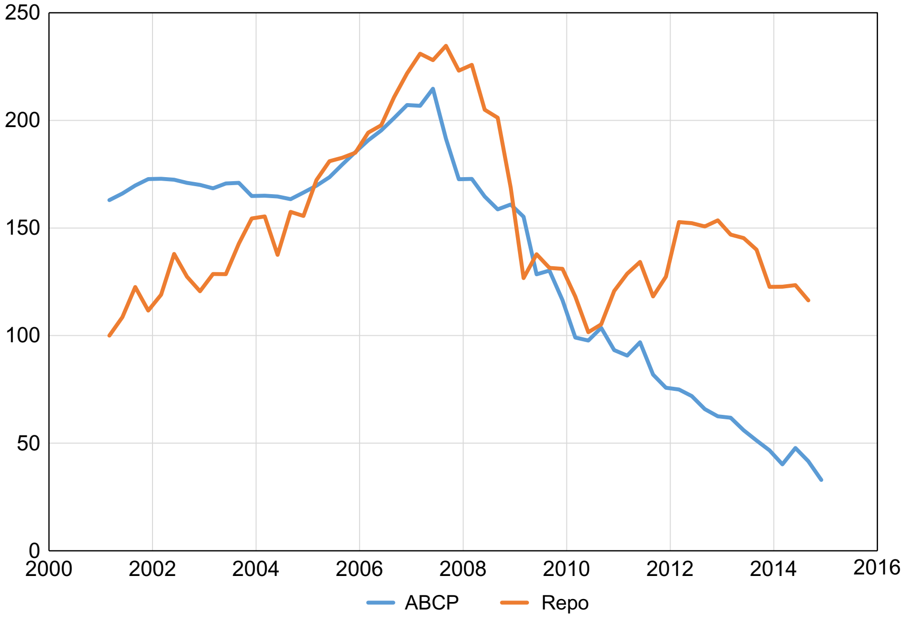


Fig. 5 Short-term wholesale funding. The graph shows the logarithm of the real value outstanding. Nominal values from Flow of Funds are deflated using the CPI.

At a minimum debt finance had the advantage of being cheaper due to the tax treatment. Debt financing was also cheaper to the extent the liabilities were liquid and thus offered a lower rate due to a liquidity premium.

Why were these assets funded in wholesale markets as opposed to retail markets? The sophistication of these assets required that creditors be highly informed to evaluate payoffs, especially given the absence of deposit insurance. The complicated asset payoff structure also suggests that having a close working relationship with borrowers is advantageous. It served to reduce the possibility of any kind of financial malfeasance. Given these considerations, it makes sense that wholesale banks obtain funding in interbank markets. In these markets lenders are sophisticated financial institutions as opposed to relatively unsophisticated households in the retail market.

Fig. 5 shows that much of the growth in leverage in wholesale banking involved short-term borrowing. The figure plots the levels of asset backed commercial paper (ABCP) and repurchase agreements (Repo). This growth reflected partly the growth in assets held by wholesale banks and partly innovation in loan securitization that made maturity transformation by wholesale banks more efficient. Also relevant, however, was a shift in retail investors demand from longer term security tranches towards short-term credit instruments as the initial fall in housing prices in 2006 raised concerns about the quality of existing securitized assets.[l,m] As we discuss next, the combination of high leverage and short-term debt is what made the wholesale banking system extremely fragile.

3. *The Crisis: The Unraveling of Wholesale Bank Funding Markets*

The losses suffered by mortgage originators due to falling housing prices in 2006 eventually created strains in wholesale funding markets. Short-term wholesale funding markets started experiencing severe turbulence in the summer of 2007. In July 2007 two Bear Sterns investment funds that had invested in subprime related products declared bankruptcy. Shortly after, BNP Paribas had to suspend withdrawals from investment funds with similar exposure. These two episodes led investors to reassess the risks associated with the collateral backing commercial paper offered by asset backed securities issuers. In August 2007 a steady contraction of Asset Backed Commercial Paper (ABCP) market began, something akin to a "slow run," in Bernanke's terminology.[n] The value of Asset Backed Commercial Paper outstanding went from a peak of 1.2 trillion dollars in July 2007 to 800 billion dollars in December of the same year and continued its descent to its current level of around 200 billion dollars.

The second significant wave of distress to hit wholesale funding markets featured the collapse of Lehman Brothers in September of 2008. Losses on short-term debt instruments issued by Lehman Brothers led the Reserve Primary Fund, a large Money Market Mutual Fund (MMMF), to "break the buck": the market value of assets fell below the value of its noncontingent liabilities. An incipient run on MMMFs was averted only by the extension of Deposit Insurance to these types of institutions. Wholesale investors,[o] however, reacted by pulling out of the Repo market, switching off the main source of funding for Security Broker Dealers. Fig. 5 shows the sharp collapse in repo financing around the time of the Lehman collapse. Indeed if the first wave of distress hitting the ABCP market had the features of a "slow run," the second, which led to the dissolution of the entire investment banking system had the features of a traditional "fast run." We emphasize that a distinctive feature of these two significant waves of financial distress is that they did not involve traditional banking institutions. In fact, the retail sector as a whole was shielded thanks to prompt government intervention that halted the run on

[l] See Brunnermeier and Oemke (2013) for a model in which investors prefer shorter maturities when release of information could lead them not to roll over debt.

[m] It is not easy to gather direct evidence on this from the aggregate composition of liabilities of wholesale banks since data from the Flow of Funds excludes the balance sheets of SIVs and CDOs from the ABS Issuers category. Our narrative is based on indirect evidence coming from ABX spreads as documented for example in Gorton (2009).

[n] Covitz et al. (2013) provide a detailed description of the run on ABCP programs in 2007. A very clear description of the role of commercial paper during the 2007–09 crisis is presented by Kacperczyk and Schnabl (2010).

[o] The poor quality of available data makes it difficult to exactly identify the identity of the investors running on Repo's. See Gorton and Metrick (2012) and Krishnamurthy et al. (2014).

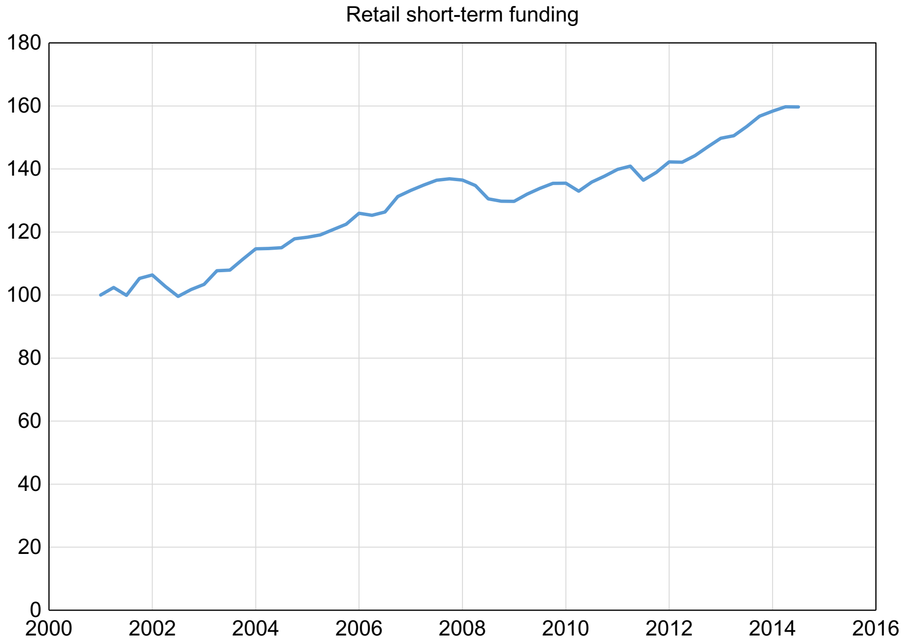


Fig. 6 Retail short-term funding. The graph shows the logarithm of the real value outstanding. Nominal values from Flow of Funds are deflated using the CPI and normalized so that the log of the normalized value of retail short-term funding in 2001 is equal to 100.

MMMFs in 2008 as well as the Troubled Asset Relief Program and other subsequent measures that supplemented the traditional safety net. In fact, total short-term liabilities of the retail sector were little affected overall (see Fig. 6). This allowed the retail banking sector to help absorb some of the intermediation previously performed by wholesale banks.

Despite the unprecedented nature and size of government intervention and the partial replacement of wholesale intermediation by retail bank lending, the distress in wholesale bank funding markets led to widespread deterioration in credit conditions. Fig. 7 plots the behavior of credit spreads and investment from 2004 to 2010. We focus on three representative credit spreads: (1) The spread between the 3 month ABCP rate and 3 month Treasury spread, (2) The financial company commercial paper spread, and (3) The Gilchrist and Zakrajsek (2012) excess bond premium. In each case, the spread is the difference between the respective rate on the private security and a similar maturity treasury security rate. The behavior of the spreads lines up with the waves of financial distress that we described. The ABCP spread jumps by 1.5% in August 2007, the beginning of the unraveling of this market. The increase in this spread implies a direct increase in credit costs for borrowing funded by ABCP including mortgages, car loans, and credit card

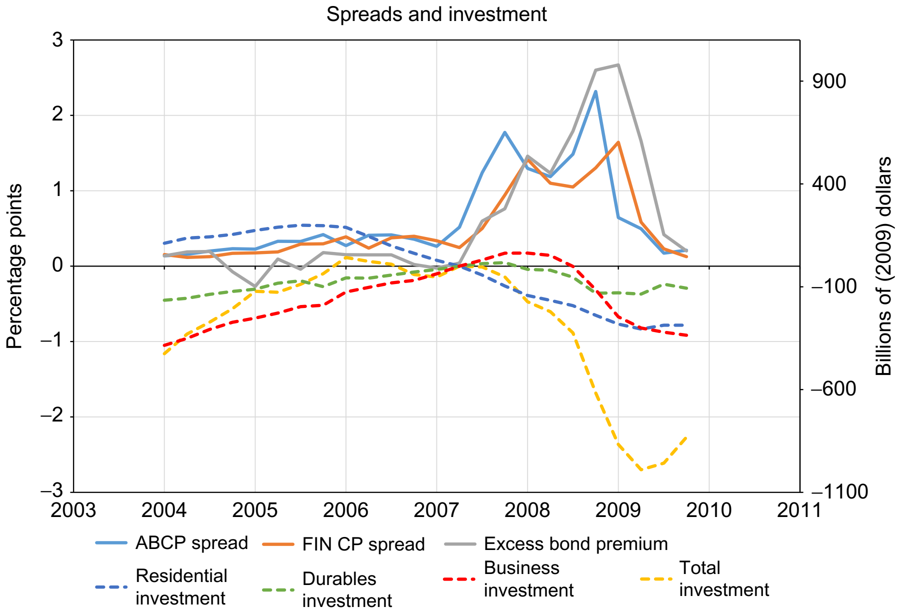


Fig. 7 Credit spreads and investment.

borrowing. As problems spread to broker dealers, the financial commercial paper spread increases reaching a peak at more than 1.5% at the time of the Lehman collapse. Increasing costs of credit for these intermediaries, in turn, helped fuel increasing borrowing costs for nonfinancial borrowers. The Gilchrist and Zakrajsek's corporate excess bond spread jumps more than 2.5% from early 2007 to the peak in late 2008.

It is reasonable to infer that the borrowing costs implied by the increased credit spreads contributed in an important way to the slowing of the economy at the onset of the recession in 2007:Q4, as well as to the sharp collapse following the Lehman failure. As shown in Fig. 7, the contraction in business investment, residential investment, durable consumption, and their sum-total investment, moves inversely with credit spreads.

In our view, there are three main conclusions to be drawn from the empirical evidence presented in this section. First, the wholesale banking sector grew into a very important component of financial intermediation by relying on securitization to reduce the risks of lending and expand the overall borrowing capacity of the financial system. Second, higher borrowing capacity came at the cost of increased fragility as high leverage made wholesale banks' net worth very sensitive to corrections in asset prices. Third, the disruptions in wholesale funding markets that took place in 2007 and 2008 seem to have played an important role in the unfolding of the Great Recession. These observations motivate our modeling approach below and our focus on interbank funding markets functioning and regulation.

3. BASIC MODEL

3.1 Key Features

Our starting point is the infinite horizon macroeconomic model with banking and bank runs developed in Gertler and Kiyotaki (2015). In order to study recent financial booms and crises, in this chapter we disaggregate banking into wholesale and retail banks. Wholesale banks make loans to the nonfinancial sector funded primarily by borrowing from retail banks. The latter use deposits from households to make loans both to the nonfinancial sector and to the wholesale financial sector. Further, the size of the wholesale banking market arises endogenously. It depends on two key factors: (1) the relative advantage wholesale banks have in managing assets over retail banks and (2) the relative advantage of retail banks over households in overcoming an agency friction that impedes lending to wholesale banks.[p]

In the previous section, we described the different layers of the wholesale sector, including origination, securitization, and funding. For tractability, in our model we consolidate these various functions into a single type of wholesale bank. Overall, our model permits capturing financial stress in wholesale funding markets which was a key feature of the recent financial crisis.

There are three classes of agents: households, retail banks, and wholesale banks. There are two goods, a nondurable good and a durable asset, "capital." Capital does not depreciate and the total supply of capital stock is fixed at $\bar{K}$. Wholesale and retail banks use borrowed funds and their own equity to finance the acquisition of capital. Households lend to banks and also hold capital directly. The sum of total holdings of capital by each type of agent equals the total supply which we normalize to unity:

$$K_t^w + K_t^r + K_t^h = \bar{K} = 1, \tag{1}$$

where K_t^w and K_t^r are the total capital held by wholesale and retail bankers and K_t^h is the amount held by households.

Agents of type j use capital and goods as inputs at t to produce output and capital at $t+1$, as follows:

$$\begin{array}{ccc} \textit{date } t & & \textit{date } t+1 \\ \left.\begin{array}{r} K_t^j \text{ capital} \\ F^j(K_t^j) \text{ goods} \end{array}\right\} & \rightarrow & \left\{\begin{array}{c} Z_{t+1}K_t^j \text{ output} \\ K_t^j \text{ capital} \end{array}\right. \end{array} \tag{2}$$

where type $j = w$, r, and h stands for wholesale banks, retail banks, and households, respectively. Expenditure in terms of goods at date t reflects the management cost of

[p] Our setup bears some resemblance to Holmstrom and Tirole (1997), which has nonfinancial firms that face costs in raising external funds from banks that in turn face costs in raising deposits from households. In our case, it is constrained wholesale banks that raise funds from constrained retail banks.

screening and monitoring investment projects. In the case of retail banks, the management costs might also reflect various regulatory constraints. We suppose this management cost is increasing and convex in the total amount of capital, as given by the following quadratic formulation:

$$F^j(K_t^j) = \frac{\alpha^j}{2}(K_t^j)^2. \tag{3}$$

In addition, we suppose the management cost is zero for wholesale banks and highest for households (holding constant the level of capital):

$$\alpha^w = 0 < \alpha^r < \alpha^h. \tag{Assumption 1}$$

This assumption implies that wholesale bankers have an advantage over the other agents in managing capital.[q] Retail banks in turn have a comparative advantage over households. Finally, the convex cost implies that it is increasingly costly at the margin for retail banks and households to absorb capital directly. As we will see, this cost formulation provides a simple way to limit agents with wealth but lack of expertise from purchasing assets during a firesale.

In our decentralization of the economy, a representative household provides capital management services both for itself and for retail banks. For the latter, the household charges retail banks a competitive price f_t^r per unit of capital managed, where f_t^r corresponds to the marginal cost of providing the service:

$$f_t^r = F^{r\prime}(K_t^r) = \alpha^r K_t^r. \tag{4}$$

Households obtain the profit from this activity $f_t^r K_t^r - F^r(K_t^r)$.

3.2 Households

Each household consumes and saves. Households save either by lending funds to bankers or by holding capital directly in the competitive market. They may deposit funds in either retail or wholesale banks. In addition to the returns on portfolio investments, every period each household receives an endowment of nondurable goods, $Z_t W^h$, that varies proportionately with the aggregate productivity shock Z_t.

Deposits held in a bank from t to $t+1$ are one period bonds that promise to pay the noncontingent gross rate of return $\bar{R}_{t+1}$ in the absence of a run by depositors. In the event of a deposit run, depositors only receive a fraction x_{t+1}^r of the promised return, where x_{t+1}^r

[q] In general, we have in mind that wholesale and retail banks specialize in different types of lending and, as a consequence, each has developed relative expertise in managing the type of assets they hold. We subsequently make this point clearer by introducing a second asset in which retail banks have a comparative advantage in intermediating. Also relevant are regulatory distortions, though we view this as a factor that leads to specialization in the first place.

is the total liquidation value of retail banks assets[r] per unit of promised deposit obligations. Accordingly, we can express the household's return on deposits, R_{t+1}, as follows:

$$R_{t+1} = \begin{cases} \bar{R}_{t+1} & \text{if no deposit run} \\ x^r_{t+1}\bar{R}_{t+1} & \text{if deposit run occurs} \end{cases} \tag{5}$$

where $0 \leq x^r_t < 1$. Note that if a deposit run occurs all depositors receive the same pro rata share of liquidated assets.

Household utility U_t is given by

$$U_t = E_t\left(\sum_{i=0}^{\infty}\beta^i \ln C^h_{t+i}\right)$$

where C^h_t is household consumption and $0 < \beta < 1$. Let Q_t be the market price of capital. The household then chooses consumption, bank deposits D_t and direct capital holdings K^h_t to maximize expected utility subject to the budget constraint

$$C^h_t + D_t + Q_t K^h_t + F^h(K^h_t) = Z_t W^h + R_t D_{t-1} + (Z_t + Q_t)K^h_{t-1} + f^r_t K^r_t - F^r(K^r_t). \tag{6}$$

Here, consumption, saving, and management costs are financed by the endowment, the returns on savings, and the profits from providing management services to retail bankers.

For pedagogical purposes, we begin with a baseline model where bank runs are completely unanticipated events. Accordingly, in this instance the household chooses consumption and saving with the expectation that the realized return on deposits, R_{t+i}, equals the promised return, $\bar{R}_{t+i}$, with certainty, and that asset prices, Q_{t+i}, are those at which capital is traded when no bank run happens. In a subsequent section, we characterize the case where agents anticipate that a bank run may occur with some likelihood.

Given that the household assigns probability zero to a bank run, the first order condition for deposits is given by

$$E_t(\Lambda_{t,t+1})R_{t+1} = 1 \tag{7}$$

where the stochastic discount factor $\Lambda_{t,\tau}$ satisfies

$$\Lambda_{t,\tau} = \beta^{\tau-t}\frac{C^h_t}{C^h_\tau}.$$

The first order condition for direct capital holdings is given by

$$E_t\left(\Lambda_{t,t+1}R^h_{kt+1}\right) = 1 \tag{8}$$

with

$$R^h_{kt+1} = \frac{Q_{t+1} + Z_{t+1}}{Q_t + F^{h\prime}(K^h_t)}$$

where $F^{h\prime}(K^h_t) = \alpha^h K^h_t$ and R^h_{t+1} is the household's gross marginal rate of return from direct capital holdings.

[r] Under our calibration only retail banks choose to issue deposits. See later.

3.3 Banks

There are two types of bankers, retail and wholesale. Each type manages a financial intermediary. Bankers fund capital investments (which we will refer to as "nonfinancial loans") by issuing deposits to households, borrowing from other banks in an interbank market and using their own equity, or net worth. Banks can also lend in the interbank market.

As we describe later, bankers may be vulnerable to runs in the interbank market. In this case, creditor banks suddenly decide to not rollover interbank loans. In the event of an interbank run, the creditor banks receive a fraction x_{t+1}^{w} of the promised return on the interbank credit, where x_{t+1}^{w} is the total liquidation value of debtor bank assets per unit of debt obligations. Accordingly, we can express the creditor bank's return on interbank loans, R_{bt+1}, as follows:

$$R_{bt+1} = \begin{cases} \bar{R}_{bt+1} & \text{if no interbank run} \\ x_{t+1}^{w}\bar{R}_{bt+1} & \text{if interbank run occurs} \end{cases} \tag{9}$$

where $0 \leq x_t^w < 1$. If an interbank run occurs, all creditor banks receive the same pro rata share of liquidated assets. As in the case of deposits, we continue to restrict attention to the case where bank runs are completely unanticipated, before turning in a subsequent section to the case of anticipated runs in wholesale funding markets.

Due to financial market frictions that we specify below, bankers may be constrained in their ability to raise external funds. To the extent they may be constrained, they will attempt to save their way out of the financing constraint by accumulating retained earnings in order to move toward 100% equity financing. To limit this possibility, we assume that bankers have a finite expected lifetime: Specifically, each banker of type j (where $j = w$ and r for wholesale and retail bankers) has an i.i.d. probability σ^j of surviving until the next period and a probability $1 - \sigma^j$ of exiting. This setup provides a simple way to motivate "dividend payouts" from the banking system in order to ensure that banks use leverage in equilibrium.

Every period new bankers of type j enter with an endowment w^j that is received only in the first period of life. This initial endowment may be thought of as the start up equity for the new banker. The number of entering bankers equals the number who exit, keeping the total constant.

We assume that bankers of either type are risk neutral and enjoy utility from consumption in the period they exit. The expected utility of a continuing banker at the end of period t is given by

$$V_t^j = E_t\left[\sum_{i=1}^{\infty} \beta^i (1-\sigma^j)(\sigma^j)^{i-1} c_{t+i}^j\right],$$

where $(1 - \sigma^j)(\sigma^j)^{i-1}$ is the probability of exiting at date $t + i$, and c_{t+i}^j is terminal consumption if the banker of type j exits at $t + i$.

The aggregate shock Z_t is realized at the start of t. Conditional on this shock, the net worth of "surviving" bankers j is the gross return on nonfinancial loans net the cost of deposits and borrowing from the other banks, as follows:

$$n_t^j = (Q_t + Z_t)k_{t-1}^j - R_t d_{t-1}^j - R_{bt} b_{t-1}^j, \tag{10}$$

where d_{t-1}^j is deposit and b_{t-1}^j is interbank borrowing at $t-1$. Note that b_{t-1}^j is positive if bank j borrows and negative if j lends in the interbank market.

For new bankers at t, net worth simply equals the initial endowment:

$$n_t^j = w^j. \tag{11}$$

Meanwhile, exiting bankers no longer operate banks and simply use their net worth to consume:

$$c_t^j = n_t^j. \tag{12}$$

During each period t, a continuing bank j (either new or surviving) finances nonfinancial loans $(Q_t + f_t^j)k_t^j$ with net worth, deposit and interbank debt as follows:

$$(Q_t + f_t^j)k_t^j = n_t^j + d_t^j + b_t^j, \tag{13}$$

where f_t^r is given by (4) and $f_t^w = 0$. We assume that banks can only accumulate net worth via retained earnings. While this assumption is a reasonable approximation of reality, we do not explicitly model the agency frictions that underpin it.[s]

To derive a limit on the bank's ability to raise funds, we introduce the following moral hazard problem: After raising funds and buying assets at the beginning of t, but still during the period, the banker decides whether to operate "honestly" or to divert assets for personal use. Operating honestly means holding assets until the payoffs are realized in period $t+1$ and then meeting obligations to depositors and interbank creditors. To divert means to secretly channel funds away from investments in order to consume personally.

To motivate the use of wholesale funding markets along with retail markets, we assume that the banker's ability to divert funds depends on both the sources and uses of funds. The banker can divert the fraction θ of nonfinancial loans financed by retained earnings or funds raised from households, where $0 < \theta < 1$. On the other hand, he/she can divert only the fraction $\theta\omega$ of nonfinancial loans financed by interbank borrowing, where $0 < \omega < 1$. Here, we are capturing in a simple way that bankers lending in the wholesale market are more effective at monitoring the banks to which they lend than are households that supply deposits in the retail market. Accordingly, the total amount of funds that can be diverted by a banker who is a net borrower in the interbank market is given by

[s] See Bigio (2015) for a model that explains why banks might find it hard to raise external equity during crises in the presence of adverse selection problems.

$$\theta[(Q+f^j)k_t^j - b_t^j + \omega b_t^j]$$

where $(Q+f^j)k_t^j - b_t^j$ equals the value of funds invested in nonfinancial loans that is financed by deposits and net worth and where $b_t^j > 0$ equals the value of nonfinancial loans financed by interbank borrowing.

For bankers that lend to other banks, we suppose that it is more difficult to divert interbank loans than nonfinancial loans. Specifically, we suppose that a banker can divert only a fraction $\theta\gamma$ of its loans to other banks, where $0 < \gamma < 1$. Here, we appeal to the idea that interbank loans are much less idiosyncratic in nature than nonfinancial loans and thus easier for outside depositors to monitor. Accordingly, the total amount a bank that lends on the interbank market can divert is given by

$$\theta[(Q_t + f_t^j)k_t^j + \gamma(-b_t^j)]$$

with $b_t^j < 0$. As we will make clear shortly, key to operation of the interbank market are the parameters that govern the moral hazard problem in this market, ω and γ.

We assume that the process of diverting assets takes time: The banker cannot quickly liquidate a large amount of assets without the transaction being noticed. For this reason, the banker must decide whether to divert at t, prior to the realization of uncertainty at $t+1$. The cost to the banker of the diversion is that the creditors can force the intermediary into bankruptcy at the beginning of the next period.

The banker's decision at t boils down to comparing the franchise value of the bank V_t^j, which measures the present discounted value of future payouts from operating honestly, with the gain from diverting funds. In this regard, rational lenders will not supply funds to the banker if he has an incentive to cheat. Accordingly, any financial arrangement between the bank and its lenders must satisfy the following set of incentive constraints, which depend on whether the bank is a net borrower or lender in the interbank market:

$$\begin{aligned} V_t^j &\geq \theta[(Q+f^j)k_t^j - b_t^j + \omega b_t^j], \text{ if } b_t^j > 0 \\ V_t^j &\geq \theta[(Q_t + f_t^j)k_t^j + \gamma(-b_t^j)], \text{ if } b_t^j < 0. \end{aligned} \tag{14}$$

As will become clear shortly, each incentive constraint embeds the constraint that the net worth n_t^j must be positive for the bank to operate: This is because the franchise value V_t^j will turn out to be proportional to n_t^j.

Overall, there are two basic factors that govern the existence and relative size of the interbank market. The first is the cost advantage that wholesale banks have in managing nonfinancial loans, as described by Assumption 1. The second is the size of the parameters ω and γ which govern the comparative advantage that retail banks have over households in lending to wholesale banks. Observe that as ω and γ decline, it becomes more attractive to channel funds through wholesale bank funding markets relative to retail markets. As ω declines below unity, a bank borrowing in the wholesale market can relax its incentive constraint by substituting interbank borrowing for deposits. Similarly, as γ declines below unity, a bank lending in the wholesale market can relax its incentive constraint by shifting its composition of assets from nonfinancial loans to interbank loans.

In what follows, we restrict attention to the case in which

$$\omega + \gamma > 1. \tag{Assumption 2}$$

In this instance, the parameters ω and γ can be sufficiently small to permit an empirically reasonable relative amount of interbank lending. However, the sum of these parameters cannot be so small as to induce a situation of pure specialization by retail banks, where these banks do not make nonfinancial loans directly but instead lend all their funds to wholesale banks.[t,u] Since in practice retail banks hold some of the same types of assets held by wholesale banks, we think it reasonable to restrict attention to this case.

We now turn to the optimization problems for both wholesale and retail bankers. Given that bankers simply consume their net worth when they exit, we can restate the bank's franchise value recursively as the expected discounted value of the sum of net worth conditional on exiting and the value conditional on continuing as:

$$\begin{aligned} V_t^j &= \beta E_t[(1-\sigma^j)n_{t+1}^j + \sigma^j V_{t+1}^j]. \\ &= E_t[\Omega_{t+1}^j n_{t+1}^j] \end{aligned} \tag{15}$$

where

$$\Omega_{t+1}^j = \beta\left(1 - \sigma^j + \sigma^j \frac{V_{t+1}^j}{n_{t+1}^j}\right). \tag{16}$$

The stochastic discount factor Ω_{t+1}^j, which the bankers use to value n_{t+1}^j, is a probability weighted average of the discounted marginal values of net worth to exiting and to continuing bankers at t+1. For an exiting banker at $t+1$ (which occurs with probability $1-\sigma^j$), the marginal value of an additional unit of net worth is simply unity, since he or she just consumes it. For a continuing banker (which occurs with probability σ^j), the marginal value is the franchise value per unit of net worth V_{t+1}^j/n_{t+1}^j (ie, Tobin's Q ratio). As we show shortly, V_{t+1}^j/n_{t+1}^j depends only on aggregate variables and is independent of bank-specific factors.

We can express the banker's evolution of net worth as:

$$n_{t+1}^j = R_{kt+1}^j\left(Q_t + f_t^j\right)k_t^j - R_{t+1}d_t^j - R_{bt+1}b_t^j \tag{17}$$

[t] See Appendix A for the formal argument that shows that under Assumption 2 pure specialization of retail bankers cannot be an equilibrium.

[u] Holmstrom and Tirole (1997) make similar assumptions on the levels and sum of the agency distortions for banks and nonfinancial firms in order to explain why bank finance arises.

where R^j_{kt+1} is the rate of return on nonfinancial loans, given by

$$R^j_{kt+1} = \frac{Q_{t+1} + Z_{t+1}}{Q_t + f^j_t} \tag{18}$$

The banker's optimization problem then is to choose $\left(k^j_t, d^j_t, b^j_t\right)$ each period to maximize the franchise value (15) subject to the incentive constraint (14) and the balance sheet constraints (13) and (17).

We defer the details of the formal bank maximization problems to Appendix A. Here, we explain the decisions of wholesale and retail banks informally. Because wholesale banks have a cost advantage over retail banks in making nonfinancial loans, the rate of return on nonfinancial loans is higher for the former than for the latter (see Eq. (18)). In turn, retail banks have an advantage over households in lending to wholesale banks due to their relative advantage in recovering assets in default. Therefore, if the interbank market is active in equilibrium, wholesale banks borrow from retail banks in the interbank market to make nonfinancial loans. Indeed the only reason retail banks directly make nonfinancial loans is because wholesale banks may be constrained in the amount of this type of loan they can make.[v]

In the text, we restrict attention to the case where the interbank market is active, with wholesale banks borrowing from retail banks, and where both types of banks are constrained in raising funds externally.

3.3.1 Wholesale banks

In general, wholesale banks may raise funds either from other banks or from households. Since the kinds of financial institutions we have in mind relied exclusively on wholesale markets for funding, we focus on this kind of equilibrium. In particular, we restrict attention to model parameterization which generate an equilibrium where the conditions for the following Lemma 1 are satisfied:

Lemma 1 $d^w_t = 0, b^w_t > 0$ *and the incentive constraint is binding if and only if*

$$0 < \omega E_t\left[\Omega^w_{t+1}\left(R^w_{kt+1} - R_{t+1}\right)\right] < E_t\left[\Omega^w_{t+1}\left(R^w_{kt+1} - R_{bt+1}\right)\right] < \theta\omega$$

We first explain why $d^w_t = 0$ in this instance. The wholesale bank faces the following trade-off in using retail deposits: If the deposit interest rate is lower than the interbank

[v] We do not mean to suggest that the only reason retail banks make nonfinancial loans in practice is because wholesale banks are constrained. Rather we focus on this case for simplicity of the basic model. Later we extend the model to allow for a second type of lending, which we refer to as commercial and industrial lending, where retail banks have a comparative advantage. In this instance, spillovers emerge where problems in wholesale banking can affect the degree of intermediation of commercial and industrial loans.

interest rate so that $E_t[\Omega^w_{t+1}(R^w_{kt+1} - R_{t+1})] > E_t[\Omega^w_{t+1}(R^w_{kt+1} - R_{bt+1})]$, then the bank gains from issuing deposits to reduce interbank loans. On the other hand, because households are less efficient in monitoring wholesale bank behavior, they will apply a tighter limit on the amount they are willing to lend than will retail banks. If ω is sufficiently low so that $\omega E_t[\Omega^w_{t+1}(R^w_{kt+1} - R_{t+1})] < E_t[\Omega^w_{t+1}(R^w_{kt+1} - R_{bt+1})]$, the cost exceeds the benefit. In this instance, the wholesale bank does not use retail deposits, relying entirely on interbank borrowing for external finance. Everything else equal, by not issuing retail deposits, the wholesale bank is able to raise its overall leverage in order to make more nonfinancial loans relative to its equity base. This incentive consideration accounts for why the wholesale bank may prefer interbank borrowing to issuing deposits, even if the interbank rate lies above the deposit rate.[w]

Next we explain why the incentive constraint is binding. If $E_t[\Omega^w_{t+1}(R^w_{kt+1} - R_{bt+1})] < \theta\omega$, then at the margin the wholesale bank gains by borrowing on the interbank market and then diverting funds to its own account. Accordingly, as the incentive constraint (14) requires, rational creditor banks will restrict lending to the point where the gain from diverting equals the bank franchise value, which is what the wholesale bank would lose if it cheated.

Given Lemma 1 we can simplify the evolution of bank net worth to

$$n^w_{t+1} = [(R^w_{kt+1} - R_{bt+1})\phi^w_t + R_{bt+1}]n^w_t \tag{19}$$

where ϕ^w_t is given by

$$\phi^w_t \equiv \frac{Q_t k^w_t}{n^w_t}. \tag{20}$$

We refer to this ratio of assets to net worth as the leverage multiple.

In turn, we can simplify the wholesale banks optimization problem to choosing the leverage multiple to solve:

$$V^w_t = \max_{\phi^w_t} E_t\{\Omega^w_{t+1}[(R^w_{kt+1} - R_{bt+1})\phi^w_t + R_{bt+1}]n^w_t\} \tag{21}$$

subject to the incentive constraint

$$\theta[\omega\phi^w_t + (1-\omega)]n^w_t \le V^w_t \tag{22}$$

[w] Under our baseline parametrization, wholesale banks borrow exclusively from retail banks. We view this as the case that best corresponds to the wholesale banking system on the eve of the Great Recession. Circumstances do exist where wholesale banks will borrow from households as well as retail banks. One might interpret his situation as corresponding to the consolidation of wholesale and retail bank in the wake of the crisis, or perhaps the period before the rapid growth of wholesale banking when retail banks were performing many of the same activities as we often observe in continental Europe and Japan.

Given the incentive constraint is binding under Lemma 1, we can combine the objective with the binding incentive constraint to obtain the following solution for ϕ_t^w:

$$\phi_t^w = \frac{E_t(\Omega_{t+1}^w R_{bt+1}) - \theta(1-\omega)}{\theta\omega - E_t[\Omega_{t+1}^w (R_{kt+1}^w - R_{bt+1})]} \tag{23}$$

Note that ϕ_t^w is increasing in $E_t(\Omega_{t+1}^w R_{kt+1}^w)$ and decreasing in $E_t(\Omega_{t+1}^w R_{bt+1})$.[x] Intuitively, the franchise value V_t^w increases when returns on assets are higher and decreases when the cost of funding asset purchases rises, as Eq. (21) indicates. Increases in V_t^w, in turn, relax the incentive constraint, making lenders will to supply more credit.

Also, ϕ_t^w is a decreasing function of both θ, the diversion rate on nonfinancial loans funded by net worth, and ω, the parameter that controls the relative ease of diverting nonfinancial loans funded by interbank borrowing relative to those funded by the other means: Increases in either parameter tighten the incentive constraint, inducing lenders to cut back on the amount of credit they supply. Later we will use the inverse relationship between ϕ_t^w and ω to help account for the growth in both leverage and size of the wholesale banking sector.

Finally, from Eq. (21) we obtain an expression from the franchise value per unit of net worth

$$\frac{V_t^w}{n_t^w} = E_t\{\Omega_{t+1}^w[(R_{kt+1}^w - R_{bt+1})\phi_t^w + R_{bt+1}]\} \tag{24}$$

where ϕ_t^w is given by Eq. (23) and Ω_{t+1}^w is given by Eq. (16). It is straightforward to show that $\frac{V_t^w}{n_t^w}$ exceeds unity: ie, the shadow value of a unit of net worth is greater than one, since additional net worth permits the bank to borrow more and invest in assets earning an excess return. In addition, as we conjectured earlier, $\frac{V_t^w}{n_t^w}$ depend only on aggregate variables and not on bank-specific ones.

3.3.2 Retail banks

As with wholesale banks, we choose a parametrization where the incentive constraint binds. In addition, as discussed earlier, we restrict attention to the case where retail banks are holding both nonfinancial and interbank loans. In particular, we consider a parametrization where in equilibrium Lemma 2 is satisfied

[x] This is because $E_t(\Omega_{t+1}^w R_{kt+1}^w) > 1 > \theta$ in equilibrium as shown in Appendix.

Lemma 2 $b_t^r < 0$, $k_t^r > 0$ *and the incentive constraint is binding if and only if*

$$0 < E_t[\Omega_{t+1}^r(R_{kt+1}^r - R_{t+1})] = \frac{1}{\gamma}E_t[\Omega_{t+1}^r(R_{bt+1} - R_{t+1})] < \theta$$

For the retail bank to be indifferent between holding nonfinancial loans vs interbank loans, the rate on interbank loans R_{bt+1} must lie below the rate earned on nonfinancial loans R_{kt+1}^r in a way that satisfies the conditions for the lemma. Intuitively, the advantage for the retail bank to making an interbank loan is that households are willing to lend more to the bank per unit of net worth than for a nonfinancial loan. Thus to make the retail bank indifferent, R_{bt+1} must be less than R_{kt+1}^r.

Let ϕ_t^r be a retail bank's effective leverage multiple, namely the ratio of assets to net worth, where assets are weighted by the relative ease of diversion:

$$\phi_t^r \equiv \frac{(Q_t + f_t^r)k_t^r + \gamma(-b_t^r)}{n_t^r}. \tag{25}$$

The weight γ on $(-b_t^r)$ is the ratio of how much a retail banker can divert from interbank loans relative to nonfinancial loans.

Given the restrictions implied by Lemma 2, we can use the same procedure as in the case of wholesale bankers to express the retail banker's optimization problem as choosing ϕ_t^r to solve:

$$V_t^r = \max_{\phi_t^r} E_t\{\Omega_{t+1}^r[(R_{kt+1}^r - R_{t+1})\phi_t^r + R_{t+1}]n_t^r\} \tag{26}$$

subject to

$$\theta\phi_t^r n_t^r \le V_t^r$$

Given Lemma 2, we can impose that incentive constraint binds, which implies

$$\phi_t^r = \frac{E_t(\Omega_{t+1}^r R_{t+1})}{\theta - E_t[\Omega_{t+1}^r(R_{kt+1}^r - R_{t+1})]}. \tag{27}$$

As with the leverage multiple for wholesale bankers, ϕ_t^r is increasing in expected asset returns on the bank's portfolio and decreasing in the diversion parameter.

Finally, from Eq. (26) we obtain an expression for the franchise value per unit of net worth

$$\frac{V_t^r}{n_t^r} = E_t\{\Omega_{t+1}^r[(R_{kt+1}^r - R_{t+1})\phi_t^r + R_{t+1}]\} \tag{28}$$

As with wholesale banks, the shadow value of a unit of net worth exceeds unity and depends only on aggregate variables.

3.4 Aggregation and Equilibrium without Bank Runs

Given that the ratio of assets and liabilities to net worth is independent of individual bank-specific factors and given a parametrization where the conditions in Lemma 1 and 2 are satisfied, we can aggregate across banks to obtain relations between total assets and net worth for both the wholesale and retail banking sectors. Let $Q_tK_t^w$ and $Q_tK_t^r$ be total nonfinancial loans held by wholesale and retail banks, D_t be retail bank deposits, B_t be total interbank debt, and N_t^w and N_t^r total net worth in each respective banking sector. Then we have:

$$Q_tK_t^w = \phi_t^w N_t^w, \tag{29}$$

$$(Q_t + f_t^r)K_t^r + \gamma B_t = \phi_t^r N_t^r, \tag{30}$$

with

$$Q_tK_t^w = N_t^w + B_t, \tag{31}$$

$$(Q_t + f_t^r)K_t^r + B_t = D_t^r + N_t^r, \tag{32}$$

and

$$E_t[\boldsymbol{\Omega}_{t+1}^r(R_{kt+1}^r - R_{t+1})] = \frac{1}{\gamma}E_t[\boldsymbol{\Omega}_{t+1}^r(R_{bt+1} - R_{t+1})]. \tag{33}$$

Eq. (33) ensures that the retail bank is indifferent at the margin between holding nonfinancial loans vs interbank loans (see Lemma 2).

Summing across both surviving and entering bankers yields the following expression for the evolution of N_t :

$$N_t^w = \sigma^w[(R_{kt}^w - R_{bt})\phi_{t-1}^w + R_{bt}]N_{t-1}^w + W^w, \tag{34}$$

$$\begin{aligned} N_t^r = \sigma^r[(R_{kt}^r - R_t)\phi_{t-1}^r + R_t]N_{t-1}^r + W^r \\ +\sigma^r\left[R_{bt} - R_t - \gamma(R_{kt}^r - R_t)\right]B_{t-1}, \end{aligned} \tag{35}$$

where $W^j = (1 - \sigma^j)w^j$ is the total endowment of entering bankers. The first term is the accumulated net worth of bankers that operated at $t - 1$ and survived to t, which is equal to the product of the survival rate σ^j and the net earnings on bank assets.

Total consumption of bankers equals the sum of the net worth of exiting bankers in each sector:

$$C_t^b = (1 - \sigma^w)\frac{N_t^w - W^w}{\sigma^w} + (1 - \sigma^r)\frac{N_t^r - W^r}{\sigma^r} \tag{36}$$

Total gross output $\bar{Y}_t$ is the sum of output from capital, household endowment Z_tW^h and bank endowment W^r and W^i :

$$\bar{Y}_t = Z_t + Z_tW^h + W^r + W^i. \tag{37}$$

Net output Y_t, which we will refer to simply as output, equals gross output minus management costs

$$Y_t = \bar{Y}_t - [F^h(K_t^h) + F^r(K_t^r)] \tag{38}$$

Eq. (38) captures in a simple way how intermediation of assets by wholesale banks improves aggregate efficiency. Finally, output is consumed by households and bankers:

$$Y_t = C_t^h + C_t^b. \tag{39}$$

The recursive competitive equilibrium without bank runs consists of aggregate quantities,

$$\left(K_t^w, K_t^r, K_t^h, B_t, D_t^r, N_t^w, N_t^r, C_t^b, C_t^h, \bar{Y}_t, Y_t\right),$$

prices

$$\left(Q_t, R_{t+1}, R_{bt+1}, f_t^r\right)$$

and bankers' variables

$$\left(\Omega_t^j, R_{kt}^j, \frac{V_t^j}{n_t^j}, \phi_t^j\right)_{j=w,r}$$

as a function of the state variables $\left(K_{t-1}^w, K_{t-1}^r, R_{bt}B_{t-1}, R_t D_{t-1}^w, R_t D_{t-1}^r, Z_t\right)$, which satisfy Eqs. (1, 4, 7, 8, 16, 18, 23, 24, and 27–39).[y]

3.5 Unanticipated Bank Runs

In this section we consider unanticipated bank runs. We defer an analysis of anticipated bank runs to Section 5. In general, three types of runs are conceivable: (i) a run on wholesale banks leaving retail banks intact, (ii) a run on just retail banks, and (iii) a run on both the wholesale and retail bank sectors. We restrict attention to (i) because it corresponds most closely to what happened in practice.

3.5.1 Conditions for a Wholesale Bank Run Equilibrium

The runs we consider are runs on the entire wholesale banking system, not on individual wholesale banks. Indeed, so long as an asset firesale by an individual wholesale bank is not large enough to affect asset prices, it is only runs on the system that will be disruptive. Given the homogeneity of wholesale banks in our model, the conditions for a run on the wholesale banking system will apply to each individual wholesale bank.

What we have in mind for a run is a spontaneous failure of the bank's creditors to roll over their short-term loans[z]. In particular, at the beginning of period t, before the

[y] In total we have a system of 23 equations. Notice that (16 and 18) have two equations. By Walras' law, the household budget constraint (6) is satisfied as long as deposit market clears as $D_t = D_t^r$.

[z] The approach follows Cole and Kehoe's (2000) model of self fulfilling debt crises.

realization of returns on bank assets, retail banks lending to a wholesale bank decide whether to roll over their loans with the bank. If they choose to "run," the wholesale bank liquidates its capital and turns the proceeds over to its retail bank creditors who then either acquire the capital or sell it to households. Importantly, both the retail banks and households cannot seamlessly acquire the capital being liquidated in the firesale by wholesale banks. The retail banks face a capital constraint which limits asset acquisition and are also less efficient at managing the capital than are wholesale banks. Households can only hold the capital directly and are even less efficient than retail banks in doing so.

Let Q_t^* be the price of capital in the event of a forced liquidation of the wholesale banking system. Then a run on the entire wholesale bank sector is possible if the liquidation value of wholesale banks assets, $(Z_t + Q_t^*)K_{t-1}^w$, is smaller than their outstanding liability to interbank creditors, $R_{bt}B_{t-1}$, so that liquidation would wipe out wholesale banks networth. In this instance, the recovery rate in the event of a wholesale bank run, x_t^w, is the ratio of $(Z_t + Q_t^*)K_{t-1}^w$ to $R_{bt}B_{t-1}$ and the condition for a bank run equilibrium to exist is that the recovery rate is less than unity, ie,

$$x_t^w = \frac{(Q_t^* + Z_t)K_{t-1}^w}{R_{bt}B_{t-1}} < 1. \tag{40}$$

Let R_{kt}^{w*} be the return on bank assets conditional on a run at t :

$$R_{kt}^{w*} \equiv \frac{Z_t + Q_t^*}{Q_{t-1}},$$

Then from (40), we can obtain a simple condition for a wholesale bank run equilibrium in terms of just two endogenous variables: (i) the ratio of R_{kt}^{w*} to the interbank borrowing rate R_{bt} and (ii) the leverage multiple ϕ_{t-1}^w :

$$x_t^w = \frac{R_{kt}^{w*}}{R_{bt}} \cdot \frac{\phi_{t-1}^w}{\phi_{t-1}^w - 1} < 1 \tag{41}$$

A bank run equilibrium exists if the realized rate of return on bank assets conditional on liquidation of assets R_{kt}^{w*} is sufficiently low relative to the gross interest rate on interbank loans, R_{bt}, and the leverage multiple is sufficiently high to satisfy condition (41). Note that the expression $\frac{\phi_{t-1}^w}{\phi_{t-1}^w - 1}$ is the ratio of bank assets $Q_{t-1}K_{t-1}^w$ to interbank borrowing B_{t-1}, which is decreasing in the leverage multiple. Also note that the condition for a run does not depend on individual bank-specific factors since R_{kt}^{w*}/R_{bt} and ϕ_{t-1}^w are the same for all in equilibrium.

Since R_{kt}^{w*}, R_{bt} and ϕ_{t-1}^w are all endogenous variables, the possibility of a bank run may vary with macroeconomic conditions. The equilibrium absent bank runs (that we described earlier) determines the behavior of R_{bt} and ϕ_{t-1}^w. The value of R_{kt}^{w*}, instead, depends on the liquidation price Q_t^*, whose determination is described in the next subsection.

3.5.2 The Liquidation Price

To determine Q_t^* we proceed as follows. A run by interbank creditors at t induces all wholesale banks that carried assets from $t - 1$ to fully liquidate their asset positions and go out of business.[aa] Accordingly they sell all their assets to retail banks and households, who hold them at t. The wholesale banking system then rebuilds itself overtime as new banks enter. For the asset firesale during the panic run to be quantitatively significant, we need there to be at least a modest delay in the ability of new banks to begin operating. Accordingly, we suppose that new wholesale banks cannot begin operating until the period after the panic run.[ab]

Accordingly, when wholesale banks liquidate, they sell all their assets to retail banks and households in the wake of the run at date t, implying

$$\bar{K} = K_t^r + K_t^h. \tag{42}$$

The wholesale banking system then rebuilds its equity and assets as new banks enter at $t + 1$ onwards. Given our timing assumptions and Eq. (34), bank net worth evolves in the periods after the run according to

$$\begin{aligned} N_{t+1}^w &= (1+\sigma^w)W^w, \\ N_{t+i}^w &= \sigma^w\left[(Z_{t+i} + Q_{t+i})K_{t+i-1}^w - R_{bt+i}B_{t+i-1}\right] + W^w, \text{ for all } i \geq 2. \end{aligned}$$

Rearranging the Euler equation for the household's capital holding (8) yields the following expression for the liquidation price in terms of discounted dividends Z_{t+i} net the marginal management cost $\alpha^h K_{t+i}^h$.

$$Q_t^* = E_t\left[\sum_{i=1}^{\infty} \Lambda_{t,t+i}(Z_{t+i} - \alpha^h K_{t+i}^h)\right] - \alpha^h K_t^h. \tag{43}$$

Everything else equal, the longer it takes for the banking sector to recapitalize (measured by the time it takes K_{t+i}^h to fall back to steady state), the lower will be the liquidation price. Note also that Q_t^* will vary with cyclical conditions. In particular, a negative shock to Z_t will reduce Q_t^*, possibly moving the economy into a regime where bank runs are possible.

[aa] Our notion of the liquidation price is related to Brunnermeier and Pedersen's (2009) concept of market liquidity. See Uhlig (2010) for an alternative bank run model with endogenous liquidation prices.

[ab] Suppose for example that during the run it is not possible for retail banks to identify new wholesale banks that are financially independent of the wholesale banks being run on. New wholesale banks accordingly wait for the dust to settle and then begin raising fund in the interbank market in the subsequent period. The results are robust to alternative timing assumptions about the entry of new banks.

4. NUMERICAL EXPERIMENTS

In this section, we examine how the long run properties of the model can account for the growth of the wholesale banking sector and then turn to studying the cyclical responses to macroeconomic shocks that may or may not induce runs. Overall these numerical examples provide a description of the tradeoff between growth and stability associated with an expansion of the shadow banking sector and illustrate the real effects of bank runs in our model.

4.1 Calibration

Here, we describe our baseline calibration. This is meant to capture the state of the economy at the onset of the financial crisis in 2007.

There are 13 parameters in the model:

$$\{\theta,\omega,\gamma,\beta,\alpha^h,\alpha^r,\sigma^r,\sigma^w,W^h,W^r,W^w,Z,\rho_z\}.$$

their values are reported in Table 2, while Table 3 shows the steady state values of the equilibrium allocation.

We take the time interval in the model to be a quarter. We use conventional values for households' discount factor, $\beta = 0.99$, and the serial correlation of dividends $\rho_z = 0.9$. We normalize the steady state level of productivity Z in order for the price of loans to be unity and set W^h so that households endowment income is twice as big as their capital income.

We calibrate managerial costs of intermediating capital for households and retail bankers, α^h and α^r, in order for the spread between the deposit rate and retail bankers' returns on loans as well as the difference between wholesale bankers and retail bankers returns on loans to be 1.2% in annual in steady state.[ac]

The fraction of divertible interbank loans $\theta\gamma$ is set in order to obtain an annualized steady state spread between deposit and interbank rates of 0.8%. The fraction of divertible assets purchased by raising deposits, θ, and interbank loans, $\omega\theta$, are set in order to get leverage ratios for retail bankers and wholesale bankers of 10 and 20, respectively.

Our retail banking sector comprises of commercial banks, open end Mutual Funds and Money Market Mutual Funds (MMMF). In the case of Mutual Funds and MMMF the computation of leverage is complicated by the peculiar legal and economic details of the relationship between these institutions, their outside investors and sponsors.[ad] Hence, our choice of 10 quite closely reflects the actual leverage ratios of commercial banks,

[ac] Philippon (2015) calculates interest rate spreads charged by financial institutions to be around 200 basis points.

[ad] On the relationship between MMFs and their sponsors see, for instance, Parlatore (2015) and McCabe (2010).

Table 2 Baseline parameters

	Parameters	
Households		
β	Discount rate	0.99
α^h	Intermediation cost	0.03
W^h	Endowment	0.006
Retail banks		
σ^r	Survival probability	0.96
σ^r	Intermediation cost	0.0074
W^r	Endowment	0.0008
θ	Divertable proportion of assets	0.25
γ	Shrinkage of divertable proportion of interbank loans	0.67
Wholesale banks		
σ^w	Survival probability	0.88
σ^w	Intermediation cost	0
W^w	Endowment	0.0008
ω	Shrinkage of divertable proportion of assets	0.46
Production		
z	Steady state productivity	0.016
ρ_z	Serial correlation of productivity shocks	0.9

Table 3 Baseline steady state

	Steady state	
Q	Price of capital	1
K^r	Retail intermediation	0.4
K^w	Wholesale intermediation	0.4
R^b	Annual interbank rate	1.048
R_r^k	Annual retail return on capital	1.052
R	Annual deposit rate	1.04
R_w^k	Annual wholesale return on capital	1.064
ϕ^w	Wholesale leverage	20
ϕ^r	Retail leverage	10
Y	Output	0.0229
C^h	Consumption	0.0168
N^r	Retail banks networth	0.0781
N^w	Wholesale banks networth	0.02

which is the only sector for which a direct empirical counterpart of leverage can be easily computed.

To set our target for wholesale leverage we decided to focus on private institutions within the wholesale banking sector that relied mostly on short-term debt. A reasonable range for the leverage multiple for such institutions goes from around 10 for some ABCP issuers[ae] to values of around 40 for brokers dealers in 2007. Our choice of 20 is a conservative target within this range.

The survival rates of wholesale and retail bankers, σ^w and σ^r, are set in order for the distribution of assets across sectors to match the actual distribution in 2007. Finally, we set W^r to make new entrants net worth being equal to 1% of total retail banks net worth and W^w to ensure that wholesale bankers are perfectly specialized.

4.2 Long Run Effects of Financial Innovation

As mentioned in Section 2, the role of wholesale banks in financial intermediation has grown steadily from the 1980s to the onset of the financial crisis. This growth was largely accomplished through a series of financial innovations that enhanced the borrowing capacity of the system by relying on securitization to attract funds from institutional investors. While our model abstracts from the details of the securitization process, we capture its direct effects on wholesale banks' ability of raising funds in interbank markets with a reduction in the severity of the agency friction between retail banks and wholesale banks, which is captured by parameter ω. Hence, in this section we study the long run behavior of financial intermediation in response to a decrease in ω and compare it to the low frequency dynamics in financial intermediation documented in Section 2.

The direct effect of ameliorating the agency problem between wholesale and retail banks is a relaxation of wholesale banks' incentive constraints. The improved ability of retail banks to seize the assets of wholesale bankers in the case of cheating allows wholesale bankers to borrow more aggressively from retail bankers.

Fig. 8 shows how some key variables depend upon ω in the steady state.[af] The general equilibrium effects of a lower ω work through various channels. For an economy with a lower interbank friction ω, the leverage multiple of the wholesale banking sector is higher, with a larger capital K^w and a larger amount interbank borrowing B by wholesale banks. Conversely, capital intermediated by retail banks K^r and households K^h tends to be lower. In the absence of bank runs, the relative shift of assets to the wholesale banking sector implies a more efficient allocation of capital and consequently a higher capital price

[ae] The same caveat as in the case of MMFs applies here because it is very complicated to factor in the various lines of credit that were provided by the sponsors of these programs.

[af] Notice that as ω increases above a certain threshold, two other types of equilibria arise: one in which wholesale bankers are imperfectly specialized and raise funds in both wholesale and retail markets; and one in which the interbank market shuts down completely. See the Appendices for details.

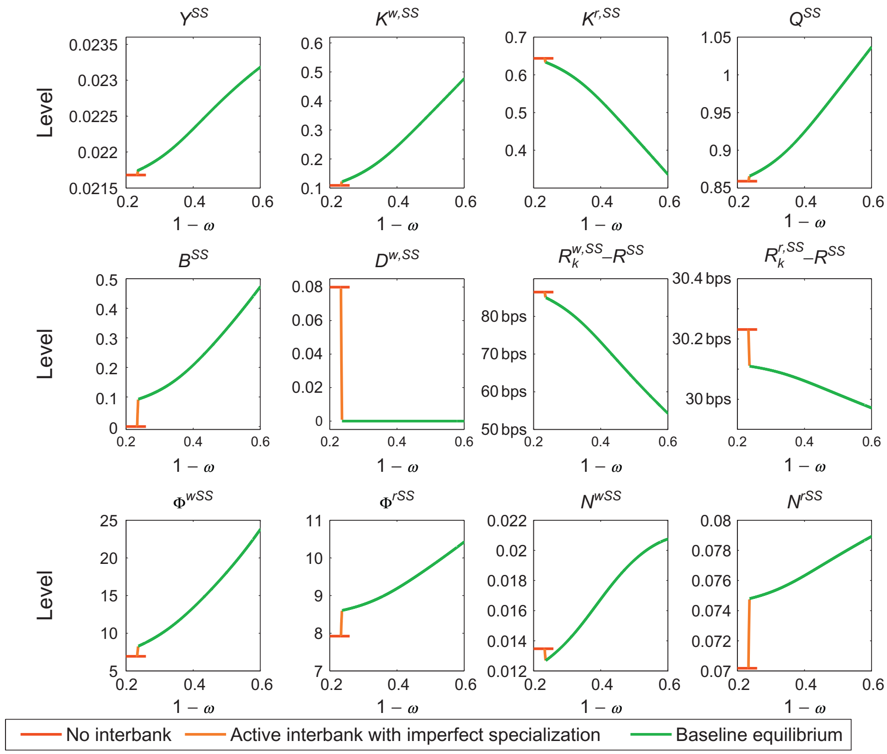


Fig. 8 Comparative statics: a reduction in ω.

Q_t. The flow of assets into wholesale banking, further, reduces the spread between the return on capital for wholesale banks and the interbank rate, as well as the spread between interbank and deposit rates. Despite lower spreads, both wholesale and retail banks enjoy higher franchise values thanks to the positive effect of higher leverage on total returns on equity. A unique aspect of financial innovation due to a lower friction in the interbank market is that the borrowing and lending among banks tends to be larger relative to the flow-of-funds from ultimate lenders (households) to ultimate nonfinancial borrowers. (See Appendix B).

Fig. 9 compares the steady state effect of financial innovations on some key measures of financial intermediation with the observed low frequency trends in their empirical counterparts. In particular, we assume that the value of ω in our baseline calibration results from a sequence of financial innovations that took place gradually from the 1980s to the financial crisis. For simplicity, we divide our sample into 2 periods of equal length and assign a value of ω to each subsample in order to match the observed percentage of intermediation of wholesale bankers over the period. In order to compute leverage of wholesale banks in Fig. 9, we compute leverage of the three sectors within the wholesale banking sector that were mainly responsible for the growth of wholesale

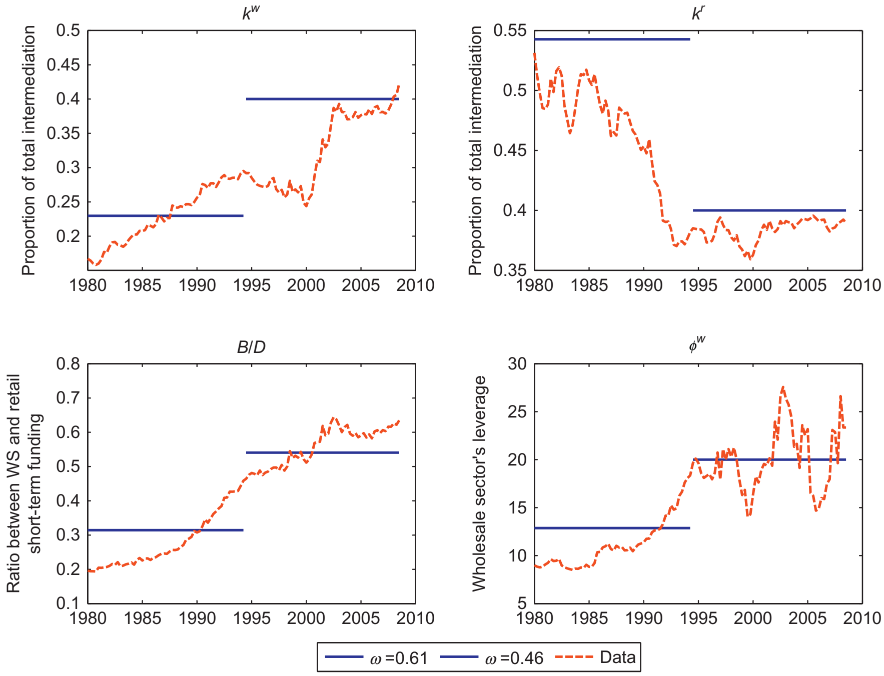


Fig. 9 Low frequency dynamics in financial intermediation.

intermediation. Overall, the steady state comparative statics capture quite well the actual low frequency dynamics in financial intermediation observed over the past few decades.[ag]

4.3 Recessions and Runs

We now turn to the cyclical behavior of our model economy. Fig. 10 shows the response of the economy to an unanticipated negative 6% shock to productivity Z_t, assuming that a run does not happen.[ah] To capture the effects of financial liberalization on the cyclical properties of the economy, we consider both our baseline parameterization and one with a higher ω which we set to be equal to the one associated with the early 1980s in Fig. 9. In both cases the presence of financial constraints activates the familiar financial accelerator mechanism of Bernanke and Gertler (1989) and Kiyotaki and Moore (1997). Leverage amplifies the effects of the drop in Z_t on bankers' net worth, inducing a tightening of

[ag] The model overstatement of the role of retail intermediation relative to household direct holding of assets can be rationalized by the lack of heterogeneity in ultimate borrowers' funding sources since, in the data, households mainly hold equities while intermediaries are responsible for most debt intermediation. Introducing a different type of asset for which intermediaries have a smaller advantage would then help to reconcile the evolution of the distribution of capital across sectors predicted by the model in response to financial innovation with the empirical one.

[ah] We choose the size of the shock to generate a fall in output similar to the one that occurred during the Great Recession.

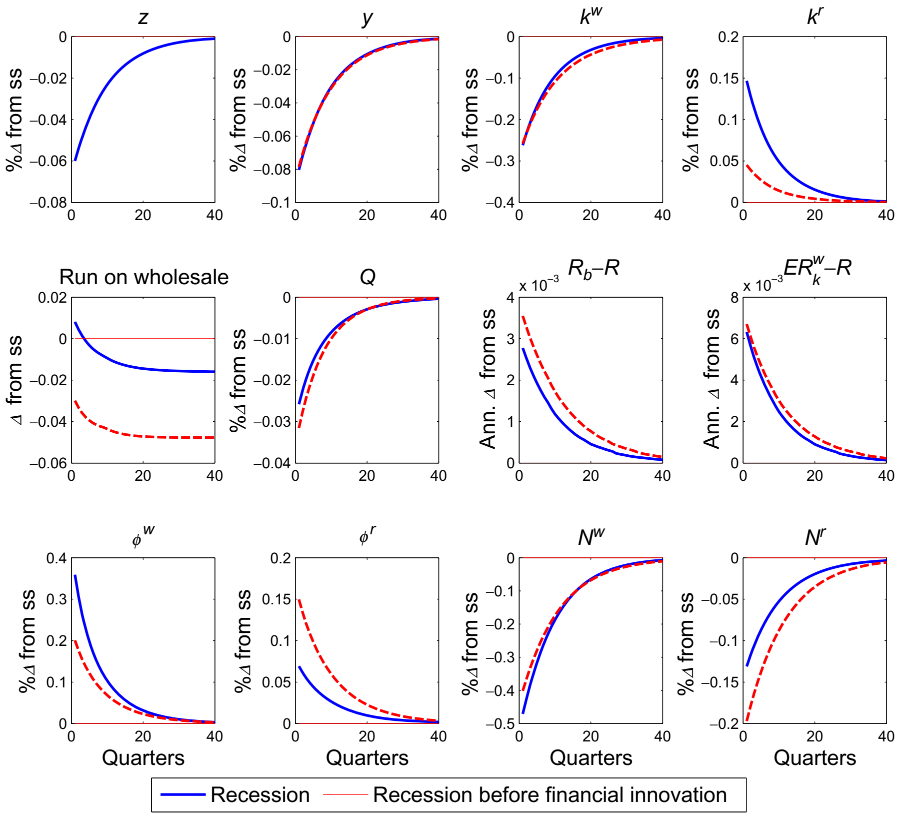


Fig. 10 A recession before and after financial innovation (NO RUN EQUILIBRIUM).

financial constraints, as reflected by an increase in credit spreads. In turn, wholesale banks sell off loans, which reduces asset prices and feeds back into lower net worth. Higher exposure to variations in Z_t and higher leverage make this effect stronger for wholesale banks that are forced into a firesale liquidation of their assets, which in turn leads them to reduce their demand for interbank loans. As a result, retail bankers increase their asset holdings and absorb, together with households, the capital flowing out of the wholesale banking sector.[ai] However, the relative inefficiency of these agents in intermediating assets makes this process costly as shown by the rise in the cost of bank credit and the amplification in the drop in output. Under our baseline calibration, spreads between gross borrowing costs for nonfinancial borrowers and the risk free rate increase by sixty basis points and output drops by 8%, which is two percentage points greater than the drop in Z_t.[aj]

[ai] The increase in households' capital holding is consistent with the shift from intermediated to unintermediated capital observed during the crisis. See, eg, Adrian et al (2012) for evidence.

[aj] Observe also that in a production economy with investment and nominal rigidities, the drop in the asset price would reduce investment and thus aggregate demand, magnifying the overall drop in output.

As we noted earlier, financial innovation makes the economy operate more efficiently in steady state. Fig. 10 shows that, absent bank runs, it also makes the economy more stable as the financial accelerator weakens. In response to the drop in Z_t, the economy with financial innovation features smaller increases in credit spreads and a smaller drop in assets prices. Intuitively, with financial innovation, retail banks provide a stronger buffer to absorb loan sales by wholesale banks, which helps stabilize asset prices. At the same time, the economy with financial innovation is more vulnerable to a bank run.

This is illustrated by the panel titled "Run on Wholesale" in Fig. 10. In this panel we plot a variable that indicates at each time t whether a run is possible at time $t + 1$. To construct this variable we define

$$Run_t^w = 1 - x_t^w$$

where x_t^w is the recovery rate on wholesale debt. Hence, in order for a run to exist the run variable must be positive.

As shown by the Run^w variable, a run on wholesale banks is not possible in the steady state under both parameterization considered. With a 6% drop in Z_t, a run equilibrium remains impossible in the economy absent financial innovation, ie, the one with a high value of ω. However, for the economy with financial innovation (ie, a low ω), the same drop in Z_t is big enough to make a run on wholesale banking possible. Intuitively, in the low ω economy, wholesale bank leverage ratios are higher than would be otherwise, and asset liquidation values are lower, which raises the likelihood that the conditions for a bank run equilibrium will be satisfied.

Fig. 11 describes the effects of bank runs. In particular we assume that two periods after the unanticipated drop in Z_t, retail investors stop rolling over short-term debt issued by wholesale banks, inducing them to liquidate all of their assets and go bankrupt.

As explained in Section 3.5.1, the run on wholesale banks forces them into bankruptcy and results in K^w dropping to 0. Households and retail banks are forced to absorb all of the wholesale banks' assets, inducing asset prices to drop by about 7% in total. The intermediation costs associated with the reallocation of assets to less efficient agents leads to an additional contraction of output of around 7%, resulting in an overall drop of about 15%.

As new wholesale bankers resume operations from the period after the run, high levels of spreads for both retail and wholesale bankers allow them to increase their leverage and recapitalize financial intermediaries thanks to above average retained earnings. The reintermediation process, however, is rather lengthy and output remains depressed for a prolonged period of time.

5. ANTICIPATED RUNS

So far, we have focused on the case in which runs are completely unexpected. In this section we study how the equilibrium changes if agents anticipate that a run will occur

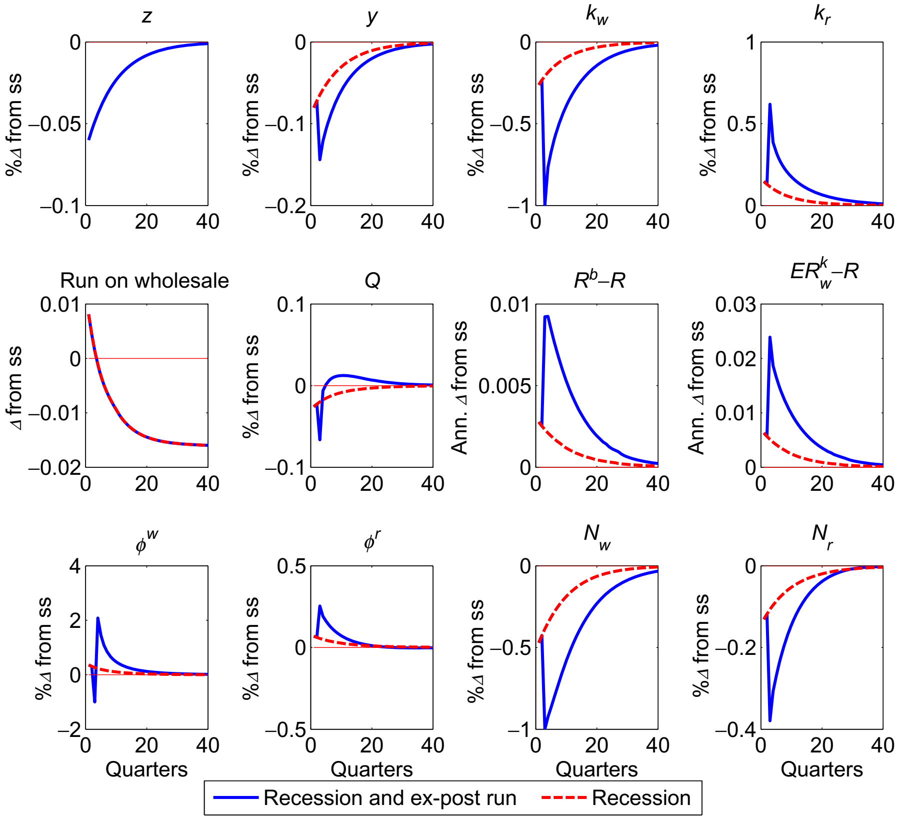


Fig. 11 A recession followed by a run on wholesale bankers.

with positive probability in the future, focusing on the more realistic case of a run on wholesale bankers only. The Appendices contains a detailed description of the equilibrium in this case.[ak] Here, we describe the key forces through which anticipation of a run in the future affects financial intermediation. To keep the analysis as simple as possible, we assume that once a negative shock to Z_t hits, Z_t obeys perfect foresight path back to steady state.

The main difference from the unanticipated case is in the market for interbank loans. In particular, once runs are anticipated, retail bankers internalize how wholesale bankers' leverage affects returns on interbank loans in case of a run and they adjust the required promised rate $\bar{R}_{bt+1}$ accordingly. We denote by p_t the time t probability that retail banks will run on wholesale banks at time t+1.[al] The indifference condition of the retail bank between making interbank loans and nonfinancial loans (33) becomes:

[ak] The analysis of anticipated runs draws heavily on Gertler and Kiyotaki (2015).

[al] The determination of this probability of "observing a sunspot" will be discussed later.

$$\begin{aligned}&E_t[(1-p_t)\Omega^r_{t+1}(\bar{R}_{bt+1}-R_{t+1})+p_t\Omega^{r*}_{t+1}(x^w_{t+1}\bar{R}_{bt+1}-R_{t+1})]\\&=\gamma E_t[(1-p_t)\Omega^r_{t+1}(R^r_{kt+1}-R_{t+1})+p_t\Omega^{r*}_{t+1}(R^{r*}_{kt+1}-R_{t+1})],\end{aligned}\tag{44}$$

where

$$\Omega^{r*}_{t+1}=\beta\left(1-\sigma+\sigma\frac{V^{r*}_{t+1}}{n^{r*}_{t+1}}\right)$$

is the value of the stochastic discount factor if a run occurs at $t+1$.

Using Eq. (41) to substitute for x^w_{t+1} in (44) we obtain a menu of promised rates:[am]

$$\begin{aligned}\bar{R}_{bt+1}(\phi^w_t)=(1-\gamma)R_{t+1}+\gamma\frac{E_t(\Omega^r_{t+1}R^r_{kt+1})}{E_t(\Omega^r_{t+1})}\\+\frac{p_t}{(1-p_t)E_t(\Omega^r_{t+1})}E_t\left\{\Omega^{r*}_{t+1}\left[(1-\gamma)R_{t+1}+\gamma R^{r*}_{kt+1}-\frac{\phi^w}{\phi^w-1}R^{w*}_{kt+1}\right]\right\}\end{aligned}\tag{45}$$

Notice that $\bar{R}_{bt+1}(\phi^w_t)$ is an increasing function ϕ^w_t. This is because as leverage increases, retail bankers suffer larger losses on interbank loans if a run occurs. This induces them to require higher returns in the event of no run, to compensate for the larger losses in the event of a run.

When choosing their portfolios, wholesale bankers will now have to factor in that changes in their leverage affect their cost of credit according to Eq. (45). This preserves homogeneity of the problem but the franchise value of the firm will change to reflect that with probability p_t the bank will be forced to liquidate assets at price Q^*_{t+1} in the subsequent period. This will have the effect of reducing the franchise value of wholesale banks, hence tightening their financial constraints.

In particular the franchise value of a wholesale bank will be given by[an]

$$\frac{V^w_t}{n^w_t}=(1-p_t)E_t\{\Omega^w_{t+1}[\phi^w_t(R^w_{t+1}-\bar{R}_{bt+1}(\phi^w_t))+\bar{R}_{bt+1}(\phi^w_t)]\}.\tag{46}$$

An increase in p_t reduces the franchise value through two channels: First, it decreases the likelihood that the bank will continue to operate next period. Second, it leads to an increase in the interbank loan rate each individual bank faces, $\bar{R}_{bt+1}(\phi^w_t)$, which reduces the franchise value even if the bank continues to operate.

[am] This is the relevant function for values of leverage high enough to induce bankruptcy in case of a run.

[an] Here, we are already assuming that wholesale bankers will choose a leverage high enough to result in bankruptcy when a run occurs. See the Appendices for a detailed description of the wholesale banker's problem when runs are anticipated. There, we derive the conditions that ensure that it is optimal for wholesale bankers to default in the event of a run.

In order to pin down a state dependent probability of a run, we follow Gertler and Kiyotaki (2015). In particular we assume that at each time t the probability of transitioning to a state where a run on wholesale banks occurs is given by a reduced form decreasing function of the expected recovery rate $E_t x^w_{t+1}$ as follows,

$$p_t = \left[1 - E_t\left(x^w_{t+1}\right)\right]^{\delta}. \tag{47}$$

Although we don't endogenize the functional dependence of p_t on the state of the economy, the above formulation allows us to capture the idea that as wholesale balance sheet positions weaken, the likelihood of a run increases. This same qualitative conclusion would follow, for example, if the probability of a run was determined endogenously by introducing imperfect information, as in the global games approach developed by Morris and Shin (1998).[ao]

Fig. 12 demonstrates how anticipation effects work to increase financial amplification of shocks in the model. The solid line is the response of the economy to an unanticipated

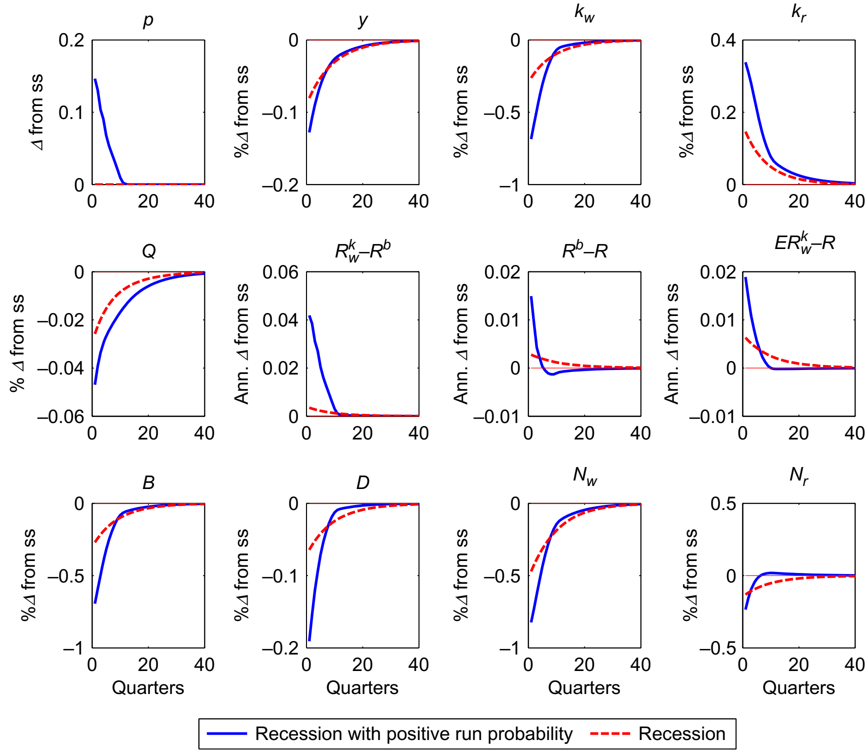


Fig. 12 A recession in the model with anticipated runs.

[ao] See Gertler et al (2016) for an alternative formulation of beliefs in a very similar setup and Goldstein and Pauzner (2005) for an application of the global games approach to bank runs.

6% shock to Z_t when agents anticipate that a run can happen at each time $t + 1$ with probability p_t as determined in Eq. (47).[ap] As we noted earlier, we assume that after the shock Z_t follows a perfect foresight path back to steady state. To isolate the effect of the anticipation of the run, we suppose in this case that the run never actually occurs ex-post. For comparison, the dotted line reports the responses of the baseline economy in which individuals assign probability zero to a bank run.

While it is still the case that in steady state a run cannot occur, the shock to Z_t leads the probability of a run to increase to 15%. As wholesale bankers' balance sheets weaken and the liquidation price decreases, retail bankers expect more losses on interbank loans in case of a run and the probability of coordinating on a run equilibrium increases as a result. The increase in p_t leads to a sharp contraction in the supply of interbank credit and a further tightening of wholesale bankers financial constraints. This, in turn, results in an overall reduction in their net worth of about 80% compared to a 50% in the baseline and to a spike in spreads between nonfinancial loan and interbank loan rates that increase by 400 basis points compared to only 30 in the baseline. As wholesale banks are forced to downsize their operations, total interbank credit falls by about 70%, more than twice the percentage drop in the baseline. These massive withdrawals of funds from wholesale markets is the model counterpart to the "slow runs" on the ABCP market in 2007. These disruptions in wholesale funding markets are then transmitted to the rest of the economy inducing a drop in asset prices of 5% and a total contraction of output of 13%.

Fig. 13 shows the case in which the run actually occurs two periods after the realization of the shock to Z_t. There are two main differences with respect to the analogous experiment performed in the case of unanticipated runs depicted in Fig. 11. First, the initial increase in the probability of a run that precedes the actual run allows the model to capture the "slow runs" followed by "fast runs" in wholesale funding markets that was a central feature of the financial crisis, as discussed in the Introduction. Second, the run induces a further increase in the probability of additional runs in the future, that goes back to about 20% the period after the run occurs. This hampers wholesale bankers ability to increase their leverage and generates higher spreads in the interbank market preventing the relatively smooth increase in asset prices that characterizes the recovery in the baseline model.

Fig. 14 shows how the model with anticipated runs can reproduce some key features of the financial disruptions that occurred in 2007 and 2008. In particular, we compare the model predicted path for interbank spreads, $\bar{R}^b_{t+1} - R_{t+1}$, and excess finance premium, $ER^w_{k,t+1} - R_{t+1}$, with their empirical counterparts over the period going from 2007Q2 to 2009Q4. For the interbank spreads we choose the ABCP spread, since the first "slow runs" in wholesale funding markets in the third quarter of 2007 took place in the ABCP market. The measure of excess borrowing costs is the Excess Bond Premium of Gilchrist

[ap] In the numerical simulations below we pick δ to be $\frac{1}{2}$.

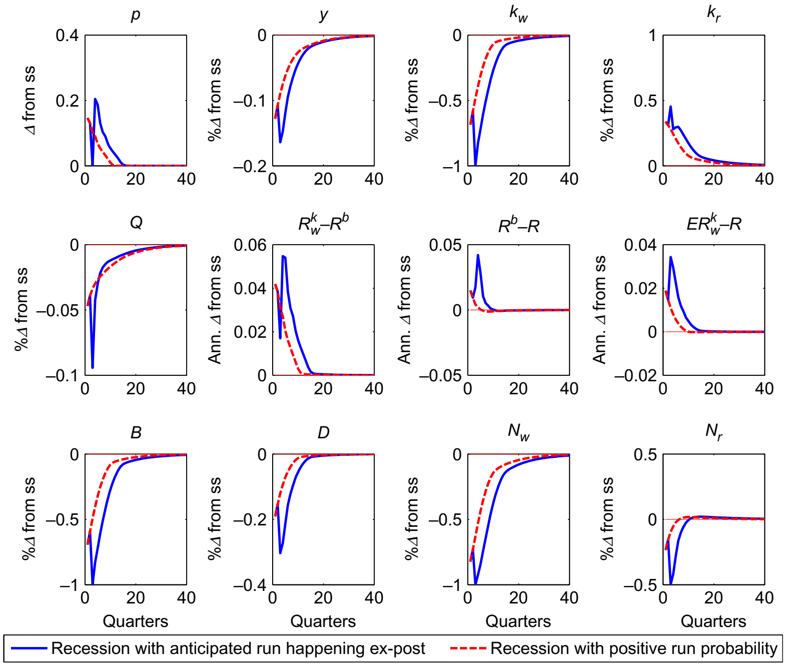


Fig. 13 A recession followed by a run in the model with anticipated runs.

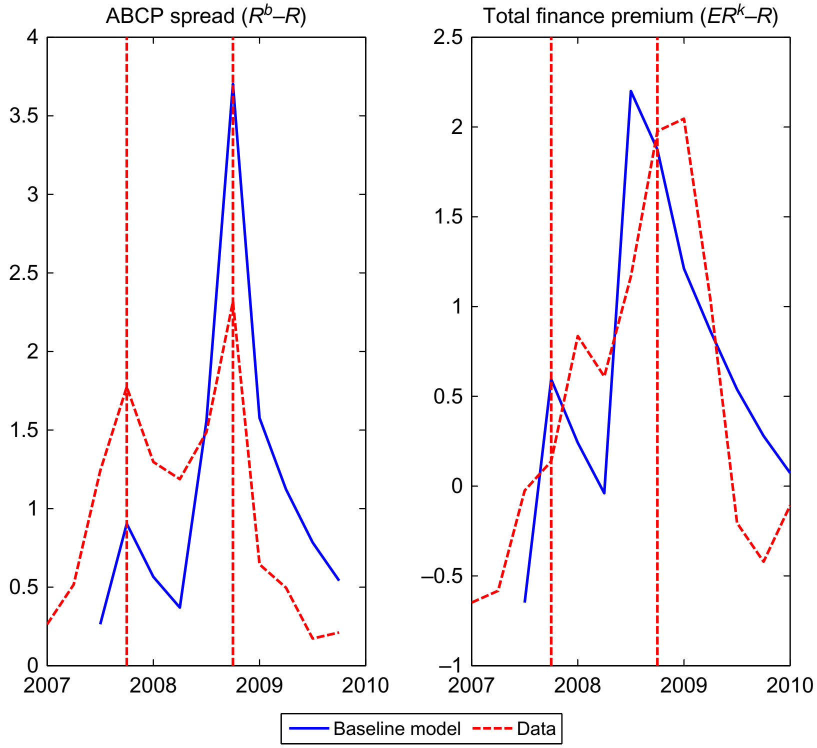


Fig. 14 Total credit spreads and interbank spreads in the model and in the data.

and Zakrajsek (2012). We assume that the economy is in steady state in 2007Q2 and the unanticipated shock hits in 2007Q3 followed by a run on wholesale banks in 2008Q3.[aq] In the data excess borrowing costs lag financial spreads, so the model predicts a stronger initial increase in $ER^{w}_{k,t+1} - R_{t+1}$ and attributes a slightly smaller proportion of the increase to interbank spreads, probably due to the behavior of the risk free rate. On the other hand, the faster decline in spreads in the data after 2009 can be attributed to the effects of government intervention in this period. Overall, the experiment can capture the credit spreads and bank equity dynamics reasonably well.

6. TWO PRODUCTIVE ASSETS AND SPILLOVER EFFECTS

In our baseline model there is only one type of capital. Wholesale banks have an efficiency advantage in holding this capital. Retail banks exist mainly because wholesale banks may be constrained by their net worth; otherwise the latter would hold all the capital. In this section we introduce a second type of capital which retail banks have an efficiency advantage in intermediating. In addition to providing a stronger motivation for the existence of retail banking, the second asset allows us to illustrate spillover effects from a crisis in wholesale banking into retail banking.

In particular, one of the salient features of the recent crisis was the strong contagion effect through which the collapse in subprime mortgage related products within the wholesale banking sector led to a deterioration in financial conditions within the commercial banking sector, ultimately affecting the flow credit through these institutions. Even though on the eve of the crisis, much of the credit provided by the retail sector had no direct reliance on shadow banks, the collapse of the latter ultimately disrupted commercial bank lending, enhancing the downturn.

As is the case with the first type of capital, we suppose the second type is fixed in supply and denote the total as $\bar{L}$. We refer to bank loans made to finance this capital as "C&I" loans (for "commercial and industrial" loans). What we have in mind are the kinds of information-intensive loans that are not easily securitized, which retail banks have historically specialized in intermediating. This contrasts with the kinds of securitized assets, involving mortgages, car loans, credit card debt, trade credit and so on, that were principally held by wholesale banks.

For simplicity, we assume that only retail banks and households fund the second type of capital. Given L_t^r and L_t^h are the amounts funded by retail banks and households, we have:

$$L_t^h + L_t^r = \bar{L} \tag{48}$$

We model retail banks' comparative advantage in making C&I loans by assuming that management costs of intermediating these loans are zero for these types of banks.

[aq] To be closer to the observed dynamics of spreads we resize the innovation to Z_t to five percentage points.

Conversely, we think of management costs for wholesale banks as being infinity. Finally, we allow households to directly fund this asset, where claims on this capital directly held by households may be thought of as corporate bonds. We suppose that households are at disadvantage to retail banks in funding the second type of capital, though at an advantage relative to wholesale banks: They must pay the management fee

$$F^L(K_t^L) = \frac{\alpha^L}{2}(K_t^L)^2$$

with $0 < \alpha^L < \infty$.

In analogy to the first type of capital, there is an exogenous dividend payout Z_t^L that obeys a stationary first order stochastic process. In addition, for simplicity we restrict attention to the case where bank runs are completely unanticipated. Accordingly, let R_{lt+1}^h be the household's rate of return from funding the second asset. Then the household's first order condition for holding the second asset is given by

$$E_t(\Lambda_{t,t+1}R_{lt+1}^h) = 1 \tag{49}$$

with

$$R_{lt+1}^h = \frac{Z_{t+1}^L + Q_{t+1}^L}{Q_t^L + \alpha_h^L L_t^h}$$

where Q_t^L is the asset price and α_h^L controls the degree of inefficiency of households in directly holding this asset.

The optimization problem of wholesale bankers is unchanged. Accordingly, we focus on retail bankers. Given retail banks now have the option of intermediating the second asset, we can rewrite the balance sheet and Flow of Funds constraints as

$$(Q_t + f_t^r)k_t^r + Q_t^L l_t^r + (-b_t^r) = n_t^r + d_t^r$$

$$n_{t+1}^r = R_{kt+1}^r\left(Q_t + f_t^r\right)k_t^r + R_{lt+1}^r Q_t^L l_t^r + R_{bt+1}(-b_t^r) - R_{t+1}d_t^r$$

where R_{lt+1}^r is the rate of return on the type L asset and is given by,

$$R_{lt+1}^r = \frac{Z_{t+1}^L + Q_{t+1}^L}{Q_t^L}.$$

Because the incentive constraint is

$$\theta[(Q_t + f_t^r)k_t^r + Q_t^L l_t^r + (-b_t^r)] \leq V_t^r,$$

the effective leverage multiple for this case ϕ_t^r now includes the holdings of the second type of capital:

$$\phi_t^r \equiv \frac{(Q_t + f_t^r)k_t^r + Q_t^L l_t^r + \gamma(-b_t^r)}{n_t^r}.$$

Proceeding as earlier to solve the retail bank's maximization problem yields a solution for ϕ^r_t which is the same as in the baseline case (see Eq. (27)). In addition, at the margin the retail bank must be indifferent between holding the types of capital, which implies the following arbitrage condition:

$$E_t[\Omega^r_{t+1}\left(R^r_{lt+1}-R^r_{kt+1}\right)]=0. \tag{50}$$

We now consider a numerical example designed to illustrate the contagion effect. The real world phenomenon that motivates the experiment is the fall in housing prices beginning in 2006 that led to the collapse of the wholesales banking sector that in turn disrupted commercial banking. In particular, we suppose that the dividend to asset L is fixed at its steady state value Z^L. Then we consider a negative shock to the dividend on the type K asset and, as in our earlier baseline experiments, allow for an unanticipated run two periods after the initial shock. Tables 4 and 5 describe the changes in the calibration for this experiment.

Table 4 Parameters in two assets model

	Parameters	
Households		
β	Discount rate	0.99
α^h	Intermediation cost	0.06
α^h_L	Intermediation cost for Cl loans	0.006
W^h	Endowment	0.016
Retail banks		
σ^r	Survival probability	0.96
α^r	Intermediation cost	0.01
α^r_L	Intermediation cost for Cl loans	0
W^r	Endowment	0.0014
θ	Divertable proportion of assets	0.27
γ	Shrinkage of divertable proportion of interbank loans	0.67
Wholesale banks		
σ^w	Survival probability	0.88
α^w	Intermediation cost	0
α^w_L	Intermediation cost for Cl loans	∞
W^w	Endowment	0.0012
ω	Shrinkage of divertable proportion of assets	0.47
Production		
Z	Steady state productivity	0.016
ρ_z	Serial correlation of productivity shocks	0.9

Table 5 Steady state in two assets model

		Steady state
Q	Price of capital	1
Q^L	Price of Cl loans	1
K^r	Retail intermediation	0.3
K^w	Wholesale intermediation	0.6
L^r	Retail holding of Cl loans	0.5
L^h	Household holding of Cl loans	0.5
R^b	Annual interbank rate	1.048
R_r^k	Annual retail return on capital	1.052
R_r^L	Annual retail return on Cl loans	1.052
R	Annual deposit rate	1.04
R_w^k	Annual wholesale return on capital	1.064
ϕ^w	Wholesale leverage	20
ϕ^r	Retail leverage	10
Y	Output	0.0466
C^h	Consumption	0.0363
N^r	Retail banks networth	0.1371
N^w	Wholesale banks networth	0.03

Fig. 15 reports the results from the experiment and demonstrates the spillover effects of shocks to Z_t on the market for L. The source of contagion in this environment is the balance sheet position of retail bankers.[ar] Losses on their capital investment and, in case of a run, on their interbank loans, result in a decrease in retail bankers' net worth and a tightening of their respective incentive constraints. As long as there are incentive costs associated with intermediating asset L, the tightening of financial constraints leads retail bankers to increase required excess returns in both markets, as shown by Eq. (50). The negative shock to returns on capital and the run on wholesale banks lead to a costly reallocation of assets to households and to an increase in spreads between returns on L_t^r and the deposit rate of about 60 basis points.

7. GOVERNMENT POLICY

In this section we study the effects of two types of policy interventions to combat banking crises: first an ex-post intervention where the central bank acts as a lender of last resort; second, an ex-ante macroprudential regulation that limits banks' risk exposure. Within the literature, these policies have largely been studied in the context of dampening negative financial accelerator effects on the economy. Here, we emphasize a somewhat

[ar] Other similar models of spillover are Bocola (2016) and Ferrante (2015b). An alternative mechanism based on market fragmentation is developed by Garleanu et al. (2015).

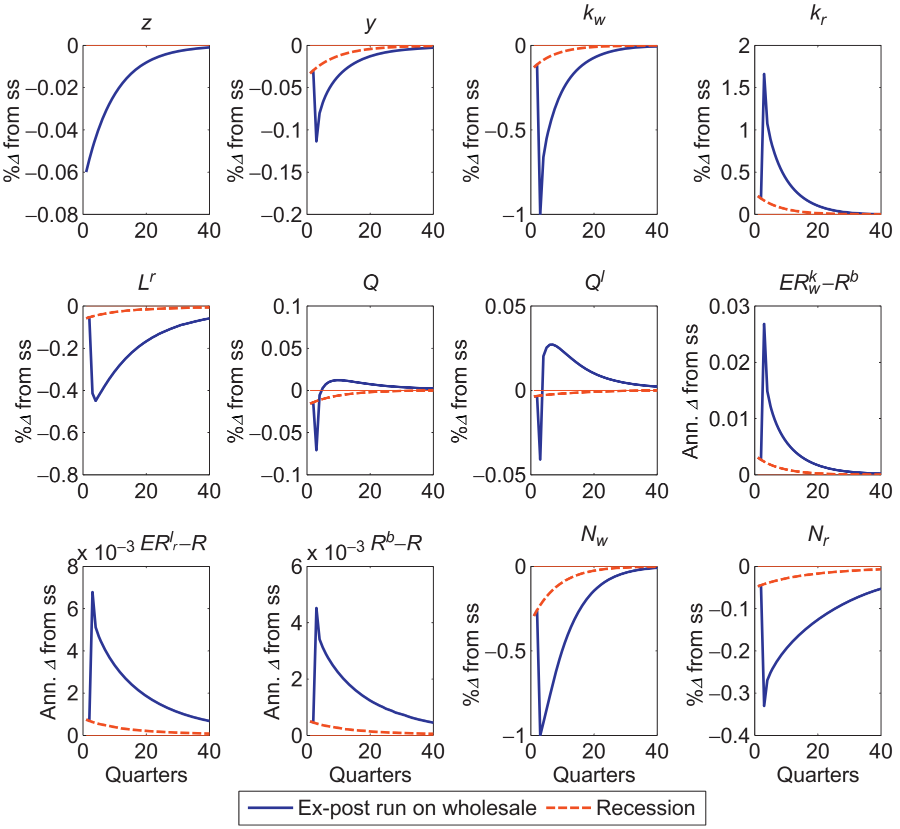


Fig. 15 Spillover.

different perspective: How these policies might be useful in reducing the likelihood of damaging bank runs? As we show, lender of last resort policy that is anticipated ex-ante in the event of an ex-post crisis reduces the likelihood of a run by raising asset liquidation prices. Macroprudential does so by reducing bank leverage.

A case for ex-ante macroprudential regulation arises because banks tend to choose an inefficiently high level of leverage in the laissez-faire economy. Roughly speaking, because individual banks ignore the consequences of their own borrowing decisions on the level of aggregate risk, they are prone to issue more debt than would be socially desirable.[as] In addition, as Farhi and Tirole (2012), Chari and Kehoe (2015), and Gertler et al. (2012) emphasize, the expectations of some type of government interventions ex-post will also encourage excessive leverage in the banking system ex-ante.

[as] See Geanakoplos and Polemarchakis (1986) for the original result of generic constrained inefficiency in a model with incomplete markets. Lorenzoni (2008) and Bianchi (2011) are recent applications to environments with financial frictions.

In this section we explore each of this kinds of policy's within our framework of Anticipated Runs of Section 5.

7.1 Ex-Post Intervention: Lender of the Last Resort

It is well known that if there are limits to arbitrage in private financial intermediation, then a central bank who plays as the lender of last resort during a financial crisis can enhance the flow of credit and in turn mitigate the economic downturn. What makes the lender of last resort effective is that the central bank can elastically obtain funds by issuing interest bearing reserves, while private financial intermediaries may be constrained in their ability to obtain funds by the condition of their balance sheets (Gertler and Karadi, 2011; Gertler and Kiyotaki, 2011).

Following the onset of the recent financial crisis, the Federal Reserve introduced a variety of lender of last resort programs. The most prominent involved large scale asset purchases (LSAPs) of high grade long-term debt, including primarily agency mortgage backed securities (AMBS), instruments that were held primarily in the shadow banking sector. The Fed announced this program in December 2008 following the collapse of the shadow banking system and began phasing it in the following March. The objective of this kind of lender of last resort intervention was to reduce the cost and thereby increase the availability of credit to the nonfinancial sector. There is evidence which suggests the Fed achieved this objective. Beyond these considerations, however, by acting as buyers in the secondary market for AMBS, the Fed raised the price and accordingly the liquidation value of these assets. As we noted, the impact of these policies on liquidation prices has important implications for banking stability. (See Eq. (40), for the condition for a bank run equilibrium.)

To model this type of intervention, we assume that the central bank can directly undertake intermediation by borrowing from retail banks and then making nonfinancial loans. The way the central bank obtains funds from retail banks is to issue interest bearing bank reserves. We assume that retail banks are unable to divert bank reserves, since they are held in an account at the Fed. Given retail banks cannot divert reserves, they are not constrained in their ability to raise deposits to fund reserves. Because there are no limits to arbitrage for banks funding reserves, the interest rate on reserves will equal the deposit rate. Therefore, when the central bank supplies interest-rate bearing reserves to retail banks, it effectively raises funds directly from households by issuing overnight government bond. What gives the central bank an advantage in intermediating assets is that, unlike retail and wholesale banks, it is not balance sheet constrained.

We also assume, following Gertler and Karadi (2011) that the central bank is less efficient than the private sector. As with retail banks and households, the government faces quadratic managerial costs $\frac{1}{2}\alpha^g (K_t^g)^2$, where K_t^g is the size of central bank's intervention and where $\alpha^h > \alpha^g > \alpha^r$. To ensure that it is desirable for the central bank to intervene only in a crisis, we also allow for inefficiency in the average performance of the government's portfolio: In particular, we assume that the return on government intermediated assets is:

$$R^g_{kt+1} = \varphi \frac{Z_{t+1} + Q_{t+1}}{Q_t + \alpha^g K^g_t} \tag{51}$$

where $\varphi \in (0,1)$ controls the relative inefficiency of central bank's intermediation for the average return on assets, independent of scale.

We assume that the central bank intervenes in credit markets whenever expected asset returns exceed its cost of borrowing. That is we posit a policy rule for central bank's intervention given by

$$\begin{array}{ll} K^g_t = 0, & \text{if } E_t\left(R^g_{kt+1} - R_{gt+1}\right) < 0 \\ E_t\left(R^g_{kt+1} - R_{gt+1}\right) = 0, & \text{if } K^g_t \geq 0 \end{array} \tag{52}$$

where R_{gt+1} is the interest paid on reserves issued to retail banks.

As we just noted, since there is no incentive problem associated with central bank intermediation, in equilibrium the interest rate on reserve R_{gt+1} must equal to the deposit rate:[at]

$$R_{gt+1} = R_{t+1}. \tag{53}$$

The key variable to which the central bank responds in determining credit market intervention is the spread between the wholesale bank's return on assets and the deposit rate, $R^w_{kt+1} - R_{t+1}$, which can be thought of as a measure of the degree of inefficiency in private financial markets. The central bank intervenes when this excess return is high.[au] In particular, the policy rule (52) prescribes that the Fed starts intermediating assets as soon as the ratio of the credit spread to the deposit rate exceeds a given threshold that varies inversely with the inefficiency parameter φ :

$$K^g_t > 0, \text{ iff } \frac{E_t(R^w_{kt+1}) - R_{t+1}}{R_{t+1}} > \frac{1-\varphi}{\varphi}.$$

[at] To see formally, first notice that, since retail bankers cannot divert reserves, their incentive constraint (14) is not affected by the amount of reserves held on their balance sheet. Hence the introduction of interest bearing reserves only affects retail bankers' optimization problem by modifying the objective function (26), which becomes

$$V^r_t = \underset{\phi^r_t, d^r_{gt}}{Max} E_t\left\{\Omega^r_{t+1}\left[\phi^r_t(R^r_{kt+1} - R_{t+1}) + R_{t+1} + d^r_{gt}(R_{gt+1} - R_{t+1})\right]n^r_t\right\}$$

where d^r_{gt} is the amount of reserves per unit of networth held by retail bankers. The optimality condition with respect to d^r_{gt} is just given by $R_{gt+1} = R_{t+1}$. Covariance terms are zero since both R_{gt+1} and R_{t+1} are known at date t.

[au] Our policy rule, which has the central bank target credit spreads, is consistent with how the central bank behaved throughout the crisis. What motivated an unconventional intervention in a given credit market was typically a sharp increase in the spread within that market.

From Eq. (52), the size of the intervention in the region where $K_t^g > 0$ is then governed by:

$$K_t^g = \frac{\varphi}{\alpha^g} Q_t \left[\frac{E_t(R_{kt+1}^w) - R_{t+1}}{R_{t+1}} - \frac{1-\varphi}{\varphi} \right].$$

We choose φ in order to ensure that the central bank only intervenes after a run happens: that is, the threshold for the credit spread to justify an intervention is reached only in the event of a run. We choose the management cost parameter α^g in order for the intervention to be around 5% of total capital.

Fig. 16 shows the response of the economy to a recession when agents anticipate that, if a run happens, the monetary authority intervenes with large scale asset purchases according to (52). Even though in this experiment the run does not happen and the central bank accordingly does not intervene, the anticipation of the intervention in the event of a run significantly dampens the downturn. It does so by reducing the probability of a run: The central bank's conditional intervention policy increases the liquidation price of

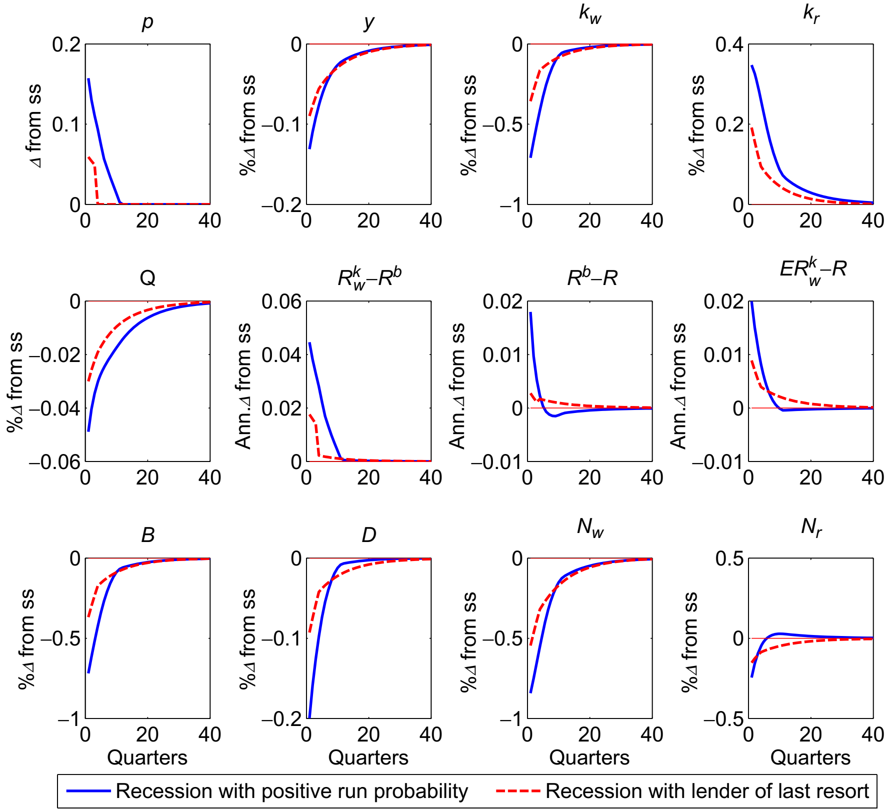


Fig. 16 Anticipation effects of government intervention.

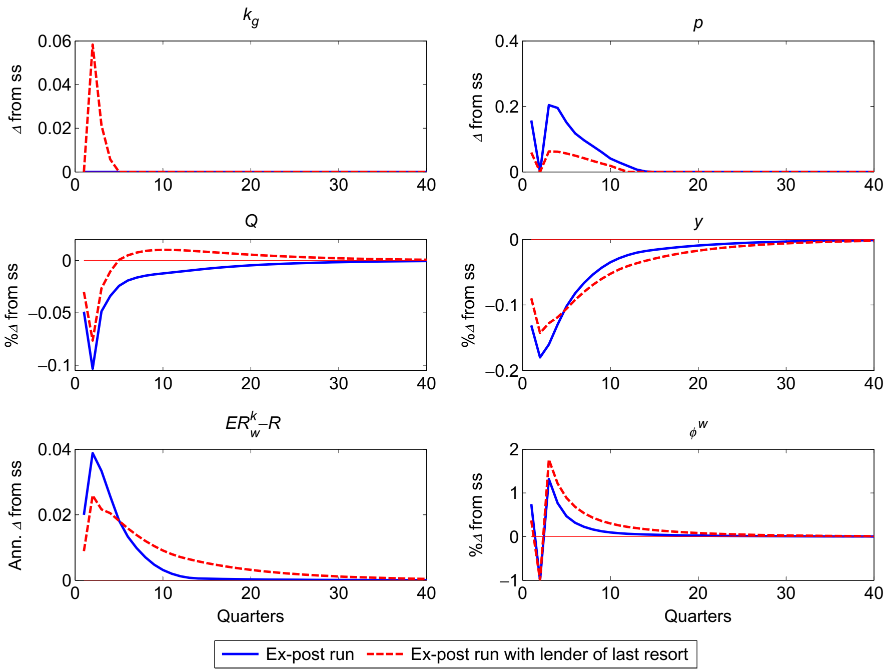


Fig. 17 Government intervention when a run happens at time 2.

wholesale banks assets. In turn, by Eq. (47), the higher recovery rate associated with higher liquidation prices decreases the probability of a run. In the experiment the probability of a run decreases by 10% in the first two periods and becomes zero thereafter. This drastic reduction in the run probability implies that, overall, anticipation of government intervention works to stimulate the economy. Notice that, even though the reduction in the run probability relaxes the incentive constraint and hence allows wholesale bankers to increase their leverage for any given level of spreads, the general equilibrium effects of asset prices on their balance sheet results in better capitalization and lower leverage in both the wholesale and retail bank sectors.

Fig. 17 illustrates the effect of the intervention when a run happens one period after the shock to Z. The intervention is around 5% of total capital and reduces the drop in asset prices and output by about 2.5 and 4%, respectively.

7.2 Ex-Ante Intervention: Macroprudential Policy

One of the most important challenges facing policy makers in the aftermath of the financial crisis is the development of financial regulations that can help prevent the recurrence of similar episodes in the future. In this respect, the most relevant innovation in the policy landscape has been the introduction of various macroprudential measures in the oversight of financial institutions, such as stress tests by central banks and the revised provisions in Basel III. These measures are aimed at ensuring that financial institutions' capital is sufficient to absorb losses during adverse economic conditions.

There is now a significant literature that analyzes the impact of capital requirements on banks for macroeconomic stability (eg, Christiano and Ikeda, 2014; Begenau, 2015; Bianchi and Mendoza, 2013; Chari and Kehoe, 2015; Gertler et al., 2012). Most of this literature analyzes how the introduction of leverage restrictions can dampen financial accelerator effects by dampening fluctuations in bank capital. The need for leverage restrictions, or equivalently capital requirements, stems from an externality that leads individual banks to fail to take into account the effect of their own borrowing on the stability of the system as a whole.[av]

Our framework offers a somewhat different perspective on the potential benefits of leverage restrictions. Not only can these restrictions dampen financial accelerator effects: Importantly, they can also make the banking system less susceptible to runs. As Eq. (41) makes clear, a bank run can only happen if the leverage ratio is high enough. Thus, by limiting the leverage ratio sufficiently, the regulatory authority can in principle eliminate the possibility of a run. The question then is what are the tradeoffs. We turn to this issue next.

We capture macroprudential policies in our model economy by introducing leverage restrictions on wholesale banks. In particular, we assume that a financial regulator can impose an upper bound on wholesale banks' leverage, $\overline{\phi}^w$. This implies that the effective limit to wholesale banks' leverage will be given by the smaller between the market imposed limit and the regulatory limit. Accordingly, constraint (22) becomes

$$\phi^w \leq \min\left\{\frac{\frac{1}{\theta}\frac{V_t^w}{n_t^w}-(1-\omega)}{\omega}, \overline{\phi}^w\right\}$$

In a fully stochastic simulation of the economy, leverage restrictions would tradeoff lower frequency of crises, resulting from reduced variation of bankers' capital, against lower average output, as the impaired ability of wholesale banks to increase their leverage would induce a costly reallocation of capital to less efficient agents. While our numerical experiments in Sections 4.2 and 4.3 provide an illustration of the tradeoff between steady state output and fragility associated to changes in the long run level of wholesale bankers'

[av] Much of the literature, following Lorenzoni (2008), features a pecuniary externality stemming from the presence of asset prices in the borrowing constraint. Farhi and Werning (2015) and Korinek and Simsek (2015) show that if aggregate demand is sensitive to aggregate leverage, a similar kind of externality can emerge.

leverage, here we focus on the conditional effects of leverage restrictions upon the occurrence of a recession that would leave the decentralized economy vulnerable to bank runs.

We focus on two possible levels for $\overline{\phi}^{w}$: the steady state level of wholesale banks' leverage and a level that is higher than steady state but still sufficiently low to prevent a run. Permitting a leverage ratio above the steady state allows banks to issue more debt in a recession, which has the overall effect of dampening the contraction in financial intermediation and thus dampening the downturn in real activity. Indeed, the more forgiving leverage restriction comes closer to mimicking the behavior of the leverage ratio in the decentralized economy, which moves countercylcially.

Figs. 18 and 19 compare the response of the economy with anticipated runs to a negative Z innovation, with and without macroprudential regulation. In Fig. 18 the regulator imposes the tighter leverage restriction, ie, $\overline{\phi}^{w}$ is set to the steady state value of wholesale leverage, while in Fig. 19 the restrictions are more lax and allow maximum regulatory leverage to exceed the steady state value by 15%. As mentioned, in both cases, the leverage restrictions are sufficient to prevent a run and hence avoid the recessionary effects associated to the endogenous increase in the probability of a run that characterizes

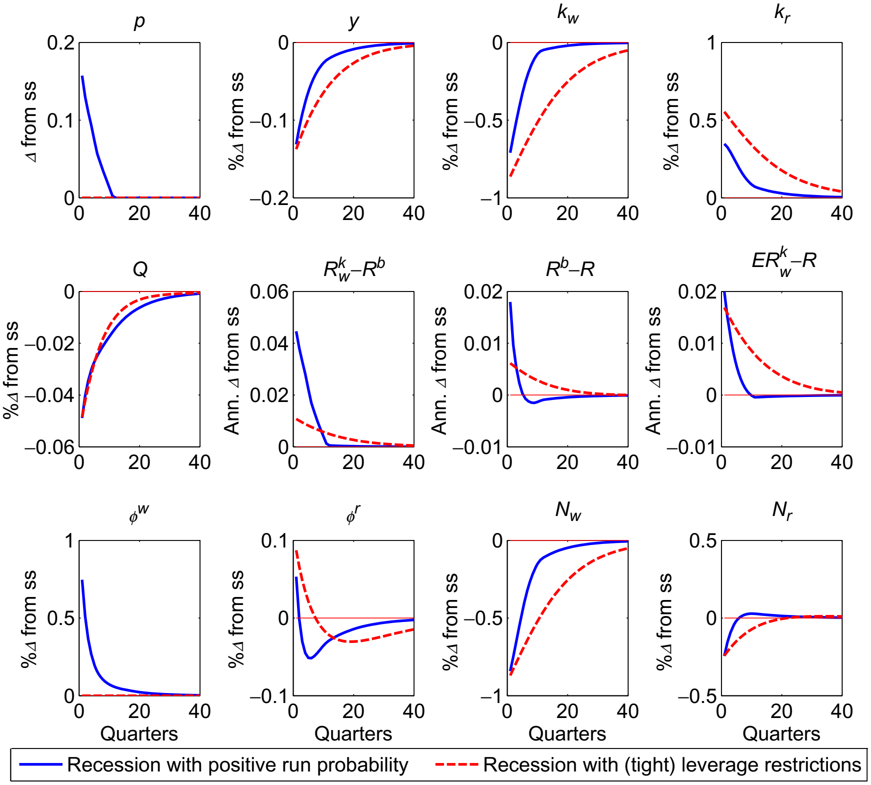


Fig. 18 Macro prudential policy: $\phi^{w} = \phi^{ss}$.

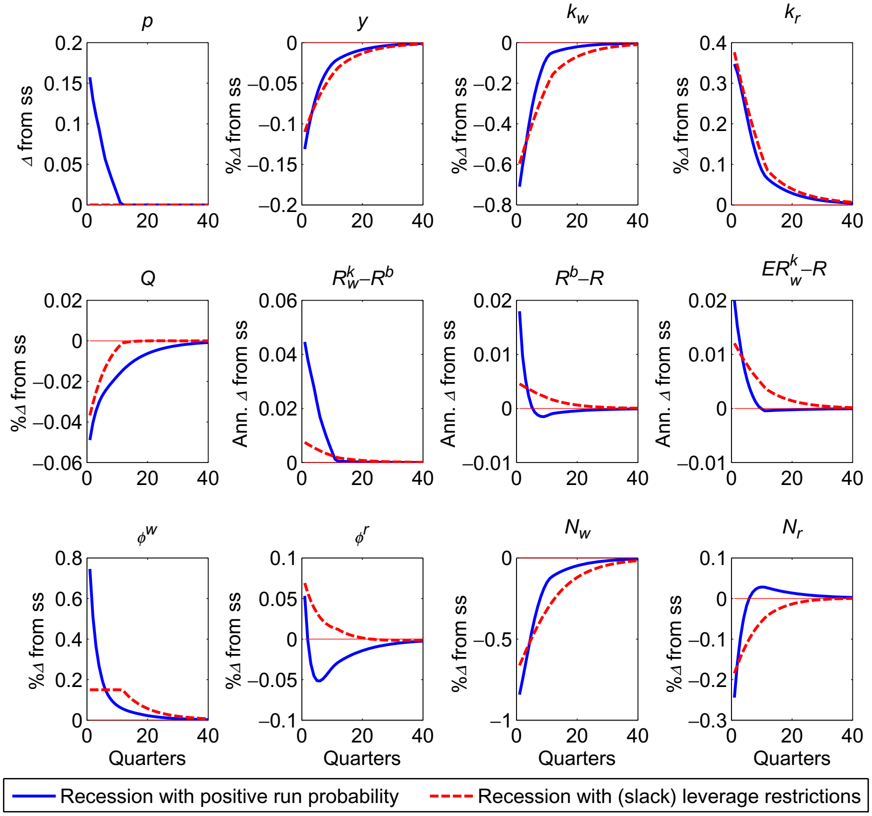


Fig. 19 Macro prudential policy: $\phi^w = 1.15\phi^{ss}$.

the unregulated economy. This results in higher asset prices in the regulated economy throughout the recession. Under the less strict requirements the stimulative effect on asset prices is significantly higher, reaching about 1.5% after the first three years of the recession. On the other hand, by constraining the ability to leverage of the most efficient intermediaries, macroprudential policies induces a costly reallocation of assets. The balance between these two contrasting forces varies overtime, in turn influencing output effects of the policy.

During the early stages of the recession, the stimulative effects of macroprudential policy are strongest because they eliminate the probability of a bank run, which in the unregulated economy is highest at this time. Under the stricter policy, the impact drop in output is very similar to the drop in the unregulated economy, while the more lax stance of policy dampens the drop in output by 2% and is stimulative throughout the first year of the recession. As time passes, the probability of a run becomes small in the unregulated economy, implying that the stimulative effects of policy decreases. On the other hand, the slower recovery of financial institutions' equity in the regulated economy that results from their impaired ability to leverage, implies a more persistent drag on output coming from financial misallocation. In both cases output costs associated with the policy peak at around 10 quarters into the recession and result in an additional drop in output of about 4% under the tighter requirements and 1.5% under the more lax stance.

8. SUMMARY AND DIRECTIONS FOR FUTURE RESEARCH

The financial crisis that triggered the Great Recession featured a disruption of wholesale funding markets, where banks lend to one another, as opposed to retail markets where banks obtain funds from depositors. It is essential to capture the roles and possible disruption of wholesale funding market to understand the financial crisis as well as to draw policy implications. Our goal in this Handbook Paper was to sketch a model based on the existing literature that provides a step toward accomplishing this objective. The model first accounts for how, through innovation in the efficiency of interbank loan markets, a wholesale banking sector emerges that intermediates loans using funds borrowed from retail banks. This wholesale sector bears a close resemblance to the shadow banking system featured in most descriptions of the crisis.

As we show, in "normal" times, the growth of the wholesale banking sector improves both efficiency and stability. Improved efficiency stems from the comparative advantage that wholesale banks having in managing certain types of loans. Improved stability arises because retail banks act as a buffer to absorb loans that wholesale banks sell off, in effect improving the liquidity of secondary loan markets. On the other hand, the growth of wholesale banking system makes the economy more vulnerable to a crisis. As occurred in practice, the high leverage of wholesale banks makes this sector susceptible to runs that can have highly disruptive effects on the economy. A contractionary disturbance that might otherwise lead to a moderate recession, can induce a run on the wholesale banking sector with devastating effects on the economy, as experienced during the Great Recession. We then describe how both lender of last resort and macroprudential policies can help reduce the likelihood of these kinds of banking crises.

Our framework also captures the buildup of safe assets prior to the crisis along with the subsequent collapse that a number of authors have emphasized (eg, Gorton and Metrick, 2015; Caballero and Farhi, 2015). The underlying mechanisms work a bit differently, in somewhat subtle ways: The "safe asset" literature points to an increased demand for safe assets as the driving force in the buildup of the shadow banking system. By making assets riskier, the crisis then reduces the ability of the shadow banking sector to create safe assets. It is this reduction in safe assets that then leads a contraction in spending, essentially for liquidity reasons. Within our framework, the increase in safe assets is a product of innovation in interbank lending markets. Indeed, this is where much of the growth in safe assets occurred. There is also a growth in households deposits as the overall banking system becomes more efficient. The crisis similarly induces a contraction in safe assets: The exact mechanism, though, is that, with an adverse shock to the net worth of banks, the probability of runs on wholesale banks becomes positive, which constrains the ability of both wholesale and retail banks to issue safe liabilities. In turn, a contraction in real activity emerges because the costs of intermediation increase, as manifested by the increase in credit spreads. In future work, it would be interesting to synthesize the role of safe assets in our framework with that in the conventional literature on this topic.

Another important area for further investigation involves the modeling of the growth of wholesale banking. Our approach was to treat this growth as the product of innovation as captured by a reduction in the agency friction in interbank lending markets. Among the factors we had in mind that motivate this reduction is technological improvements that permit less costly monitoring, such as the development of asset-backed securities and repo lending. Of course, more explicit modeling of this phenomenon would be desirable. Also important is integrating regulatory considerations. While financial innovation was important for the development of shadow banking, regulatory factors also played an important role. For example, tightening of capital requirements on commercial banks in conjunction with innovation in asset securitization induced movement of a considerable amount of mortgage lending from the retail to the wholesale banking sector. A careful integration of the roles of regulation and innovation in the development of wholesale banking would be highly desirable.

Finally, consistent with what occurred in the recent crisis, what makes the financial system within our model so vulnerable is high degree of leverage in the form of short-term debt. Here, we simply rule out a richer set of state-contingent financial contracts that would permit banks to hedge against the systemic risk implied by this liability structure. Why in practice we don't seem to observe the kind of seemingly desirable hedging is an important question for future research.[aw]

APPENDICES

Appendix A Details of the Equilibrium

From (13, 15–17), we get

$$
\begin{aligned}
\frac{V_t^j}{n_t^j} &= E_t\left(\Omega_{t+1}^j \cdot \frac{n_{t+1}^j}{n_t^j}\right)\\
&= E_t\left\{\Omega_{t+1}^j\left[R_{kt+1}^j + \left(R_{kt+1}^j - R_{t+1}\right)\frac{d_t^j}{n_t^j} + \left(R_{kt+1}^j - R_{bt+1}\right)\frac{b_t^j}{n_t^j}\right]\right\}\\
&= \nu_{kt}^j + \mu_{dt}^j\frac{d_t^j}{n_t^j} + \mu_{bt}^j\frac{b_t^j}{n_t^j},
\end{aligned}
$$

[aw] Some efforts to address this issue include Krishnamurthy (2003), Di Tella (2014), Gertler et al. (2012), and Dang et al. (2012).

where

$$\nu^j_{kt} \quad = E_t(\Omega^j_{t+1} R^j_{kt+1}) \tag{A.1}$$

$$\mu^j_{dt} \quad = E_t\left[\Omega^j_{t+1}\left(R^j_{kt+1} - R_{t+1}\right)\right] \tag{A.2}$$

$$\mu^j_{bt} \quad = E_t\left[\Omega^j_{t+1}\left(R^j_{kt+1} - R_{bt+1}\right)\right]. \tag{A.3}$$

From (13), the incentive constraint (14) can be written as

$$V^j_t \geq \theta\left[n^j_t + d^j_t + \omega b^j_t \cdot I_{b^j_t > 0} + (1-\gamma) b^j_t \cdot I_{b^j_t < 0}\right],$$

where $I_{b^j_t>0} = 1$ if $b^j_t > 0$ and $I_{b^j_t>0} = 0$ otherwise, (and $I_{b^j_t<0} = 1$ if $b^j_t < 0$ and $I_{b^j_t<0} = 0$ otherwise).

In order to save the notations, we normalize $n^j_t = 1$ and suppress the suffix and time subscript. The generic choice of a bank is given by

$$\psi = \underset{b,\,d}{Max}(\nu_k + \mu_d d + \mu_b b) \tag{A.4}$$

subject to

$$\theta[1 + d + \omega b \cdot I_{b>0} + (1-\gamma) b \cdot I_{b<0}] \leq \nu_k + \mu_d d + \mu_b b, \tag{A.5}$$

$$d \geq 0,$$

$$1 + d + b \geq 0.$$

Figs. A.1 and A.2 depict the Feasible set and an Indifference Curve for Wholesale Bankers and Retail Bankers under our baseline.

Defining λ and λ_k as Lagrangian multipliers of the incentive constraint and the non-negativity constraint of capital, we have the Lagrangian as

$$\mathcal{L} = (1+\lambda)(\nu_k + \mu_d d + \mu_b b) - \lambda\theta[1 + d + \omega b \cdot I_{b>0} + (1-\gamma) b \cdot I_{b<0}] + \lambda_k(1 + d + b).$$

For the case of $b \geq 0$, we know $\lambda_k = 0$ and the first order conditions are

$$(1+\lambda)\mu_b \leq \lambda\theta\omega,$$

where $=$ holds if $b > 0$, and $<$ implies $b = 0$.

$$(1+\lambda)\mu_d \leq \lambda\theta,$$

where $=$ holds if $d > 0$, and $<$ implies $d = 0$.

In the following we restrict the attention to the case of $\mu_d > 0$, and will verify the inequality later. Thus for the case of $b > 0$, we learn

$$d > 0, \text{ if } \frac{\mu_b}{\mu_d} = \omega,$$

$$d = 0, \text{ if } \frac{\mu_b}{\mu_d} > \omega.$$

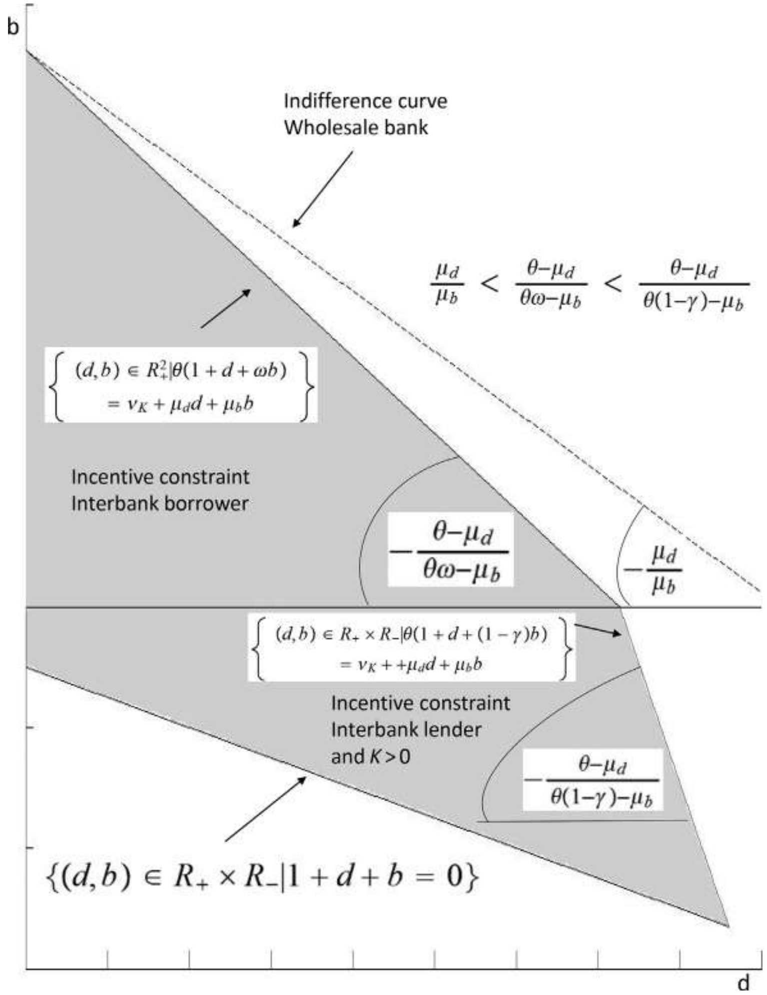

b
Indifference curve
Wholesale bank
$\frac{\mu_d}{\mu_b} < \frac{\theta-\mu_d}{\theta\omega-\mu_b} < \frac{\theta-\mu_d}{\theta(1-\gamma)-\mu_b}$
$\left\{ \begin{matrix} (d,b) \in R_+^2 | \theta(1+d+\omega b) \\ = \nu_K + \mu_d d + \mu_b b \end{matrix} \right\}$
Incentive constraint
Interbank borrower
$-\frac{\theta-\mu_d}{\theta\omega-\mu_b}$
$-\frac{\mu_d}{\mu_b}$
$\left\{ \begin{matrix} (d,b) \in R_+ \times R_- | \theta(1+d+(1-\gamma)b) \\ = \nu_K + +\mu_d d + \mu_b b \end{matrix} \right\}$
Incentive constraint
Interbank lender
and $K>0$
$-\frac{\theta-\mu_d}{\theta(1-\gamma)-\mu_b}$
$\{(d,b) \in R_+ \times R_- | 1+d+b=0\}$
d

Fig. A.1 Wholesale banker's optimization.

For the case of $b \leq 0$, the first order conditions are

$$(1+\lambda)\mu_b + \lambda_k \geq \lambda\theta(1-\gamma),$$

where $=$ holds if $b<0$, and $>$ implies $b=0$.

$$(1+\lambda)\mu_d + \lambda_k \leq \lambda\theta,$$

where $=$ holds if $d>0$, and $<$ implies $d=0$.

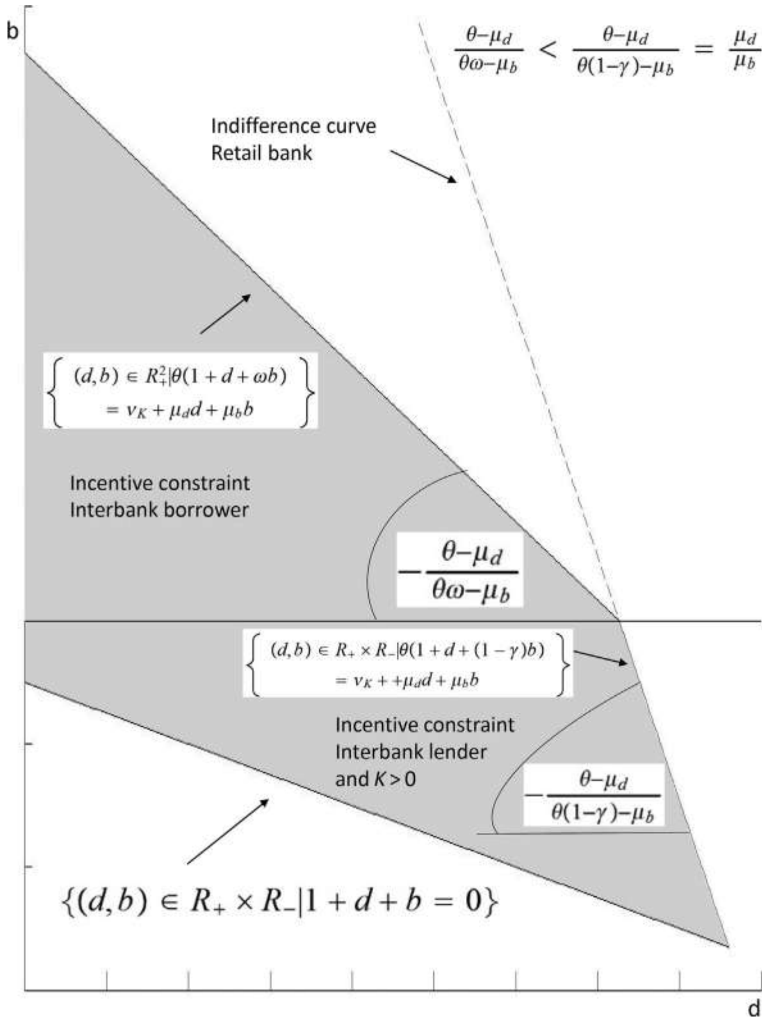

b
$\frac{\theta-\mu_d}{\theta\omega-\mu_b} < \frac{\theta-\mu_d}{\theta(1-\gamma)-\mu_b} = \frac{\mu_d}{\mu_b}$
Indifference curve
Retail bank
$\left\{ \begin{array}{c} (d,b) \in R_+^2 \vert \theta(1+d+\omega b) \\ = v_K + \mu_d d + \mu_b b \end{array} \right\}$
Incentive constraint
Interbank borrower
$-\frac{\theta-\mu_d}{\theta\omega-\mu_b}$
$\left\{ \begin{array}{c} (d,b) \in R_+ \times R_- \vert \theta(1+d+(1-\gamma)b) \\ = v_K + +\mu_d d + \mu_b b \end{array} \right\}$
Incentive constraint
Interbank lender
and $K > 0$
$-\frac{\theta-\mu_d}{\theta(1-\gamma)-\mu_b}$
$\{(d,b) \in R_+ \times R_- \vert 1+d+b=0\}$
d

Fig. A.2 Retail banker's optimization.

Thus for the case of $b < 0$ and $d > 0$, we learn

$$k > 0, \text{ if } \frac{\mu_b}{\mu_d} = 1-\gamma,$$

$$k = 0 \text{ and } \lambda_k > 0, \text{ if } \frac{\mu_b}{\mu_d} < 1-\gamma.$$

Therefore, under Assumption 2: $\omega + \gamma > 1$, we can summarize the bank's choice as:

(i) $b > 0,\ d = 0,\ k > 0$, if $\mu_b > \omega\mu_d$

(ii) $b > 0,\ d > 0,\ k > 0$, implies $\mu_b = \omega\mu_d$

(iii) $b = 0$, $d > 0$, $k > 0$, if $(1-\gamma)\mu_d < \mu_b < \omega\mu_d$
(iv) $b < 0$, $d > 0$, $k > 0$, implies $\mu_b = (1-\gamma)\mu_d$
(v) $b < 0$, $d > 0$, $k = 0$, if $\mu_b < (1-\gamma)\mu_d$.
In the steady state equilibrium, we know

$$\frac{\mu_b}{\mu_d} = \frac{R_k - R_b}{R_k - R}.$$

Because we know $R_k^w \geq R_k^r$ and $R_b \geq R$, we learn

$$\frac{\mu_b^w}{\mu_d^w} \geq \frac{\mu_b^r}{\mu_d^r}.$$

Therefore, market clearing for interbank loans implies that, if the interbank market is active wholesale bankers' choice can only be (i) or (ii) and retail banker's choice (iv) or (v). Otherwise both types must choose according to (iii) and the interbank market is inactive. That is, we have only the following possible patterns of equilibrium in the neighborhood of the steady state.

(A) Perfect Specialization with active Interbank Market: $d^w = 0, k^r = 0, b^w > 0 > b^r$
(B) Perfect Specialized Retail Banks with active Interbank Market: $d^w > 0, k^r = 0, b^w > 0 > b^r$
(C) Perfect Specialized Wholesale Banks with active Interbank Market: $d^w = 0, k^r > 0, b^w > 0 > b^r$
(D) Imperfect Specialization with active Interbank Market: $d^w > 0, k^r > 0, b^w > 0 > b^r$
(E) Inactive Interbank Market: $d^w > 0, k^r > 0, b^w = 0 = b^r$.
We can show that, under Assumption 2, there is no equilibrium of type (A) nor (B):

Proof. Equilibrium of type (A) and (B) require $\mu_b^w \geq \omega\mu_d^w$ and $(1-\gamma)\mu_d^r \geq \mu_b^r$. Thus

$$R_b \leq \omega R + (1-\omega)R_k^w,$$
$$R_b \geq (1-\gamma)R + \gamma R_k^r = (1-\gamma)R + \gamma R_k^w, \text{ as } K^r = 0 \text{ in (A)and (B)}.$$

This implies

$$\omega R + (1-\omega)R_k^w \geq (1-\gamma)R + \gamma R_k^w,$$

or

$$(\omega + \gamma - 1)R \geq (\omega + \gamma - 1)R_k^w.$$

But this is a contradiction as $\omega + \gamma > 1$ and $R_k^w > R$ (as $\mu_d^w > 0$ under our conjecture).

Equilibrium C and D: Active Interbank Market
Suppose that $0 < \mu_{bt}^{w} < \theta\omega$. We will verify this numerically after we characterize the equilibrium. Then the incentive constraint (A.5) holds with equality for wholesale banks. Together with Bellman equation (A.4), we have

$$\begin{aligned}\psi_t^w &= \nu_{kt}^w + \mu_{dt}^w d_t^w + \mu_{bt}^w b_t^w \\ &= \theta\left(1 + d_t^w + \omega b_t^w\right),\end{aligned}$$

or

$$b_t^w = \frac{1}{\theta\omega - \mu_{bt}^w}\left[\nu_{kt}^w - \theta - (\theta - \mu_{dt}^w) d_t^w\right],$$

$$\psi_t^w = \frac{\theta}{\theta\omega - \mu_{bt}^w}\left[\omega\nu_{kt}^w - \mu_{bt}^w + (\omega\mu_{dt}^w - \mu_{bt}^w) d_t^w\right].$$

Maximizing Tobin's Q, ψ_t^w, with respect to $d_t^w \geq 0$, we learn

$$d_t^w = 0, \text{ if } \mu_{dt}^w < \frac{1}{\omega}\mu_{bt}^w$$

$$d_t^w > 0 \text{ implies } \mu_{dt}^w = \frac{1}{\omega}\mu_{bt}^w.$$

This proves Lemma 1 and the argument in the text follows for wholesale banks, noting that we normalize $n_t^w = 1$ above.

Suppose also that $0 < \mu_{dt}^r < \theta$. We will verify this numerically after we characterize the equilibrium. Then the incentive constraint (A.5) holds with equality for retail banks. Together with Bellman equation (A.4), we have

$$\begin{aligned}\psi_t^r &= \nu_{kt}^r + \mu_{dt}^r d_t^r + \mu_{bt}^r b_t^r \\ &= \theta[1 + d_t^r + (1-\gamma) b_t^r].\end{aligned}$$

Then we get

$$d_t^r = \frac{1}{\theta - \mu_{dt}^r}\{\nu_{kt}^r - \theta + [\theta(1-\gamma) - \mu_{bt}^r](-b_t^r)\},$$

$$\psi_t^r = \frac{\theta}{\theta - \mu_{dt}^r}\left[\nu_{kt}^r - \mu_{dt}^r + (\mu_{dt}^r - \mu_{bt}^r - \gamma\mu_{dt}^r)(-b_t^r)\right].$$

Maximizing Tobin's Q, ψ_t^r, with respect to $k_t^r \geq 0$ and $b_t^r \leq 0$, we learn

$$k_t^r > 0 \text{ and } b_t^r < 0 \text{ imply } \mu_{dt}^r - \mu_{bt}^r = \gamma\mu_{dt}^r$$

$$k_t^r = 0 \text{ and } b_t^r < 0 \text{ if } \mu_{dt}^r - \mu_{bt}^r > \gamma\mu_{dt}^r.$$

This proves Lemma 2 and the argument in the text follows for retail banks, noting that we normalize $n_t^r = 1$ above.

Therefore the argument in the text follows for the aggregate equilibrium.

Equilibrium E: No Active Interbank Market $b_t^w = b_t^r = 0$

From Bellman equation and the incentive constraint of each bank (A.4, A.5) with $\left(Q_t + f_t^j k_t^j\right) k_t^j = 1 + d_t^j$, we have

$$\psi_t^j = \theta\left(Q_t + f_t^j k_t^j\right) k_t^j = \nu_{kt}^j - \mu_{dt}^j + \mu_{dt}^j \left(Q_t + f_t^j k_t^j\right) k_t^j,$$

or

$$\begin{aligned} \left(Q_t + f_t^j k_t^j\right) k_t^j &= \frac{\nu_{kt}^j - \mu_{dt}^j}{\theta - \mu_{dt}^j}, \\ \psi_t^j &= \theta \frac{\nu_{kt}^j - \mu_{dt}^j}{\theta - \mu_{dt}^j} \end{aligned} \tag{A.6}$$

The aggregate balance sheet conditions of wholesale and retail banking sectors are

$$Q_t K_t^w = \frac{\nu_{kt}^w - \mu_{dt}^w}{\theta - \mu_{dt}^w} N_t^w = N_t^w + D_t^w \tag{A.7}$$

$$\left(Q_t + f_t^r K_t^r\right) K_t^r = \frac{\nu_{kt}^r - \mu_{dt}^r}{\theta - \mu_{dt}^r} N_t^r = N_t^r + D_t^r. \tag{A.8}$$

The recursive competitive equilibrium without bank runs consists of 24 variables-aggregate quantities $\left(K_t^w, K_t^r, K_t^h, D_t^w, D_t^r, N_t^w, N_t^r, C_t^b, C_t^h, \bar{Y}_t, Y_t\right)$, prices $\left(Q_t, R_{t+1}, f_t^r\right)$ and bankers' franchise values and leverage multiples $\left(\Omega_t^j, R_{kt}^j, \nu_{kt}^j, \mu_{dt}^j, \psi_t^j\right)_{j=w,r}$- as a function of the state variables $\left(K_{t-1}^w, K_{t-1}^r, R_t D_{t-1}^w, R_t D_{t-1}^r, Z_t\right)$, which satisfy 24 equations (1, 4, 7, 8, 16, 18, 34–39, A.1, A.2, A.6–A.8) where each of (16, 18, A.1, A.2, A.6–A.8) contain two equations.

After finding the equilibrium, we need to check the inequalities

$$\begin{aligned} \mu_{bt}^w &< \omega \mu_{dt}^w, \\ \mu_{bt}^r &> (1-\gamma)\mu_{dt}^r. \end{aligned}$$

In the neighborhood of the steady state, it is sufficient to show

$$(1-\omega)E_t\left(\frac{Q_{t+1} + Z_{t+1}}{Q_t}\right) + \omega R_{t+1} < \gamma E_t\left(\frac{Q_{t+1} + Z_{t+1}}{Q_t + \alpha^r K_t^r}\right) + (1-\gamma)R_{t+1}. \tag{A.9}$$

Appendix B Steady State of the Economy Without Run

In order to characterize the steady state of (C,D,E), define x^j as the growth rate of the net worth of continuing bank j in the steady state:

$$x^j = \frac{n^j_{t+1}}{n^j_t} = R^j_k \frac{(Q+f^j)k^j}{n^j} - R_b \frac{b^j}{n^j} - R\frac{d^j}{n^j}$$

$$= \left(R^j_k - R_b\right)\frac{b^j}{n^j} + \left(R^j_k - R\right)\frac{d^j}{n^j} + R^j_k.$$

Then we have the aggregate net worth of bank j as

$$N^j = \sigma^j x^j N^j + W^j$$

$$= \frac{W^j}{1-\sigma^j x^j} \equiv N^j(x^j),$$

if $\sigma^j x^j < 1$, which we guess and verify later. Tobin's Q of bank j is

$$\psi^j = \beta(1-\sigma^j + \sigma^j\psi^j)x^j$$

$$= \frac{\beta(1-\sigma^j)x^j}{1-\beta\sigma^j x^j} \equiv \psi^j(x^j).$$

The ratio of bank loans to net worth is

$$\frac{Qk^w}{n^w} = \frac{\psi^w(x^w)}{\theta\omega} - \frac{1-\omega}{\omega}\left(1+\frac{d^w}{n^w}\right), \text{ if } b^w > 0,$$

$$\frac{Qk^w}{n^w} = \frac{\psi^w(x^w)}{\theta}, \text{ if } b^w = 0,$$

$$\frac{(Q+f^r)k^r}{n^r} = \frac{\psi^r(x^r)}{\theta} - \gamma\left(-\frac{b^r}{n^r}\right).$$

Case of Active Interbank Market: C and D

From the condition for the retail banks, we have

$$1-\gamma = \frac{\mu^r_b}{\mu^r_d} = \frac{R^r_k - R_b}{R^r_k - R},$$

or

$$R_b = \gamma R^r_k + (1-\gamma)R.$$

$$\begin{aligned}
x^r - R &= (R_k^r - R_b)\frac{b^r}{n^r} + \left(R_k^r - R\right)\left(1 + \frac{d^r}{n^r}\right) \\
&= \left(R_k^r - R\right)\left[1 + \frac{d^r}{n^r} + (1-\gamma)\frac{b^r}{n^r}\right] \\
&= \left(R_k^r - R\right)\left[\frac{(Q+f^r)k^r}{n^r} + \gamma\left(-\frac{b^r}{n^r}\right)\right] \\
&= \left(R_k^r - R\right)\frac{\psi^r(x^r)}{\theta}.
\end{aligned}$$

Thus from $R = \beta^{-1}$,

$$\beta(R_k^r - R) = \theta\frac{\beta x^r - 1}{\psi^r(x^r)} = \theta\frac{(\beta x^r - 1)(1-\sigma^r\beta x^r)}{(1-\sigma^r)\beta x^r} \equiv \varphi^r(\beta x^r),$$

$$\beta(R_b - R) = \gamma\theta\frac{\beta x^r - 1}{\psi^r(x^r)} = \gamma\varphi^r(\beta x^r).$$

Thus R_k^r and R_b are functions of only x^r :

$$R_k^r = R_k^r(x^r), R_b = R_b(x^r).$$

Differentiating log of the right hand side (RHS) of the above equation with respect to x^r, we learn

$$\begin{aligned}
\frac{d\ \ln\varphi^r(\beta x^r)}{d(\beta x^r)} &= \frac{1}{\beta x^r - 1} - \frac{\sigma^r}{1-\beta\sigma^r x^r} - \frac{1}{\beta x^r} \\
&\propto 1 - \sigma^r(\beta x^r)^2 \\
&> 0, \text{ iff } \sigma^r(\beta x^r)^2 < 1.
\end{aligned}$$

Thus if $\sigma^r(\beta x^r)^2 < 1$, R_k^r and R_b are increasing functions of only x^r :

$$R_k^r = R_k^r(x^r), R_k^{r\prime}(\cdot) > 0,$$

$$R_b = R_b(x^r), R_b'(\cdot) > 0.$$

Similarly

$$\begin{aligned}
x^w - R_b &= (R_k^w - R_b)\left(1 + \frac{b^w}{n^w}\right) + \left(R_k^w - R\right)\frac{d^w}{n^w} \\
&= (R_k^w - R_b)\left(1 + \frac{b^w}{n^w}\right) + \frac{1}{\omega}\left(R_k^w - R_b\right)\frac{d^w}{n^w} \\
&= \left(R_k^w - R_b\right)\left(\frac{Qk^w}{n^w} + \frac{1-\omega}{\omega}\frac{d^w}{n^w}\right) \\
&= \left(R_k^w - R_b\right)\left(\frac{1}{\omega\theta}\psi^w - \frac{1-\omega}{\omega}\right).
\end{aligned}$$

Thus

$$\begin{aligned}
R_k^w - R_b &= \omega\theta\frac{x^w - R_b}{\psi^w - \theta(1-\omega)}, \\
R_k^w - R &= \frac{1}{\psi^w - \theta(1-\omega)}\left[\omega\theta(x^w - R) + (\psi^w - \theta)(R_b - R)\right].
\end{aligned}$$

Because

$$\begin{aligned}
&\frac{d}{dx^w}\ \ln\left[\frac{\omega\theta(x^w - R)}{\psi^w - \theta(1-\omega)}\right] \\
&\propto \frac{1}{\beta x^w - 1} - \frac{\sigma^w}{1 - \sigma^w\beta x^w} - \frac{\Delta}{\Delta\beta x^w - \theta(1-\omega)}, \text{ where } \Delta = 1 - \sigma^w + \theta(1-\omega)\sigma^w \\
&\propto (1-\sigma^w)\left[1 - \sigma^w(\beta x^w)^2\right] - \theta(1-\omega)(1 - \sigma^w\beta x^w)^2,
\end{aligned}$$

R_k^w is an increasing function of x^w and x^r

$$R_k^w = R_k^w(x^w, x^r),$$

if

$$\begin{gathered}
(1-\sigma^w)\left[1 - \sigma^w(\beta x^w)^2\right] > \theta(1-\omega)(1 - \sigma^w\beta x^w)^2, \\
\sigma^r(\beta x^r)^2 < 1.
\end{gathered}$$

In the following we assume these conditions to be satisfied.

In the steady state, we know the rates of returns on capital for wholesale and retail banks and households are

$$R_k^w = \frac{Z+Q}{Q}$$

$$R_k^r = \frac{Z+Q}{Q+\alpha^r K^r}$$

$$R_k^h = \frac{Z+Q}{Q+\alpha^h K^h} = R.$$

Thus we have

$$Q = \frac{Z}{R_k^w - 1},$$

$$\alpha^r K^r = \frac{Z-(R_k^r-1)Q}{R_k^r} = Z\frac{R_k^w - R_k^r}{R_k^r(R_k^w-1)},$$

$$\alpha^h K^h = \frac{Z-(R-1)Q}{R} = Z\frac{R_k^w - R}{R(R_k^w-1)},$$

and Q, K^r and K^w are functions of (x^w, x^r).

Equilibrium C: $D^w = 0$

Here, the market clearing condition of capital is given by

$$\begin{aligned} QK^w &= \frac{Qk^w}{n^w}N^w \\ &= \frac{\psi^w(x^w) - \theta(1-\omega)}{\theta\omega}N^w(x^w) \\ &= Q(x^w,x^r)\left[\bar{K} - K^r(x^w,x^r) - K^h(x^w,x^r)\right] \end{aligned} \tag{B.1}$$

The market clearing condition of interbank credit is given by

$$\begin{aligned} B &= \left(\frac{Qk^w}{n^w} - 1\right)N^w \\ &= \frac{\psi^w(x^w) - \theta}{\theta\omega}N^w(x^w) \\ &= \frac{1}{\gamma}\left\{\frac{\psi^r(x^r)}{\theta}N^r(x^r) - [Q(x^w,x^r) + \alpha^r K^r(x^w,x^r)]\cdot K^r(x^w,x^r)\right\} \end{aligned} \tag{B.2}$$

The equilibrium value of $(x^w, x^r)$is given by (x^w, x^r) which satisfies (B.1 and B.2) simultaneously.

In order to verify $\mu_d^w > 0$ and $\mu_d^r > 0$, it is sufficient to check the inequalities

$$x^w > x^r > R = \beta^{-1}.$$

For the other inequality $\mu_b^w > \omega \mu_d^w$, it is sufficient to check

$$R_k^w - R_b > \omega\left(R_k^w - R\right),$$

or

$$(1-\omega)(R_k^w - R) > R_b - R.$$

This is equivalent with

$$(1-\omega)\frac{\beta x^w - 1}{\psi^w(x^w)} > \gamma \frac{\beta x^r - 1}{\psi^r(x^r)}. \tag{B.3}$$

Equilibrium D: $D^w > 0$

For this type of equilibrium, we need $\mu_{kb}^w = \omega \mu_d^w$, or

$$R_k^w - R_b = \omega\left(R_k^w - R\right).$$

Thus

$$\begin{aligned} x^w - R &= \left(R_k^w - R\right)\left(1 + \frac{d^w}{n^w} + \omega \frac{b^w}{n^w}\right) \\ &= \left(R_k^w - R\right)\frac{\psi^w}{\theta}, \end{aligned}$$

Thus being similar to the expression for $\beta(R_k^r - R)$, we get

$$\beta(R_k^w - R) = \theta \frac{\beta x^w - 1}{\psi^w(x^w)} = \theta \frac{(\beta x^w - 1)(1 - \sigma^w \beta x^w)}{(1-\sigma^w)\beta x^w} \equiv \varphi^w(\beta x^w).$$

R_k^w is an increasing function of x^w if $\sigma^w(\beta x^w)^2 < 1$.

Also we learn

$$R_b - R = (1-\omega)\left(R_k^w - R\right) = \gamma\left(R_k^r - R\right),$$

or

$$(1-\omega)\varphi^w(\beta x^w) = \gamma \varphi^r(\beta x^r), \tag{B.4}$$

and thus x^r is an increasing function of x^w. We can solve Q and K^h as functions of x^w as

$$\begin{aligned} Q &= \frac{Z}{R_k^w - 1} \\ &= \frac{\beta Z}{\varphi^w(\beta x^w) + 1 - \beta} \equiv Q(x^w), \end{aligned}$$

$$\begin{aligned} K^h &= \frac{1}{\alpha^h}[\beta Z - (1-\beta)Q] \\ &= \frac{1}{\alpha^h}\frac{\beta Z \varphi^w(\beta x^w)}{\varphi^w(\beta x^w) + 1 - \beta} \equiv K^h(x^w). \end{aligned}$$

We also get

$$\begin{aligned} K^r &= \frac{1}{\alpha^r}\frac{Z - (R_k^r - 1)Q}{R_k^r} = \frac{Z}{\alpha^r}\frac{R_k^w - R_k^r}{R_k^r(R_k^w - 1)} \\ &= \frac{1}{\alpha^r}\frac{\beta Z \varphi^w(\beta x^w)}{\varphi^w(\beta x^w) + 1 - \beta}\frac{\gamma + \omega - 1}{\gamma + (1-\omega)\varphi^w(\beta x^w)} \\ &= \frac{\gamma + \omega - 1}{\gamma + (1-\omega)\varphi^w(\beta x^w)}\frac{\alpha^h}{\alpha^r}K^h \equiv K^r(x^w) \end{aligned}$$

The capital market equilibrium is given by

$$\begin{aligned} QK^w &= \frac{1}{\theta\omega}\psi^w N^w - \frac{1-\omega}{\omega}(N^w + D^w) \\ &= \frac{1}{\theta\omega}\psi^w N^w - \frac{1-\omega}{\omega}(QK^w - B) \\ &= \frac{1}{\theta}\psi^w N^w + (1-\omega)B \\ &= \frac{1}{\theta}\psi^w N^w + \frac{1-\omega}{\gamma}\left[\frac{\psi^r}{\theta}N^r - (Q + \alpha^r K^r)K^r\right] \\ &= Q\left(\bar{K} - K^h - K^r\right). \end{aligned}$$

Thus

$$\begin{aligned} &\frac{\psi^w}{\theta}N^w + \frac{1-\omega}{\gamma}\frac{\psi^r}{\theta}N^r \\ &= \frac{\psi^w}{\theta}\left[N^w + \frac{\beta x^r - 1}{\beta x^w - 1}N^r\right], (\because (\text{B.4})) \\ &= Q\left[\bar{K} - K^h - K^r + \frac{1-\omega}{\gamma}\frac{Q + \alpha^r K^r}{Q}K^r\right] \\ &= Q\left[\bar{K} - K^h - K^r + \frac{1-\omega}{\gamma}\frac{R_k^w}{R_k^r}K^r\right] \\ &= Q\left[\bar{K} - K^h - \frac{\gamma + \omega - 1}{\gamma + (1-\omega)\varphi^w(\beta x^w)}K^r\right], \end{aligned}$$

or

$$\begin{aligned}&\frac{\psi^w(x^w)}{\theta}\left[N^w(x^w)+\frac{\beta x^r-1}{\beta x^w-1}N^r(x^r)\right]\\&=Q(x^w)\left[\overline{K}-K^h(x^w)-\frac{\gamma+\omega-1}{\gamma+(1-\omega)\varphi^w(\beta x^w)}K^r(x^w)\right].\end{aligned}\tag{B.5}$$

The equilibrium is given by (x^r, x^w) which satisfies (B.4 and B.5).

We need to check $D^w > 0$, or

$$0<\left(\frac{\psi^w}{\theta\omega}-\frac{1-\omega}{\omega}\right)N^w-\frac{1}{\theta}\psi^w N^w-\frac{1-\omega}{\gamma}\left[\frac{\psi^r}{\theta}N^r-(Q+\alpha^r K^r)K^r\right],$$

or

$$\gamma\left[\frac{\psi^w(x^w)}{\theta}-1\right]N^w(x^w)>\omega\left[\frac{\psi^r(x^r)}{\theta}N^r(x^r)-[Q(x^w)+\alpha^r K^r(x^w)]\cdot K^r(x^w)\right].$$

Equilibrium E: No Active Interbank Market

We have for $j = w, r$ that

$$\frac{(Q+f^j)k^j}{n^j}=\frac{\psi^j(x^j)}{\theta},$$

$$x^j-R=\left(R_k^j-R\right)\frac{(Q+f^j)k^j}{n^j}=\left(R_k^j-R\right)\frac{\psi^j(x^j)}{\theta},$$

or

$$R_k^j-R=\theta\frac{x^j-R}{\psi^j(x^j)},$$

or

$$R_k^j=R_k^j\left(x^j\right),\ R_k^{j\prime}(\cdot)>0$$

if $\sigma^w(\beta x^j)^2 < 1$. Thus

$$Q=Q(x^w),\ Q'(\cdot)<0$$

$$K^h=K^h(x^w), K^{h\prime}(\cdot)>0.$$

The aggregate capital of retail banks satisfies

$$\begin{aligned}QK^r&=Q\frac{Z-(R_k^r-1)Q}{\alpha^r R_k^r}=Q(x^w)\frac{Z}{\alpha^r}\frac{R_k^w(x^w)-R_k^r(x^r)}{R_k^r(x^r)[R_k^w(x^w)-1]}\\&=\frac{\psi^r(x^r)}{\theta}N^r(x^r)\end{aligned}\tag{B.6}$$

The capital market clearing condition is

$$\begin{aligned} QK^w &= \frac{\psi^w(x^w)}{\theta} N^w(x^w) \\ &= Q(x^w)\left[\bar{K} - K^r(x^r, x^w) - K^h(x^w)\right] \end{aligned} \tag{B.7}$$

The equilibrium is given by (x^r, x^w) which satisfies (B.6 and B.7).

Appendix C Anticipated Bank Run Case

Here, we describe the conditions determining agents policy functions in the case of anticipated runs. As in the text, we focus on the case in which variation in Z_{t+1} is negligible. Moreover, we follow the notation by which for any given variable $\widetilde{\xi}_t$

$$E_t^*\left(\widetilde{\xi}_{t+1}\right) = (1-p_t)\xi_{t+1} + p_t\xi_{t+1}^*$$

where ξ_{t+1}^* is the value taken by $\widetilde{\xi}_{t+1}$ when a run occurs.

Appendix C.1 Households

Households optimal choices of capital holdings and deposits are given by

$$E_t^*\left(\widetilde{\Lambda}_{t,t+1}\right)R_{t+1} = 1$$

$$E_t^*\left(\widetilde{\Lambda}_{t,t+1}\widetilde{R}_{kt+1}^h\right) = 1$$

Appendix C.2 Retail Bankers

The conditions in Lemma 2 that guarantee that retail banks are constrained are now modified as follows:

Lemma C.1 $b_t^r < 0$, $k_t^r > 0$ *and the incentive constraint is binding off*

$$0 < E_t^*\left[\widetilde{\Omega}_{t+1}^r\left(\widetilde{R}_{kt+1}^r - R_{t+1}\right)\right] = \frac{1}{\gamma}E_t^*\left[\widetilde{\Omega}_{t+1}^r\left(\widetilde{R}_{bt+1} - R_{t+1}\right)\right] < \theta.$$

The optimal choice of leverage is

$$\phi_t^r = \frac{E_t^*\left(\widetilde{\Omega}_{t+1}^r\right)R_{t+1}}{\theta - E_t^*\left[\widetilde{\Omega}_{t+1}^r\left(\widetilde{R}_{kt+1}^r - R_{t+1}\right)\right]}.$$

Appendix C.3 Wholesale Bankers

The optimization problem of wholesale banks when bank runs are anticipated is complicated by the fact that the banker can avoid bankruptcy by reducing its leverage in case a run materializes. Here, we derive conditions under which he does not wish to do this. For simplicity, we focus on the problem of a wholesale banker that only funds himself in the interbank market.

In this case we can derive a threshold level for leverage, ϕ_t^{wM}, under which the banker will survive a bank run, which is given by

$$\bar{R}_{bt+1} = R_{f,t+1} \equiv \frac{E_t^*\left(\widetilde{\Omega}_{t+1}^r \widetilde{R}_{\gamma,t+1}^r\right)}{E_t^*\left(\widetilde{\Omega}_{t+1}\right)} = R_{kt+1}^{w*} \frac{\phi_t^{wM}}{\phi_t^{wM}-1}$$

where

$$\widetilde{R}_{\gamma,t+1}^r \equiv \gamma \widetilde{R}_{kt+1}^r + (1-\gamma)R_{t+1}$$

and $R_{f,t+1}$ is the risk free interbank rate that satisfies Eq. (44) with $x_{t+1}^w = 1$.

The objective function of wholesale bankers displays a kink at ϕ_t^{wM}, so that in order to derive their optimal leverage choice we need to study separately the optimal choice in the region where leverage is high enough to induce bankruptcy when a run happens, $[\phi_t^{wM}, \infty)$, and in the region where bankruptcy is avoided even if a run happens, $\left[0, \phi_t^{wM}\right]$. As long as wholesale bankers objective is strictly increasing in leverage in both of these regions, the incentive constraint holds with equality.

In the bankruptcy region, $[\phi_t^{wM}, \infty)$, (45) with deterministic Z_{t+1} is simplified to

$$\bar{R}_{bt+1}(\phi_t^w) = R_{\gamma,t+1}^r + \frac{p_t}{1-p_t}\frac{\Omega_{t+1}^{r*}}{\Omega_{t+1}^r}\left(R_{\gamma,t+1}^{r*} - \frac{\phi_t^w}{\phi_t^w - 1}R_{t+1}^{w*}\right).$$

Then the objective function of a wholesale bank with one unit of networth is given by

$$\begin{aligned}
\psi^w\left(\phi_t^w\right) &= (1-p_t)\left\{\Omega_{t+1}^w\left[\phi_t^w\left(R_{kt+1}^w - \bar{R}_{bt+1}\left(\phi_t^w\right)\right) + \bar{R}_{bt+1}\left(\phi_t^w\right)\right]\right\} \\
&= (1-p_t)\Omega_{t+1}^w\left[\phi_t^w\left(R_{t+1}^w - R_{\gamma,t+1}^r\right) + R_{\gamma,t+1}^r\right] \\
&\quad + p_t\Omega_{t+1}^w\frac{\Omega_{t+1}^{r*}}{\Omega_{t+1}^r}\left[\phi_t^w\left(R_{k,t+1}^{w*} - R_{\gamma,t+1}^{r*}\right) + R_{\gamma,t+1}^{r*}\right]
\end{aligned}$$

which is strictly increasing in ϕ_t^w if and only if

$$(1-p_t)\left(R_{kt+1}^w - R_{\gamma,t+1}^r\right) + p_t\frac{\Omega_{t+1}^{r*}}{\Omega_{t+1}^r}\left(R_{kt+1}^{w*} - R_{\gamma,t+1}^{r*}\right) > 0 \qquad \text{(C.1)}$$

Notice that condition (C.1) is implied by the condition that guarantees that retail bankers are constrained, $E_t^*\left[\tilde{\Omega}_t^r\left(\tilde{R}_{kt+1}^r - R_{t+1}\right)\right] > 0$, together with the fact that retail bankers are less efficient at intermediating capital than wholesale bankers $\alpha^r > 0$:

$$
\begin{aligned}
&(1-p_t)\left(R_{kt+1}^w - R_{\gamma,t+1}^r\right) + p_t\frac{\Omega_{t+1}^{r*}}{\Omega_{t+1}^r}\left(R_{kt+1}^{w*} - R_{\gamma,t+1}^{r*}\right)\\
&> (1-p_t)\left(R_{kt+1}^r - R_{\gamma,t+1}^r\right) + p_t\frac{\Omega_{t+1}^{r*}}{\Omega_{t+1}^r}\left(R_{k,t+1}^{r*} - R_{\gamma,t+1}^{r*}\right)\\
&= \frac{(1-\gamma)}{\Omega_{t+1}^r}E_t^*\left\{\tilde{\Omega}_t^r\left(\tilde{R}_{kt+1}^r - R_{t+1}\right)\right\} > 0
\end{aligned}
$$

In the region where the banker is able to avoid bankruptcy even when a run happens, $\left[0, \phi_t^{wM}\right]$, the objective is instead

$$
\begin{aligned}
\psi^{w,n}\left(\phi_t^w\right) &= E_t^*\left\{\tilde{\Omega}_{t+1}^w\left[\phi_t^w\left(\tilde{R}_{kt+1}^w - R_{f,t+1}\right) + R_{f,t+1}\right]\right\}\\
&= \begin{array}{c} (1-p_t)\left\{\Omega_{t+1}^w\left[\phi_t^w\left(R_{kt+1}^w - R_{f,t+1}\right) + R_{f,t+1}\right]\right\}\\ +p_t\left\{\Omega_{t+1}^{w*}\left[\phi_t^w\left(R_{kt+1}^{w*} - R_{f,t+1}\right) + R_{f,t+1}\right]\right\} \end{array}
\end{aligned}
$$

and the condition that guarantees that the objective is strictly increasing in ϕ_t^w in this region is

$$
E_t^*\left[\tilde{\Omega}_{t+1}^w\left(\tilde{R}_{kt+1}^w - R_{f,t+1}\right)\right] > 0. \tag{C.2}
$$

Given this we can modify the conditions in Lemma 1 as follows:

Lemma C.2 *Under the conditions of Lemma C.1, the incentive constraint is binding off*

$$
0 < E_t^*\left[\tilde{\Omega}_{t+1}^w\left(\tilde{R}_{kt+1}^w - R_{f,t+1}\right)\right]
$$

$$
\theta\omega > (1-p_t)\left(R_{kt+1}^w - R_{\gamma,t+1}^r\right) + p_t\frac{\Omega_{t+1}^{r*}}{\Omega_{t+1}^r}\left(R_{kt+1}^{w*} - R_{\gamma,t+1}^{r*}\right).
$$

Appendix D Measurement

We use data from the Flow of Funds in order to construct empirical counterparts of the financial flows in the simplified intermediation process described in Fig. 1. The first step in constructing our time series is a definition of the wholesale and retail sector within the broad financial business sector.

Our classification is based on the sectors and instruments reported in the Flow of Funds. We use the liability structure of the different sectors included in the "Financial

Business" sector of the Flow of Funds in order to aggregate them into a Retail sector, a Wholesale sector, and Others. To do this, we proceed in two steps: we first classify the funding instruments in the Flow of Funds into four categories that we name Retail Funding, Wholesale Funding, Intermediated Assets, and Other Instruments; then we assign financial intermediaries to the Retail/Wholesale sector if the funding instruments they mostly rely on belong to the Retail/Wholesale category.

Table D.1 describes the four categories of funding we use. The labels in parentheses are the identifiers in the Flow of Funds.

The criterion we use to define the above categories is the composition of demand and supply for each instrument. Instruments that are supplied by financial intermediaries and demanded by households fall in the Retail category, while instruments that are mainly traded among financial intermediaries are included in Wholesale Funding. Intermediated Assets consist of all of the claims issued by domestic nonfinancial business and households. Others is a residual category.

To define our Retail and Wholesale sectors, we start by excluding some types of intermediaries from the ones that we are trying to study in our model economy. These are the intermediaries listed in the "Others" category in Table D.2. The remaining financial intermediaries appearing in the Flow of Funds are included in the Retail/Wholesale sector if they mostly rely on Retail/Wholesale funding. The resulting aggregation is described in Table D.2.

Table D.1 Classification of instruments in the flow of funds

Retail funding	Checkable deposits and currency (L.204) Time and saving deposits (L.205) Money market mutual fund shares (L.206) Mutual fund shares (L.214)	
Wholesale funding	*Short term*	Repurchase agreements (L.207) Security credit (L.224) Financial open market paper (L.208) Agency/GSE backed securities (L.210)
	Long term	Financial corporate bonds (L.212) Retail loans to wholesale (L.215)
Intermediated assets	Non-financial corporate bonds (L.212) Non-financial equity (L.213) Non-financial open market paper (L.208) Retail loans to non-financial (L.215) Mortgages (L.217) Consumer credit (L.222) Other loans (L.216)	
Other types of funding	All other instruments in the flow of funds	

Table D.2 Aggregation of financial sectors in the flow of funds

Retail sector	Private depository institutions (L.110) Money market mutual funds (L.121) Mutual funds (L.122)
Wholesale sector	Security brokers dealers (L.129) ABS issuers (L.126) GSE and GSE mortgage pools (L.124–125) Real estate investment trusts (L.128) Finance companies (L.127) Funding corporations (L.131) Holding companies (L.130)
Other intermediaries	Monetary authority (L.109) Private and public pension funds (L.117) Closed end and exchange traded funds (L.123) Insurance companies (L.115–116) Government (L.105–106) Rest of the world (L.132)
Households	L.101
Firms	L.102

Given this we construct the following measures:

1. K_t^h, K_t^r, K_t^w

   The intermediation shares are constructed by computing aggregate short and long positions of Households, Retail Banks, and Wholesale banks in the markets that make up the Intermediated Assets category in Table D.1. The matrix below describes each sectors' activity in each market. If sector J has a long/short position in market X the corresponding entry is given by X_+^J / X_-^J. If sector J has both long and short positions in market X, the corresponding entry also displays its net position, $X_{net}^J(+)/X_{net}^J(-)$.

Markets	**Bonds** *L*.**212**	**Equity** *L*.**213**	**Comm paper** *L*.**208**	**Loans** *L*.**215**	**Mortgages** *L*.**208**	**Consumer credit** *L*.**222**
Sectors Retail banks	BO_+^R BO_-^R $BO_{net}^R(+)$	EQ_+^R NA ?	CP_+^R CP_-^R $CP_{net}^R(+)$	L_+^R	M_+^R	CC_+^R
Wholesale banks	BO_+^W BO_-^W $BO_{net}^W(-)$	EQ_+^W NA ?	CP_+^W CP_-^W $CP_{net}^W(-)$	L_-^W	M_+^W	CC_+^W
Other item	BO_+^O BO_-^O $BO_{net}^O(+)$	EQ_+^O NA ?	CP_+^O CP_-^O $CP_{net}^O(+)$	L_-^O	M_+^O	CC_+^O

Markets	Bonds *L*.212	Equity *L*.213	Comm paper *L*.208	Loans *L*.215	Mortgages *L*.208	Consumer credit *L*.222
Households	BO^H_+	EQ^H_+	0 0	L^H_-	M^H_-	CC^H_-
Firms	BO^F_-	EQ^F_-	CP^F_+ CP^F_- $\mathrm{CP}^F_{net}(-)$	L^F_-	M^F_+ M^F_- $\mathrm{M}^F_{net}(-)$	CC^F_+

We make several assumptions in order to conduct our measures.First, in the markets for bonds and commercial paper, some positions are potentially inconsistent with our intermediation model. This is because some sectors within the retail category are short in these markets and some in wholesale are long, $\mathrm{BO}^R_- > 0$, $\mathrm{CP}^R_- > 0$, $\mathrm{BO}^W_+ > 0$ and $\mathrm{CP}^W_+ > 0$. This allows for the possibility that retail banks were borrowing from wholesale in these markets. However, we rule out this possibility in constructing our measures for two reasons: given the heavy reliance on these types of instruments in financial transactions among industries within the respective categories and among financial firms within the same industry, it is reasonable to assume that the vast majority of these offsetting positions were actually arising from cross holdings among firms within the same category; moreover, the actual size of BO^R_- and CP^R_- with respect to BO^W_- and CP^W_- was very small, ie, $\frac{CP^R_-}{CP^W_-} \simeq 0.1\%$ and $\frac{CP^R_-}{CP^W_-} \simeq 3\%$ in 2007. This implies that we can safely work with the net positions for wholesalers and retailers. Given the assumptions we make in these markets we can construct model consistent measures from bonds and commercial paper data by assuming that households lend to nonfinancial firms, which is part of K^h, while retail banks (and Other intermediaries) lend to both Wholesale banks, which is part of B, and firms, which is part of K^r.[ax] We also assume that portfolio weights on nonfinancial and financial issued instruments in these markets are the same for retail banks and other intermediaries.[ay] That is, letting $F^{i,F}_{bo}$ and $F^{i,F}_{cp}$ be the proportions of lender's $i's$ holdings of bonds and commercial paper that are issued by nonfinancial firms, we have

[ax] The Households' sector in the Flow of Funds is a residual category that includes Hedge Funds, private equity funds and personal trusts, which are intermediaries that our model does not directly capture. In any case, households' intermediation in bonds and commercial paper market is a small component of household intermediation so that very little would change if we instead made different assumptions about households positions in these markets.

[ay] We include long positions of nonfinancial firms in the commercial paper within intermediation performed by "Others."

$$F^{H,F}_{bo} = 1; F^{R,F}_{bo} = \left(\frac{BO^F_- - BO^H_+}{BO^F_- + BO^W_{net} - BO^H_+}\right);$$[az]

Similarly for commercial paper: $F^{H,F}_{cp} = 0; F^{R,F}_{cp} = \frac{CP^F_-}{CP^F_- + CP^W_{net}}$ Second, for corporate equities the Flow of Funds does not report a disaggregated measure of equity issued by individual industries or the type of equity held by the various industries. Since we use this market only in measuring K^i, we simply assume that each sector holds a scaled version of the same equity portfolio consisting of the three sectors for which we have issuance data: Foreign equities, Financial Business equities, and Non-Financial Business Equities, denoted by EQ^{ROW}, EQ^{FIN}, and EQ^{NFI}, respectively. That is, in order to compute how many funds flow to nonfinancial firms from each other sector we simply scale their total equity holdings by

$$\eta = \frac{EQ^{NFI}}{EQ^{NFI} + EQ^{FIN} + EQ^{ROW}}$$

Given this we can compute

$$K^h_t = \eta EQ^H + BO^H_+$$

$$K^r_t = \eta EQ^R + F^{R,F}_{bo} BO^R_{net} + F^{R,F}_{cp} CP^R_{net}$$
$$+ L^F_- + L^H_- + M^R_+ + CC^R_+$$

$$K^W_t = \eta EQ^W + M^W_+ + CC^W_+$$

2. *B*,*D*

   B is simply computed as wholesale net borrowing in all of the short-term wholesale instruments: Repo, Commercial Paper, Agency Debt, and Security credit. *D* is given by Households and nonfinancial Business holdings of retail funding instruments.

3. Leverage multiple for broker dealers, finance companies, and GSE

   We compute financial leverage multiple for these three sectors by dividing total financial assets by financial assets minus financial liabilities plus equity investment by holding companies. We do not have a measure of nonfinancial assets in the Flow of Funds so the leverage multiple reported here overstates financial leverage multiple that would include nonfinancial assets in the computation. We compute average leverage multiple by using time varying weights corresponding to the relative sizes of these three sectors as measured by total financial assets.

[az] Notice that we attribute all household's lending in this market, BO^H_+, to "nonfinancial loans" K^h; we then allocate retail bankers supply of funds in this market to nonfinancial loans, K^r proportionally to the weight of nonfinancial firms demand for funds that is not met by households, $BO^F_- - BO^H_+$, in the total demand for funds that is not met by households, $BO^F_- + BO^W_{net} - BO^H_+$

Appendix E Computation

It is convenient for computations to introduce the ex-ante optimal values of surviving bankers at time t in the two sectors:

$$\begin{aligned}\bar{V}_t^w &= \left[1-\sigma+\sigma\theta\left(1-\omega+\omega\phi_t^w\right)\right]\frac{N_t^w - W^w}{\sigma^w}\\ &= \Omega_t^w \frac{N_t^w - W^w}{\sigma^w}\end{aligned} \tag{E.1}$$

$$\begin{aligned}\bar{V}_t^r &= \left[1-\sigma+\sigma\theta\phi_t^r\right]\frac{N_t^r - W^r}{\sigma^r}\\ &= \Omega_t^r \frac{N_t^r - W^r}{\sigma^r}\end{aligned} \tag{E.2}$$

Let the state of the economy if a run has not happened be denoted by $x=(N^w, N^r, Z)$, and the state in case a run has happened be denoted by $x^* = (0, N^r, Z)$. We use time iteration in order to approximate the functions

$$\left\{\mathbf{Q}(x), \mathbf{C}^h(x), \bar{\mathbf{V}}^r(x), \bar{\mathbf{V}}^w(x), \Gamma(x)\right\} \quad x\in[W^w, \bar{N}^w]\times[W^r, \bar{N}^r]\times[(0.95)Z, Z]$$

and

$$\left\{\mathbf{Q}^*(x), \mathbf{C}^{h*}(x^*), \bar{\mathbf{V}}^{r*}(x^*), \Gamma^*(x^*)\right\} \quad x^*\in\{0\}\times[W^r, \bar{N}^r]\times[(0.95)Z, Z]$$

where $\Gamma(x)$ and $\Gamma^*(x^*)$ are the laws determining the stochastic evolution of the state (see later).

The computational algorithm proceeds as follows:

1. Determine a functional space to use for approximating equilibrium functions. (We use piecewise linear).
2. Fix a grid of values for the state in case no run happens $G\subset[W^w, \bar{N}^w]\times[W^r, \bar{N}^r]\times[0.95, 1]$ and for the state in case a run happens $G^*\subset\{0\}\times[W^r, \bar{N}^r]\times[0.95, 1]$.
3. Set $j=0$ and guess initial values for

$$NRPol_{t,j} = \left\{Q_{t,j}(x),\, C_{t,j}^h(x),\, \bar{V}_{t,j}^r(x),\, \bar{V}_{t,j}^w(x), \Gamma_{t,j}(x)\right\}_{x\in G}$$

and

$$RPol_{t,j} = \left\{Q_{t,j}^*(x),\, C_{t,j}^{h*}(x^*),\, \bar{V}_{t,j}^{r*}(x^*), \Gamma_{t,j}^*(x^*)\right\}_{x^*\in G^*}.$$

The guess for $\Gamma_{t,j}(x)$ involves guessing $\left\{p_{t,j}(x), N_{t,j}^{r\prime}(x), N_{t,j}^{w\prime}(x), N_{t,j}^{r\prime *}(x), Z'(x)\right\}$ which implies

$$\Gamma_{t,j}(x) = \begin{cases}\left(N_{t,j}^{w\prime}(x), N_{t,j}^{r\prime}(x), Z'(Z)\right) & w.p.\ 1-p_{t,j}(x)\\ \left(0, N_{t,j}^{r\prime *}(x), Z'(Z)\right) & w.p.\ p_{t,j}(x)\end{cases}.$$

We denote by $x_{t,j}^{\prime NR}(x)=\left(N_{t,j}^{w\prime}(x),\, N_{t,j}^{r\prime}(x), Z^{\prime}(Z)\right)$ the state evolution if there is no run in the following period and $x_{t,j}^{\prime R}(x)=\left(0,\, N_{t,j}^{r\prime *}(x), Z^{\prime}(Z)\right)$ the evolution if a run happens in the following period.

Similarly the guess for $\Gamma_{t,j}^{*}(x^{*})$ involves guessing $\left\{\hat{N}_{t,j}^{r\prime}(x^{*}), Z^{\prime}(Z)\right\}$ which implies

$$\Gamma_{t,j}^{*}(x^{*})=\left((1+\sigma^{w})W^{w},\, \hat{N}_{t,j}^{r\prime}(x^{*}), Z^{\prime}(Z)\right)$$

4. Assume that $NRPol_{t,j}$ and $RPol_{t,j}$ have been found for $j\leq i<M$ where M is set to 10,000. To find $NRPol_{t,i+1}$ and $RPol_{t,i+1}$, first use $NRPol_{t,i}$ and $RPol_{t,i}$ to find functions in the approximating space that take on these values on the grid, eg, $\mathbf{Q}_{i}\text{:}\,[W^{w},\bar{N}^{w}]\times[W^{r},\bar{N}^{r}]\times[0.95,1]\rightarrow\mathbf{R}$ is the price function that satisfies $\mathbf{Q}_{i}(x)=Q_{t,i}(x)$ for each $x\in G$.
5. Derive $NRPol_{t,i+1}$ and $RPol_{t,i+1}$ by assuming that from time $t+1$ onwards equilibrium outcomes are determined according to the functions associated to $NRPol_{t,i}$ and $RPol_{t,i}$ found in step 4:
   - NO RUN SYSTEM

   At any point $x_{t}=\left(N_{t}^{w},N_{t}^{r},Z_{t}\right)\in G$ the system determining $\left\{\phi_{t}^{w},\phi_{t}^{r},B_{t},Q_{t},C_{t}^{h},K_{t}^{h},K_{t}^{r}\right\}$ is given by

$$\theta\left[1-\omega+\omega\phi_{t}^{w}\right]N_{t}^{w}=\beta(1-\mathbf{p}_{i}(x_{t}))\bar{\mathbf{V}}_{i}^{w}\left(\mathbf{x}_{i}^{\prime NR}(x_{t})\right)$$

$$\left(\phi_{t}^{w}-1\right)N_{t}^{w}=B_{t}$$

$$\phi_{t}^{w}N_{t}^{w}=Q_{t}\left(1-K_{t}^{r}-K_{t}^{h}\right)$$

$$\theta\phi_{t}^{r}N_{t}^{r}=\beta\left[(1-\mathbf{p}_{i}(x_{t}))\bar{\mathbf{V}}_{i}^{r}\left(\mathbf{x}_{i}^{\prime NR}(x)\right)+\mathbf{p}_{i}(x_{t})\bar{\mathbf{V}}_{i}^{r*}\left(\mathbf{x}_{i}^{\prime R}(x)\right)\right]$$

$$\phi_{t}^{r}N_{t}^{r}=\left(Q_{t}+\alpha^{r}K_{t}^{r}\right)K_{t}^{r}+(1-\gamma)B_{t}$$

$$\beta E_{i}\left\{\frac{C_{t}^{h}}{\widetilde{\mathbf{C}}_{i}^{h}(\Gamma_{i}(x))}\left(\mathbf{Z}^{\prime}(Z_{t})+\widetilde{\mathbf{Q}}_{i}(\Gamma_{i}(x))\right)\right\}=Q_{t}+\alpha^{h}K_{t}^{h}$$

$$C_{t}^{h}+\frac{(1-\sigma_{w})\left(N_{t}^{r}-W^{w}\right)}{\sigma_{w}}+\frac{(1-\sigma_{r})\left(N_{t}^{r}-W^{r}\right)}{\sigma_{r}}+\frac{\alpha^{h}\left(K_{t}^{h}\right)^{2}}{2}+\frac{\alpha^{r}\left(K_{t}^{r}\right)^{2}}{2}= \\ Z_{t}\left(1+W^{h}\right)+W^{r}+W^{w}=$$

where E_{i} is the expectation operator associated with the stochastic realization of a run according to $\mathbf{p}_{i}$ and tildes denote random variables whose values depend on the realization of the sunspot. For instance,

$$\widetilde{\mathbf{C}}_i^h(\Gamma_i(x)) = \begin{cases} \mathbf{C}_i^h\left(\mathbf{N}_i^{w\prime}(x), \mathbf{N}_i^{r\prime}(x), \mathbf{Z}'(Z)\right) & w.p.\ 1 - \mathbf{p}_i(x) \\ \mathbf{C}_i^{h*}\left(\mathbf{N}_i^{r\prime *}(x), \mathbf{Z}'(Z)\right) & w.p.\ \mathbf{p}_i(x) \end{cases}$$

One can then find $\left\{R_t, \bar{R}_t^b\right\}$ from

$$R_t = \frac{1}{\beta E_i \left\{ \dfrac{C_t^h}{\widetilde{\mathbf{C}}_i^h(\Gamma_i(x))} \right\}}$$

$$\bar{R}_t^b = \frac{E_i \left\{ \widetilde{\Omega}^r(\Gamma_i(x)) \left(\gamma \dfrac{\left(\mathbf{Z}'(Z_t) + \widetilde{\mathbf{Q}}_i(\Gamma_i(x))\right)}{Q_t + \alpha^r K_t^r} + (1-\gamma) R_t \right) \right\}}{(1 - \mathbf{p}_i(x_t)) \boldsymbol{\Omega}^r(\mathbf{x}_i^{\prime NR}(x_t))} - \frac{-\mathbf{p}_i \boldsymbol{\Omega}^{r*}\left(x_i^{\prime R}(x_t)\right) \left(\dfrac{\left(\mathbf{Z}'(Z_t) + \widetilde{\mathbf{Q}}_i(\Gamma_i(x))\right)}{Q_t} \dfrac{\phi_t^w}{\phi_t^w - 1} \right)}{(1 - \mathbf{p}_i(x_t)) \boldsymbol{\Omega}^r(\mathbf{x}_i^{\prime NR}(x_t))}$$

where

$$\widetilde{\Omega}^r(\Gamma_i(x)) = \begin{cases} \sigma^r \dfrac{\bar{\mathbf{V}}_i^r\left(\mathbf{N}_i^{w\prime}(x), \mathbf{N}_i^{r\prime}(x), \mathbf{Z}'(Z)\right)}{\mathbf{N}_i^{w\prime}(x) - W} & w.p.\ 1 - \mathbf{p}_i(x) \\ \sigma^r \dfrac{\bar{\mathbf{V}}_i^{r*}\left(\mathbf{N}_i^{r\prime *}(x), \mathbf{Z}'(Z)\right)}{\mathbf{N}_i^{w\prime}(x) - W} & w.p.\ \mathbf{p}_i(x) \end{cases}$$

and finally $\left\{\bar{V}_t^r, \bar{V}_t^w, t\right\}$ are given by

$$\bar{V}_t^w = \left[1 - \sigma + \sigma\theta\left(1 - \omega + \omega\phi_t^w\right)\right] \frac{N_t^w - W^w}{\sigma^w}$$

$$\bar{V}_t^r = \left[1 - \sigma + \sigma\theta\phi_t^r\right] \frac{N_t^r - W^r}{\sigma^r}$$

$$\Gamma_t = \begin{cases} \left(N_{t+1}^w, N_{t+1}^r, Z'(Z)\right) & w.p.\ 1 - p_t \\ \left(0, N_{t+1}^{r*}, Z'(Z)\right) & w.p.\ p_t \end{cases}$$

where

$$N_{t+1}^w = \sigma^w N_t^w \left[\phi_t^w \left(\frac{\mathbf{Z}'(Z_t) + \mathbf{Q}_i\left(\mathbf{x}_i^{\prime NR}(x)\right)}{Q_t} - \bar{R}_t^b \right) + -R_t^b \right] + W^w$$

$$N_{t+1}^r = \sigma^r \left(\left[\mathbf{Z}'(Z_t) + \mathbf{Q}_i\left(\mathbf{x}_i^{\prime NR}(x)\right)\right] K_t^r + B_t \bar{R}_t^b - D_t R_t \right) + W^w$$

$$N_{t+1}^{r*} = \sigma^r\left(\left[Z'(Z_t) + \mathbf{Q}_i^*\left(\mathbf{x}_i'^R(x)\right)\right]\left(K_t^r + K_t^w\right) - D_t R_t\right) + W^w$$

$$p_t = \left[1 - \frac{\dfrac{Z'(Z_t) + \mathbf{Q}_i^*\left(\mathbf{x}_i'^R(x)\right)}{Q_t}}{\bar{R}_{bt}} \cdot \frac{\phi_t^w}{\phi_t^w - 1}\right]^\delta$$

- RUN SYSTEM

 Analogously at a point $x_t^* = \left(0, N_t^r, Z_t\right) \in G^*$ the system determining $\left\{\phi_t^{r*}, Q_t^*, C_t^{h*}, K_t^{h*}\right\}$ is given by

$$\theta\phi_t^{r*} N_t^r = \beta - \mathbf{V}_i^r\left(\Gamma_i^*\left(x_t^*\right)\right)$$

$$\phi_t^{r*} N_t^r = \left(Q_t^* + \alpha^r K_t^{r*}\right) K_t^{r*}$$

$$\beta\left\{\frac{C_t^{h*}}{\mathbf{C}_i^h\left(\Gamma_i^*\left(x_t^*\right)\right)}\left(Z'(Z_t) + \mathbf{Q}_i\left(\Gamma_i^*\left(x_t^*\right)\right)\right)\right\} = Q_t^* + \alpha^h K_t^{h*}$$

$$C_t^{h*} + \frac{(1-\sigma_r)}{\sigma_r}\left(N_t^r - W^r\right) + \frac{\alpha^h}{2}\left(K_t^{h*}\right)^2 + \frac{\alpha^r}{2}\left(1 - K_t^{h*}\right)^2 = Z_t\left(1 + W^h\right) + W^r$$

and $\left\{R_t^*, \bar{V}_t^{r*}, \Gamma_t^*\right\}$ are given by

$$R_t^* = \frac{1}{\beta E_i\left\{\dfrac{C_t^{h*}}{\mathbf{C}_i^h\left(\Gamma_i^*\left(x_t^*\right)\right)}\right\}}$$

$$\bar{V}_t^{r*} = \left[1 - \sigma + \sigma\theta\phi_t^{r*}\right]\frac{N_t^r - W^r}{\sigma^r}$$

$$\Gamma_i^*(x^*) = \left((1 + \sigma^w)W^w, \hat{N}_{t+1}^r, Z'(Z)\right)$$

$$\hat{N}_{t+1}^r = \sigma^r N_r^r\left[\phi_t^{r*}\left(\frac{Z'(Z_t) + \mathbf{Q}_i\left(\Gamma_i^*\left(x_t^*\right)\right)}{Q_t} - R_t^*\right) + R_t^*\right] + W^r$$

6. Compute the maximum distance between $NRPol_t = \left\{Q_t, \bar{V}_t^r, \bar{V}_t^w, C_t^h, p_t, N_{t+1}^r, N_{t+1}^w, N_{t+1}^{r*}\right\}$ and $NRPol_{t,i}$

$$dNR = \max_{x_t \in G} \max\left|NRPol_t - NRPol_{t,i}\right|$$

and similarly for $RPol_t = \left\{Q_t^*, -V_t^{r*}, C_t^{h*}, \hat{N}_{t+1}^r\right\}$ and $RPol_{t,i}$

$$dR = \max_{x_t \in G^*} \max\left|RPol_t - RPol_{t,i}\right|$$

if dNR and dR are small enough, in our case $e - 6$, set

$$NRPol_{t,i+1} = NRPol_{t,i}$$
$$RPol_{t,i+1} = RPol_{t,i}$$

Otherwise set

$$NRPol_{t,i+1} = \alpha NRPol_{t,i} + (1-\alpha)NRPol_t$$
$$RPol_{t,i+1} = \alpha RPol_{t,i} + (1-\alpha)RPol_t$$

where $\alpha \in (0,1)$.

REFERENCES

Adrian, T., Ashcraft, A., 2012. Shadow banking: a review of the literature. In: The New Palgrave Dictionary of Economics, 2012 Version, second ed. [internet]. Palgrave Macmillan, Basingstoke.

Adrian, T., Colla, P., Shin, H., 2012. Which financial frictions? Paring the evidence from financial crisis of 2007-9. In: Acemoglu, D., Parker, J., Woodford, M. (Eds.), NBER Macroeconomic Annual 2012, vol. 27, May 2013, pp. 159–214.

Allen, F., Gale, D., 2007. Understanding Financial Crises. Oxford University Press, Oxford.

Angeloni, I., Faia, E., 2013. Capital regulation and monetary policy with fragile banks. J. Monet. Policy 60, 3111–3382.

Begenau, J., 2015. Capital requirements, risk choice, and liquidity provision in a business cycle model. Harvard Business School Working Paper, no. 15-072.

Bernanke, B., 2010. Causes of the recent financial and economic crisis. Statement before the Financial Crisis Inquiry Commission, Washington, September 2.

Bernanke, B., Gertler, M., 1989. Agency costs, net worth and business fluctuations. Am. Econ. Rev. 79, 14–31.

Bianchi, J., 2011. Overborrowing and systemic externalities in the business cycle. Am. Econ. Rev. 101, 3400–3426.

Bianchi, J., Mendoza, E., 2013. Optimal time-consistent macroprudential policy. NBER Working Paper 19704.

Bigio, S., 2015. Financial risk capacity. Working Paper.

Bocola, L., 2016. The Pass-Through of Sovereign Risk. J. Polit. Econ. forthcoming.

Boissay, F., Collard, F., Smets, F., 2013. Booms and systemic banking crises. Mimeo.

Brunnermeier, M.K., Oemke, M., 2013. Maturity rat race. J. Finance 68, 483–521.

Brunnermeier, M.K., Pedersen, L., 2009. Market liquidity and funding liquidity. Rev. Financ. Stud. 22, 2201–2238.

Brunnermeier, M.K., Sannikov, Y., 2014. A macroeconomic model with a financial sector. Am. Econ. Rev. 104, 379–421.

Caballero, R., Farhi, E., 2015. The safety trap. Working Paper.

Chari, V., Kehoe, P., 2015. Bailouts, time inconsistency, and optimal regulation: a macroeconomic view. Federal Reserve Bank of Minneapolis, Research Department Staff Report 481.

Christiano, L., Ikeda, D., 2014. Leverage restrictions in a business cycle model. In: Macroeconomic and Financial Stability: Challenges for Monetary Policy.

Cole, H., Kehoe, T., 2000. Self-fulfilling debt crises. Rev. Econ. Stud. 67, 91–161.

Cooper, R., Ross, T., 1998. Bank runs: liquidity costs and investment distortions. J. Monet. Econ. 41, 27–38.

Covitz, D., Liang, N., Suarez, G., 2013. Evolution of a financial crisis: collapse of the asset-backed commercial paper market. J. Finance 68, 815.

Curdia, V., Woodford, M., 2010. Credit spreads and monetary policy. J. Money Credit Bank. 42 (6), 3–35.

Dang, T., Gorton, G., Holmstrom, B., 2012. Ignorance, debt and financial crises.

Diamond, D., Dybvig, P., 1983. Bank runs, deposit insurance, and liquidity. J. Polit. Econ. 91, 401–419.

Di Tella, S., 2014. Uncertainty shocks and balance sheet recessions. Working Paper.

Eggertsson, G., Krugman, P., 2012. Debt, Deleveraging, and Liquidity Trap: a Fisher-Minsky-Koo Approach, Q. J. Econ. 127 (3), 1469–1513.

Ennis, H., Keister, T., 2003. Economic growth, liquidity, and bank runs. J. Econ. Theory 109, 220–245.

Farhi, E., Tirole, J., 2012. Collective moral hazard, maturity mismatch and systemic bailouts. Am. Econ. Rev. 102 (1), 60–93.

Farhi, E., Werning, I., 2015. A theory of macroprudential policies in the presence of nominal rigidities. Working Paper.

Farmer, R., 1999. The Macroeconomics of Self-Fulfilling Prophecies. MIT Press.

Ferrante, F., 2015a. A model of endogenous loan quality and the collapse of the shadow banking system. Finance and Economics Discussion Series 2015-021, Federal Reserve Board.

Ferrante, F., 2015b. Risky mortgages, bank leverage and credit policy. Working Paper.

Garleanu, N., Panageas, S., Yu, J., 2015. Financial entanglement: a theory of incomplete integration, leverage, crashes and contagion. Am. Econ. Rev. 105 (7), 1979–2010.

Geanakoplos, J., Polemarchakis, H., 1986. Existence, regularity, and constrained suboptimality of competitive allocations when the asset market is incomplete. In: Uncertainty, Information, and Communication: Essays in Honor of K. J. Arrow, III. Cambridge University Press, Cambridge.

Gertler, M., Karadi, P., 2011. A model of unconventional monetary policy. J. Monet. Econ. 58 (1), 17–34.

Gertler, M., Kiyotaki, N., 2011. Financial Intermediation and Credit Policy in Business Cycle Analysis. In: Friedman, B.M., Woodford, M. (Eds.), Handbook of Monetary Economics, vol. 3A. Elsevier Science, Amsterdam, pp. 547–599.

Gertler, M., Kiyotaki, N., 2015. Banking, liquidity and bank runs in an infinite horizon economy. Am. Econ. Rev. 105 (7), 2011–2043

Gertler, M., Kiyotaki, N., Prestipino, A., 2016. Anticiapted Banking Panics. Am. Econ. Rev. Pap. Proc. 106 (5), 554–559.

Gertler, M., Kiyotaki, N., Queralto, A., 2012. Financial crises, bank risk exposure and government financial policy. J. Monet. Econ. 59, S17–S34.

Gilchrist, S., Zakrajsek, E., 2012. Credit spread and business cycle fluctuations. Am. Econ. Rev. 102, 1692–1720.

Giroud, X., Mueller, H., 2015. Firm leverage and unemployment during the great recession. Mimeo.

Goldstein, I., Pauzner, A., 2005. Demand-deposit contracts and the probability of bank runs. J. Finance 60, 1293–1327.

Gorton, G., 2009. Information, liquidity and the (ongoing) panic of 2007. Am. Econ. Rev. Pap. Proc. 99 (2), 567–572.

Gorton, G., Metrick, A., 2012. Who ran on repo? NBER Working Paper 18455.

Gorton, G., Metrick, A., 2015. The safe asset share. Am. Econ. Rev. Pap. Proc. 102 (3), 101–106.

Guerrieri, V., Lorenzoni, G., 2011. Credit crises, precautionary savings and the liquidity trap. NBER Working Paper 17583.

Gurley, J., Shaw, E., 1960. Money in Theory of Finance. Brookings Institution, Washington, DC.

He, Z., Krishnamurthy, A., 2013. Intermediary asset pricing. Am. Econ. Rev. 103 (2), 732–770.

He, Z., Krishnamurthy, A., 2014. A macroeconomic framework for quantifying systemic risk. University of Chicago and Stanford University, Working Paper.

Holmstrom, B., Tirole, J., 1997. Financial intermediation, loanable funds and the real sector. Q. J. Econ. 112 (3), 663–691.

Holmstrom, B., Tirole, J., 2011. Inside and Outside Liquidity. MIT Press, Cambridge, MA.

Iacoviello, M., 2005. House prices, borrowing constraints and monetary policy in the business cycle. Am. Econ. Rev. 95 (3), 739–764.

Kacperczyk, M., Schnabl, P., 2010. When safe proved risky: commercial paper during the financial crisis of 2007-2009. J. Econ. Perspect. 24 (1), 29–50.

Kiyotaki, N., Moore, J., 1997. Credit cycles. J. Polit. Econ. 105, 211–248.

Korinek, A., Simsek, A., 2015. Liquidity trap and excessive leverage. Working Paper.

Krishnamurthy, A., 2003. Collateral constraints and the amplification mechanism. J. Econo. Theory 111 (2), 277–292.

Krishnamurthy, A., Nagel, S., Orlov, D., 2014. Seizing up repo. J. Finance 69 (6), 2381–2417.

Lorenzoni, G., 2008. Inefficient credit boom. Rev. Econ. Stud. 75, 809–833.

Martin, A., Skeie, D., Thadden, E.V., 2014. Fragility of short-term secured funding. J. Econ. Theory 149, 15–42.

Martin, A., Skeie, D., Thadden, E.V., 2014. Repo runs. Rev. Financ. Stud. 27, 957–989.

McCabe, P., 2010. The cross section of money market fund risks and financial crises. Finance and Economics Discussion Series 2010-51, Federal Reserve Board.

Mendoza, E., 2010. Sudden stops, financial crises, and leverage. Am. Econ. Rev. 100, 1941–1966.

Midrigan, T., Philippon, T., 2011. A macroeconomic framework for quantifying systemic risk. NBER Working Paper 19885.

Morris, S., Shin, H., 1998. Unique equilibrium in a model of self-fulfilling currency attacks. Am. Econ. Rev. 88, 587–597.

Parlatore, C., 2015. Fragility in money market funds: sponsor support and regulation. Working Paper.

Philippon, T., 2015. Has the US Finance Industry Become Less Efficient? On the Theory and Measurement of Financial Intermediation. Am. Econ. Rev. 105 (4), 1408–1438.

Pozsar, Z., Adrian, T., Ashcraft, A., Boesky, H., 2013. Shadow banking. Fed. Reserv Bank. New York Econ. Policy Rev 19 (2), 1–16.

Robatto, R., 2014. Financial crises and systematic bank runs in a dynamic model of banking.

Uhlig, H., 2010. A model of a systemic bank run. J. Monet. Econ. 57, 78–96.

CHAPTER 17

Housing and Credit Markets: Booms and Busts

V. Guerrieri[*,†], H. Uhlig[*,†,‡]
[*]University of Chicago, Chicago, IL, United States
[†]NBER, Cambridge, MA, United States
[‡]CEPR, London, United Kingdom

Contents

Abstract

Prompted by the recent US experience, in this chapter, we study the interaction between cycles in credit markets and cycles in housing markets. There is a large growing literature exploring two different approaches: on the one hand, a boom–bust in house prices can generate a boom–bust in credit market and, on the other hand, a boom–bust in credit markets can generate a boom–bust in house prices. We start by presenting a stark mechanical model to formalize the interaction between housing prices

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.06.001

and credit markets and explore these two channels in a mechanical way. Next, we present two simple models that highlight the two approaches. First, we propose a catastrophe model, where an increase in credit availability can generate first a boom and then a bust in mortgage markets because of multiple equilibria due to adverse selection: as lending expands, the composition of borrowers worsens and at some point this can generate a crash in credit market. Second, we propose a sentiment model, where house prices increase above fundamentals because investors buy assets under the irrational belief that there is always going to be an ever more foolish buyer, willing to buy at a higher price. In the course of the chapter, we relate our simple models to the large existing literature on these topics. At the end, we also point to some empirical papers that propose related facts.

Keywords

Housing prices, Credit markets, Cycles, Leverage, Adverse selection, Bubbles, Sentiments

JEL Classification Codes:

D82, D84, E44, G21

1. INTRODUCTION

In the recent years, the United States has experienced, at the same time, a boom–bust episode in house price and a boom–bust episode in credit markets, as reflected in Figs. 1 and 2.

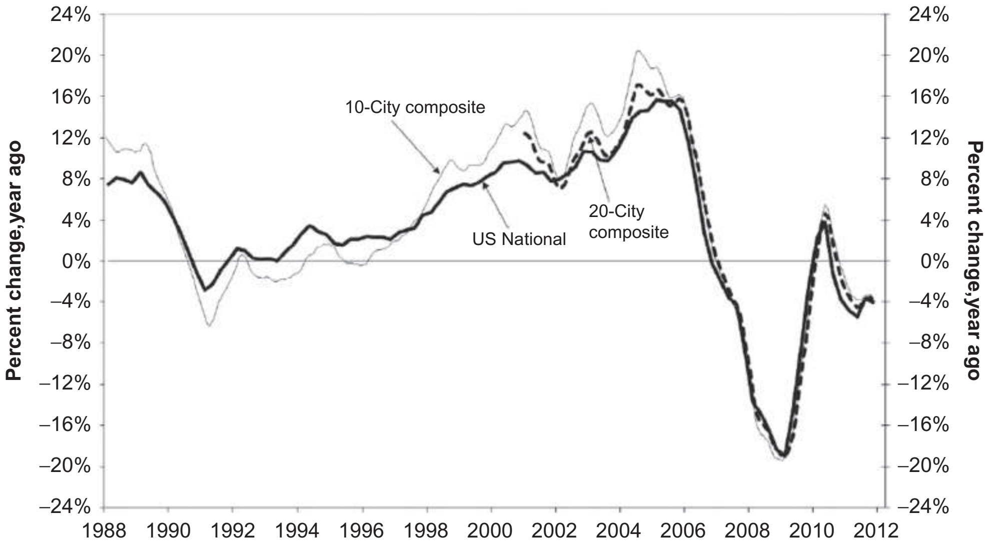
The purpose of this chapter is to explore the connection between financial markets and the housing market and its effects on the macroeconomic activity. There is a large and growing literature that separately explores credit cycles and house price bubbles and busts. In this chapter, we will try to connect these two streams of literature and understand the potential feedbacks between the two.

In particular, we will explore two different broad approaches to think about this connection:

1. the house price boom–bust generates the credit boom–bust;
2. the credit boom–bust generates the house price boom–bust.

Moreover, we embrace the view that, in both cases, these connected boom–bust episodes generate a boom–bust episode in aggregate activity, which, in turns, can feedback and amplify the boom and bust in the financial and housing markets. Given that the relationship between the house price boom–bust episode and the credit boom–bust episode can be itself quite rich, in this chapter we will mostly focus on that, and less on the connection with the real economy.

We start the chapter by discussing a simple mechanical baseline model in Section 2, which is meant to describe the interaction between the credit cycle and house prices, highlighted above. On purpose, it makes a number of stark assumptions to avoid several thorny issues that arise in a fully specified equilibrium model. In particular, we take as given the dynamics of both leverage and house prices. We then perform two types of



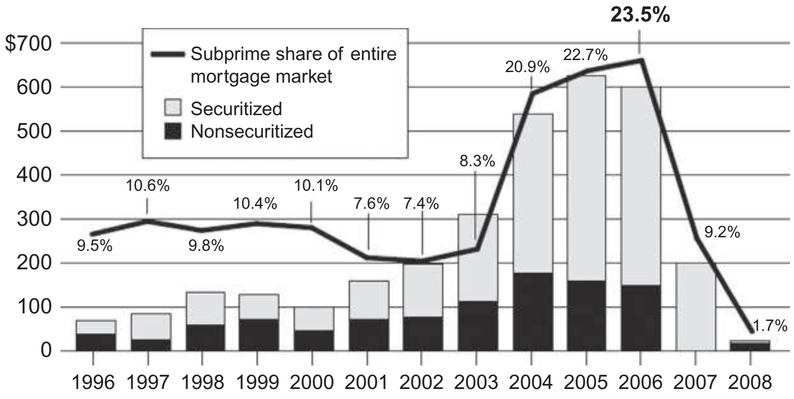
Source: S&P Indices & FiServ

Fig. 1 The S&P/Case-Shiller home price indices. *http://www.worldpropertyjournal.com/north-america-residential-news/spcase-shiller-home-price-indices-report-fordecember-2011-case-schiller-home-price-index-median-home-prices-the-national-composite-10-city-composite-index-20-citycomposite-home-price-indices-david-m-blitzer-5351.php*

Subprime Mortgage Originations

In 2006, $600 billion of subprime loans were originated, most of which were securitized. That year, subprime lending accounted for 23.5% of all mortgage originations.

In billions of dollars



Note: Percent securitized is defined as subprime securities issued divided by originations in a given year. In 2007, securities issued exceeded originations.

Source: Inside Mortgage Finance

Fig. 2 Subprime mortgage originations. *The Financial Crisis Inquiry Report, National Commission, January 2011.*

exercises: first, we keep house prices constant and study the response to a boom and bust in leverage; and second, we keep the leverage constant and study the response to a boom and bust in house prices. In the rest of the chapter, we try to go deeper in understanding where these dynamics are coming from and connect these two types of exercises to the existing literature.

We next explore the idea that the boom and bust in the credit market was the fundamental shock that spilled over the housing market and the real economy. We first discuss several papers related to this idea. In Section 3, we discuss a large class of papers that explore in particular the role of leverage in the households' sector. We focus on papers that study the effects of credit constraints on house prices and, more generally, on the real economy.

In this spirit, in Section 4, we completely abstract from the housing price dynamics and focus on the boom and crash in the credit market. In particular, we propose a stylized static model of households who borrow to become homeowners and intermediaries who lack information about the quality of the borrowers. The main idea is that a simple increase in the credit availability in the economy, what we interpret as "saving glut," can endogenously generate first a boom and then a crush in lending activity because of multiple equilibria due to adverse selection issues. The basic idea that a saving glut can generate an endogenous credit cycle because of multiplicity of equilibria is inspired by Boissay et al. (2016), although the mechanism is quite different. In our model, the increase in credit availability first increases the lending activity and hence increases the "subprime market." However, as the quality of the pool of borrowers decreases, good borrowers may decide to pay a cost to separate themselves and get better credit terms. This, in turns, may make the subprime market to collapse.

Next, we move to the idea that the boom and crash in housing prices is the key element of the interaction between housing and credit markets. In particular, if house prices are expected to rise, banks are more willing to lend, although this means to lend to worse creditors, eg, subprime borrowers. Moreover, speculating households are more willing to buy as house prices appreciate. While appealing, formalizing this intuition tends to run into the "conundrum of the single equilibrium": if prices are expected to be high tomorrow, then demand for credit and thus housing demand should be high today, and that should drive up prices today, making it less likely that prices will increase. Or, put differently, as prices rise, they eventually must get to a near maximum at some date: call it "today." At that point, prices are expected to decline in the future. But if so, banks and speculating households are less likely to buy today: but then, the price should not be high today. The issue is that in a rational expectations equilibrium there should be no "fool" willing to buy at the highest price, when prices can only go down from there.

In Section 5, we discuss two strands of literature that focus on two types of bubble models. In Section 5.1, we refer to a class of bubble models, where the interest rate demanded on assets is below or at most equal to the growth rate of the economy. This can give rise to rational bubbles and stochastically bursting bubble in, essentially, dynamically inefficient (or borderline efficient) overlapping generations models, as in Carvalho et al. (2012) or Martin and Ventura (2012). These authors employ various versions of OLG models, in which, ideally, resources should be funneled from inefficient investors or savers to efficient investors or entrepreneurs. They assume that there is a lending friction: entrepreneurs cannot promise repayment. They can only issue securities, where the buyer hopes that someone else buys them: call them "bubble," "cash," or "worthless pieces of paper." Equilibria then exist, where newborn entrepreneurs create "bubble" paper. The existing bubble paper in the hands of old agents and bubble paper created by newborn entrepreneurs get sold to savers. Savers find investing in these bubbles more attractive than investing in their own inefficient technology. This technology needs to be inefficient enough so that its return is on average below the growth rate of the economy, creating the dynamic inefficiency for bubbles to arise. In that case, the "fundamental" value of any asset paying even a tiny amount per period is actually infinite. Or, put differently, the last fool to buy the bubble at the highest price is happy to do so, since the value of the bubble next period will not have gone down too much and since that fool is desperate to save.

In Section 5.2, we discuss a second class of bubble models, where the interest rate demanded on assets is above the growth rate of the economy. Here, an aggregate bubble eventually must stop growing, being bounded by the resources in the hands of "newborn" agents purchasing these assets. Rationality considerations typically rule out such bubbles, see the "conundrum" above. We therefore investigate models with irrational optimism and changing sentiments. A benchmark example in the literature, exploiting changing sentiments, is the "disease" bubble model of Burnside et al. (2013). There is some intrinsically worthless bubble component, which could be part of the price of a house. An initially pessimistic population may gradually become infected to be "optimistic" and believe that the bubble component actually has some intrinsic value: once, everyone is optimistic (forever, let's say), there is some constant price that everyone is willing to pay. However, "truth" may be revealed with some probability every period and reveals that the bubble component is worthless indeed. Then, during the pessimistically dominated population epoch, prices rise during the nonrevelation phase, since the rise in prices compensates the pessimistic investors for the risk of ending up with a worthless bubble piece, in case the truth gets revealed. The price will rise until the marginal investor is optimistic: at that point, the maximum price may be reached.

In the spirit of this stream of sentiment literature, in Section 6, we propose a simple model where prices are above fundamentals because investors buy assets under the irrational belief that there is always going to be an ever more foolish buyer, willing to buy for a higher price.

Finally, in Section 7, we juxtapose our findings to some lessons we have drawn from the existing literature regarding the empirical evidence.

We wish this chapter will trigger further research and thinking on this important connection. As shall become clear, the issues are far from resolved.

2. A STARK MODEL

In the financial crisis of 2008, the following interplay might be at work, amplifying any initial shock: (1) as house prices fell, banks became more reluctant to lend to new home buyers, and (2) as banks became more reluctant to lend to new home buyers, demand for houses and thus prices for houses fell as a consequence. In particular, banks became more reluctant to lend because the drop in house prices negatively impacted their balance sheets, hence generating a more general credit crunch, depressing real activity.

In this section, we introduce a very simple mechanical model, featuring some of that interplay, but without describing the deep reasons for some key elements. The model features a potential mismatch between the long-term assets in the form of a pool of mortgages and the short-term assets in the form of saver deposits. It allows us to study the evolution of bank balance sheets during a house price boom. The model is useful for providing some key insights regarding price crashes and leverage crashes, and their impact on the financial system. In particular, we will use it to conduct numerical experiments, illustrating the two channels emphasized in the introduction:

1. House prices are constant and there is a boom and bust in the leverage ratio.
2. The leverage ratio is constant, and there is a boom and bust in house prices.

Furthermore, the model and its analysis sets the stage for discussing the related literature and for the latter sections of the chapter.

Time is discrete and infinite, $t = \ldots, -1, 0, 1, \ldots$. There is a continuum of households, who are the borrowers. There is a competitive sector of bankers, each operating a bank. There is a group of savers who exogenously supply deposits to the banks, and a government which assumes the role of a special saver. Finally, there is a numeraire consumption good and a housing good (or simply houses).

Each period, a fraction λ of the households exit (or "die"), and they are replaced with a fraction λ of newborn households. Each period all alive households earn some exogenously fixed income y and consume a nonnegative amount of goods, while newborn households earn some initial income $\widetilde{y}$ and buy a house. We allow $\widetilde{y}$ to differ from y, reflecting a potential period of saving-up before purchasing the first home. Just before a household exits, it sells its house. Houses are in fixed supply and are identical to each other.

We assume that a household born at time s is willing to buy a house for any price $p_s \leq \bar{p}_s$, where the process for $\bar{p}_s \geq 0$ is exogenously given.[a] When $\widetilde{y}$ is not large enough, households have to borrow in order to make that purchase. Restrictions to borrowing may then imply that the newly born households have less than $\bar{p}_s$ resources at hand.

[a] In principle, one could introduce preferences giving rise to this behavior.

We assume that the sellers get to extract all the rents, ie, we assume that the newly born households pay the lesser between the resources available and $\bar{p}_s$. Thus, only borrowing restrictions may force the market price p_s below $\bar{p}_s$.

To buy their house, households borrow from a banking sector. Consider a household born at date s who buys the house at the prevailing market price p_s. We assume that the following mortgage contract is the only type of contract offered by banks and available to households. In the initial period, households have to make a down payment of $\theta < \tilde{y}$ and borrow the remainder $l_s = \max\{p_s - \theta, 0\}$. We shall focus on parameter specifications such that in equilibrium $p_s \geq \theta$ for all s. The contract requires that households repay the principal l_s when they exit and sell their house. Failing that, they pay all resources available to them in that exiting period. In all other periods, including the period of purchase, households pay a flow interest r per unit of principal borrowed. We treat θ and r as parameters of the model. We assume that $r > 0$, while we do not necessarily restrict θ to be positive, allowing for a cash out at the time of purchase of a house when $\theta < 0$.

We will focus the analysis on equilibria with $l_s = p_s - \theta \geq 0$, where equality is the autarkic case when households do not borrow from banks. Hence, the consumption of a household born at time s in her first period of life is equal to $c_{s;s} = \tilde{y} - \theta - r(p_s - \theta)$, where the first index of $c_{s;s}$ refers to the date of consumption and the second index refers to the year of birth. As we do not allow for negative consumption, that is, $c_{s;s} \geq 0$, prices are bounded above by

$$p_s \leq p^{\max} = \frac{\tilde{y} - (1 - r)\theta}{r}. \tag{1}$$

In any subsequent period, the household will learn if she exits at the end of that period. The nonexiting households then consumes $c_{t;s} = y - r(p_s - \theta) \geq 0$, imposing another constraint on house prices:

$$p_s \leq \frac{y}{r} + \theta. \tag{2}$$

We will concentrate on parameter specifications, where (2) is tighter than (1), that is, we assume that $\tilde{y} - y > \theta$. If the household exits at time t, she sells her house at current market price p_t. If $p_t + y \geq (1 + r)(p_s - \theta)$, she can repay the interest and the principal, and before exiting can consume $c_{t;s}^{f} = p_t + y - (1 + r)(p_s - \theta)$, where f is meant to indicate her "final period." If $p_t + y < (1 + r)(p_s - \theta)$, the household defaults, consumes zero, and the bank receives $p_t + y$ in total, which one can split into $r(p_s - \theta)$ as the interest portion and $p_t + y - r(p_s - \theta)$ as the partial repayment of principal. One can then calculate the fraction $\phi_{t;s}$ of principal repaid by households born at date s and exiting at date t by solving for $\phi_{t;s}$ the following equation:

$$\phi_{t;s}(p_s - \theta) = \min\{p_s - \theta, p_t + y - r(p_s - \theta)\}. \tag{3}$$

The default rate is then $1 - \phi_{t;s}$.

We assume that banks discount future periods at the same rate r that they charge as interest payments on the mortgages: this is the easiest case to analyze. Consider a scenario in which households never default. Then, the date-t value $v_{t;s}$ of a contract signed at date s is independent of t, $v_{t;s} \equiv v_s$ and satisfies the recursion

$$v_s = \frac{1}{1+r}\left(r(p_s - \theta) + (1-\lambda)v_s + \lambda(p_s - \theta)\right),$$

which gives

$$v_s = p_s - \theta. \tag{4}$$

Banks only invest in mortgages. We assume that banks allow newly born households to borrow as much as they wish to borrow, provided banks have the resources to let them do that.

On the liability side, we assume that banks have deposits d_t by a group of savers as well as a deposit or loan L_t by the government. The bank pays some rate r_D per unit of deposit by savers. On the government loans, the bank pays an interest r_L, which is treated as an exogenous parameter. Additionally, banks are required to repay an exogenously given fraction μ of the principal.

To close the model, we need to specify the evolution of d_t and L_t. We choose an exogenous process for the banks' leverage ratio and set d_t to match such a process, given the endogenous value of the banks' assets. This is meant to be a simple stand-in for the view that banks finance projects by maximizing the amount of outside financing, subject to constraints on their leverage from regulatory restrictions or repayment concerns by depositors.

To calculate the value of a bank's assets, we need to take a stand on how the bank or, implicitly, some (unmodeled) regulator values the portfolio of its mortgages. We shall assume v_s to be the **book value** of a mortgage issued in period $s \leq t$ and which has not been repaid, even if the expected value or **market value** of this mortgage has been declining, due to house price decline and default considerations. Let a_t be the sum of all end-of-period book values of remaining mortgages, that is,

$$\begin{aligned} a_t &= \sum_{j=0}^{\infty} \lambda(1-\lambda)^j (p_{t-j} - \theta) \\ &= \lambda(p_t - \theta) + (1-\lambda)a_{t-1}, \end{aligned} \tag{5}$$

given that only young households, that is, a fraction λ of the population, purchase a home in each period and given Eq. (4). For example, if prices are constant forever, $p_t \equiv p^*$, then

$$a_t \equiv p^* - \theta. \tag{6}$$

Consider the balance sheet at the end of the period. We assume that the liabilities are recorded at their face value. The differences between assets and liabilities is the net worth

n_t of the bank. The (book value) net worth n_t of the bank then results from the balance sheet equation

$$a_t = d_t + L_t + n_t \tag{7}$$

We define the capital requirement or net worth requirement κ_t per

$$\kappa_t a_t = n_t + L_t, \tag{8}$$

or

$$(1 - \kappa_t) a_t = d_t, \tag{9}$$

effectively treating the government loan L_t as a perfect substitute for net worth. We choose an exogenous stochastic process for $\kappa_t \in [0, 1]$ and assume that d_t is set so as to satisfy Eq. (9). Note that $1/\kappa_t$ is the book-value leverage ratio on $n_t + L_t$.

For the evolution of L_t, we consider two alternative versions of the model. The central issue is how to treat a shortfall of funds, should it occur. For simplicity, we seek specifications of the model that avoid potential defaults on depositors, although it would be interesting to explore an extension of the model with default. In the baseline version of the model, we assume that bankers themselves inject any needed funds and hence we assume $L_t \equiv 0$. In the alternative version of the model, we shut down the channel of the injection of bank equity, and instead assume that the government provides loans, if necessary, to avoid a default on depositors and to avoid a shortfall of regulatory capital. For both versions, we need to calculate the evolution of the balance sheet.

Consider the beginning of a new period, after exiting households have sold their houses to newly born households. Let us trace out the impact of each transaction on the residual net worth. The bank receives interest payments ra_{t-1} on all outstanding mortgages, increasing net worth by that amount. A fraction λ of outstanding mortgages exits. Let us define ϕ_t the fraction of principal exiting mortgages that is repaid to the bank, so that the bank receives $\phi_t \lambda a_{t-1}$ in total. Using Eq. (3), we obtain

$$\phi_t a_{t-1} = \sum_{j=0}^{\infty} \lambda (1-\lambda)^j \min\{p_{t-1-j} - \theta, p_t + \gamma - r(p_{t-1-j} - \theta)\} \tag{10}$$

The resulting net worth loss is $(1 - \phi_t) a_{t-1}$, as the book value a_{t-1} of the exiting mortgages is replaced by their payoff $\phi_t a_{t-1}$. In particular, if the current market price is at least as high as all past market prices, then $\phi_t = 1$ and there is no change in net worth.

The bank also receives an inflow of new deposits $d_t - d_{t-1}$, new government loans L_t, and makes new mortgage investments $\lambda(p_t - \theta)$ which do not change net worth, but just lengthen the balance sheet.

On the liability side, banks pay the market interest rate r_D per unit of deposit, so that net worth decreases by the total payments $r_D d_{t-1}$. Furthermore, the bank pays $(r_L + \mu)$ L_{t-1}, the interest and a fraction μ of the principal on the beginning-of-period government loans. After all these transactions, but excluding the new government loan position L_t, the bank has a residual cash position m_t on the asset side, expressed in units of the consumption good. This position may be negative and can be expressed as follows:

$$m_t = (r + \phi_t \lambda) a_{t-1} + d_t - (1 + r_D) d_{t-1} - (\mu + r_L) L_{t-1} - \lambda(p_t - \theta).$$

Finally, we assume that the banker consumes some amount $c_{b,t}$, reducing the net worth of its bank by that amount. It may be useful to think of this consumption as a payment to bank shareholders. In the baseline version of the model, we assume that $L_t = 0$ and that $c_{b,t} = m_t$, that is it exactly equals the cash position, so that the postbanker consumption cash position is equal to zero. Since that cash position can be negative, we must allow $c_{b,t}$ to be negative as well. One might wish to think of this as an injection of equity by the existing bank owners.

The equilibrium of the baseline model with the assumption that $L_t = 0$ can be characterized by the following equations:

$$a_t = \lambda(p_t - \theta) + (1 - \lambda) a_{t-1}, \tag{11}$$

$$d_t = (1 - \kappa_t) a_t, \tag{12}$$

$$m_t = (r + \phi_t \lambda) a_{t-1} + d_t - (1 + r_D) d_{t-1} - \lambda(p_t - \theta), \tag{13}$$

$$c_{b,t} = m_t, \tag{14}$$

$$n_t = a_t - d_t, \tag{15}$$

where ϕ_t is given by Eq. (10). Note that substituting for a_t and d_t using Eqs. (11) and (12) in (13) we obtain

$$\begin{aligned} m_t = & [r - r_D + r_D \kappa_{t-1} + \lambda \kappa_t + \kappa_{t-1} - \kappa_t - (1 - \phi_t)\lambda] a_{t-1} \\ & - \kappa_t \lambda (p_t - \theta) \end{aligned} \tag{16}$$

This equation has an intuitive appeal. Consider the bracket, multiplying a_{t-1}. The first term, $r - r_D$ is the interest arbitrage collected. The second term $r_D \kappa_{t-1}$ is the interest earned on the net worth portion of a_{t-1}. The third term, $\lambda \kappa_t$ concerns the repayment of principal. The difference $\kappa_{t-1} - \kappa_t$ means that cash is freed up, if the capital requirement κ_t decreases. The final term $(1 - \phi_t)\lambda$ reduces cash flow only if there are defaults, $\phi_t < 1$.

Moreover, substituting for a_t using Eq. (11) into (15), and using Eq. (15) one period backward, after some manipulation we obtain

$$n_t = n_{t-1} - \lambda a_{t-1} + \lambda(p_t - \theta) - d_t + d_{t-1}. \tag{17}$$

One can use this equation to examine the evolution of net worth. As one special case, suppose that the evolution for the exogenous process κ_t implies that deposits are constant, $d_{t-1} = d_t = d$. Then,

$$n_t = n_{t-1} - \lambda a_{t-1} + \lambda(p_t - \theta), \tag{18}$$

ie, the change in net worth is given by the book value difference between newly created and exiting mortgages. At a superficial look, it would appear that net worth is "magically" created by higher prices and that default on exiting mortgages does not matter. However, it needs to be recognized that these movements find their counterpart in the banker's consumption $c_{b,t}$, see (14): to keep d unchanged, higher prices for new houses as well as larger defaults on old mortgages reduce these shareholder payouts or even require equity injection.

Eq. (18) also reveals that net worth stays constant, if deposits are constant and prices are constant, as, according to Eq. (6), constant prices imply $a_{t-1} = p^{*} - \theta$. In this case, Eq. (14) implies that bankers' consumption is equal to

$$c_b^* = r(p^* - \theta) - r_D d^*. \tag{19}$$

This simply says that the interest payments on the assets, reduced by the interest payments on the liabilities, are the flow profits in this steady state situation.

Finally, the house price is easy to characterize in this baseline version of the model:

Proposition 1 *Assume that $\bar{p}_t \leq p^{\max}$, defined in Eq. (1). In the baseline version of the model, the house price is then always equal to the exogenous process, that is, $p_t = \bar{p}_t$.*

Proof This follows from the assumption that sellers extract all the rent from buying households, ie, newly born households are willing to borrow up to $\bar{p}_t - \theta$, and the assumption that banks let them do so, potentially financing the needed resources with negative banker consumption. □

For the alternative version of the model, we impose the restriction that $c_{b,t} \geq 0$, that is, the banks cannot raise equity from their owners. We assume that the government provides loans L_t, making up for any potential shortfall. The equations characterizing the equilibrium are now:

$$a_t = \lambda(p_t - \theta) + (1 - \lambda)a_{t-1}, \tag{20}$$

$$d_t = (1 - \kappa_t)a_t, \tag{21}$$

$$\begin{aligned} m_t = {} & (r + \phi_t \lambda)a_{t-1} + d_t - (1 + r_D)d_{t-1} \\ & - \lambda(p_t - \theta) - (r_L + \mu)L_{t-1}, \end{aligned} \tag{22}$$

$$c_{b,t} = \max\{0; m_t\}, \tag{23}$$

$$L_t = (1 - \mu)L_{t-1} - \min\{0, m_t\}, \tag{24}$$

$$n_t = a_t - d_t - L_t, \tag{25}$$

where again ϕ_t is given by Eq. (10). Eq. (22), compared to (13), includes the payment of the interest and of a portion μ of the principal on the outstanding government loans. Eq. (23) encodes the nonnegativity of $c_{b,t}$, compared to (14). With that, one needs to add Eq. (24) for the evolution of the government loans, which are reduced by the repayment of the principal portion, but are increased by any need of repayment for a shortfall of funds $m_t < 0$.

The alternative model is not yet complete, however. Note that larger p_t in (22) can now be compensated for by correspondingly larger loans L_t by the government. Indeed, there is potentially an interesting range of policies to consider. At the one and most generous extreme, the government may provide sufficiently large loans so as to reestablish the maximal price $p_t = \bar{p}_t$, which households are willing to pay. At the other and most stingy extreme, the government may only provide loans to assure nonnegative banker consumption, with house prices reduced all the way to $p_t = \theta$ and thus not requiring bank loans for purchases (assuming $\theta \geq 0$). In the numerical exercise for the exogenous price crash below, we shall investigate the implications of the latter extreme. Put differently, we pick the highest price p_t with $\theta < p_t \leq \bar{p}_t$, subject to the restriction that the resulting m_t in (22) is nonnegative, provided such a price exists. Note that this price will either equal $\bar{p}_t$ or result in $m_t = 0$. If no such price exists, then $p_t = \theta$, $m_t < 0$ and the newly issued loan will equal $-\min\{0, m_t\}$, as stated in Eq. (24). With that, the house price becomes endogenous, and Proposition 1 ceases to hold. In the case of a bust in the leverage ratio, we assume that the government provides a loan to the banks to make up the missing equity. A better interpretation is to view this as a partial stake in the banking system, at a required rate of return for the government. This stake is then reduced over time at the assumed required rate of the loan repayment.

2.1 Numerical Experiments: Overview

We now conduct two sets of numerical experiments to highlight the two approaches we discussed in the introduction:

1. We assume that house prices are constant and assume a boom and bust in the leverage ratio κ;
2. We assume that the leverage ratio is constant, and assume a boom and bust in house prices dynamics.

For both exercises, we consider the implications both for the baseline specification, when bankers can inject fresh equity, and the alternative specification, when they cannot and when, potentially, government loans are required to cover shortfalls of resources.

The numerical exercises are meant to be illustrative, and are not intended as careful calibrations. The parameters are picked to be broadly reasonable, but the results are quite sensitive to their choices. We shall think of a period as 1 year. An overview of the parameters is in Table 1.

Everything scales with income y, so we arbitrarily set income $y = 1$. Hence, one can read all quantities such as banker consumption, government loans or assets, as multiples of annual GDP. We assume a down payment (relative to income) of $\theta = 2$, which should be assumed to be "saved up" from prior income before agents are born and enter the

Table 1 Parameter values for the numerical experiments

γ	1
p^*	5 y
κ	0.05 (precrash, experiment 1)
	0.2 (postcrash, experiment 1)
	0.1 (always, experiment 2)
θ	2 y
r	0.04
r_D	0.03
r_L	0.03
μ	0.05
λ	0.1
γ	1.13 (experiment 2)
α	19 y (experiment 2)

housing market. That is, we assume that $\tilde{\gamma}$ is high enough, so that (2) is tighter than (1). Since $\tilde{\gamma}$ does not play a role otherwise, we have not listed an explicit value in Table 1. The exit probability λ has been set equal to 0.1, implying a turnover of a house on average every 10 years. We assume that banks earn 4% on their assets and pay 3% on their liabilities, be they depositors or government loans. We assume that government loans have a maturity of 20 years, ie, that the fraction $\mu = 0.05$ of the outstanding bonds need to be repaid each period.

For the first set of numerical experiments, we set the maximal willingness to pay constant at $\bar{p}_t \equiv p^*$, where $p^* = 5$ y, and thus as five times (annual) income. To model the boom and subsequent bust in leverage, we assume that the required capital ratio is initially at $\kappa = 0.05$ until some date $t = -1$, implying a leverage ratio of 20, and then unexpectedly rises to $\kappa = 0.2$ at date $t = 0$, implying a leverage ratio of 5.

For the second set of numerical experiments, we keep the required capital ratio constant at $\kappa = 0.1$. To capture an initial run-up of house prices and subsequent crash, we assume that the maximal house prices $\bar{p}_t$ increase exponentially until $t = -1$ and then drop to some constant level $\bar{p}_t \equiv p^* \geq \theta$, where $p^*/\gamma = 5$, which is also comparable to the distant past. That is, we assume that

$$\bar{p}_t = p^* + \alpha\gamma^t \tag{26}$$

for $t < 0$, and $\bar{p}_t = p^*$ for $t \geq 0$, where $\gamma \geq 1$.

2.2 An Exogenous Crash in Leverage

Let us examine the first set of numerical experiments, with a constant maximal price $\bar{p}_t = p^*$ and an exogenous crash in leverage. "Case A" is the benchmark version of the model, where we assume that bankers supply fresh equity, if needed. This is

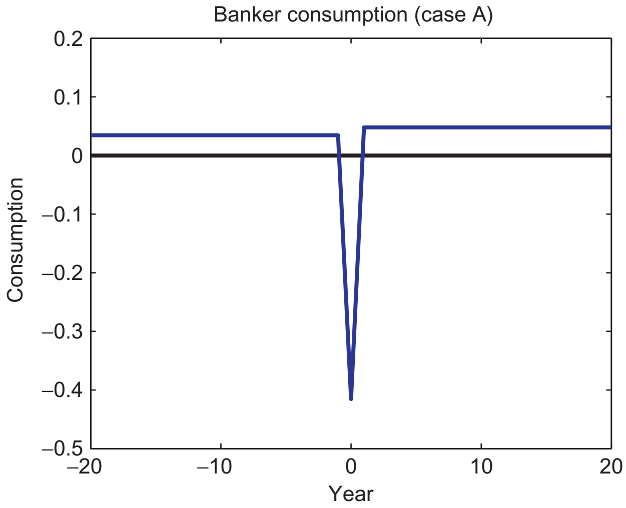


Fig. 3 An exogenous crash in leverage: Implication for banker consumption. Negative values should be interpreted as the injection of fresh bank equity.

shown in Fig. 3. There is a single period, when the leverage ratio suddenly changes, necessitating an infusion of extra cash, modeled as negative banker consumption. Once the fresh equity is injected, everything continues as before, except that banker consumption is now higher, given the new and lower leverage. The price for houses remains at $p_t = \bar{p}$.

Matters are more dramatic for "case B," the alternative specification of the model, where there is no fresh infusion of bank equity. The results are shown in Fig. 4. Prices crash endogenously, as can be seen in the top left panel. There is a fairly brief period of default, as shown in the top right panel. The government makes up the missing equity by, essentially, obtaining a partial stake in the banking system, at a required rate of return for the government. This stake is then reduced over time at the assumed required rate of the loan repayment. If the payments for interest and repayments are less than the revenue of the banking system, the consumption of the bankers are positive, as can be seen here. Finally, banks gradually rebuild their net worth to the required new ratio, as indicated by the red-dashed line in the left panel of the third row.

2.3 An Exogenous Crash in House Prices

For the second set of numerical experiments, we seek to investigate an exogenous crash in house prices, following a phase of increasing house prices, given by (26), while keeping leverage κ constant.

Consider first the run-up phase for house prices, $t < 0$. In the benchmark specification of the model, houses are always sold at the maximum price that newborn home buyers are

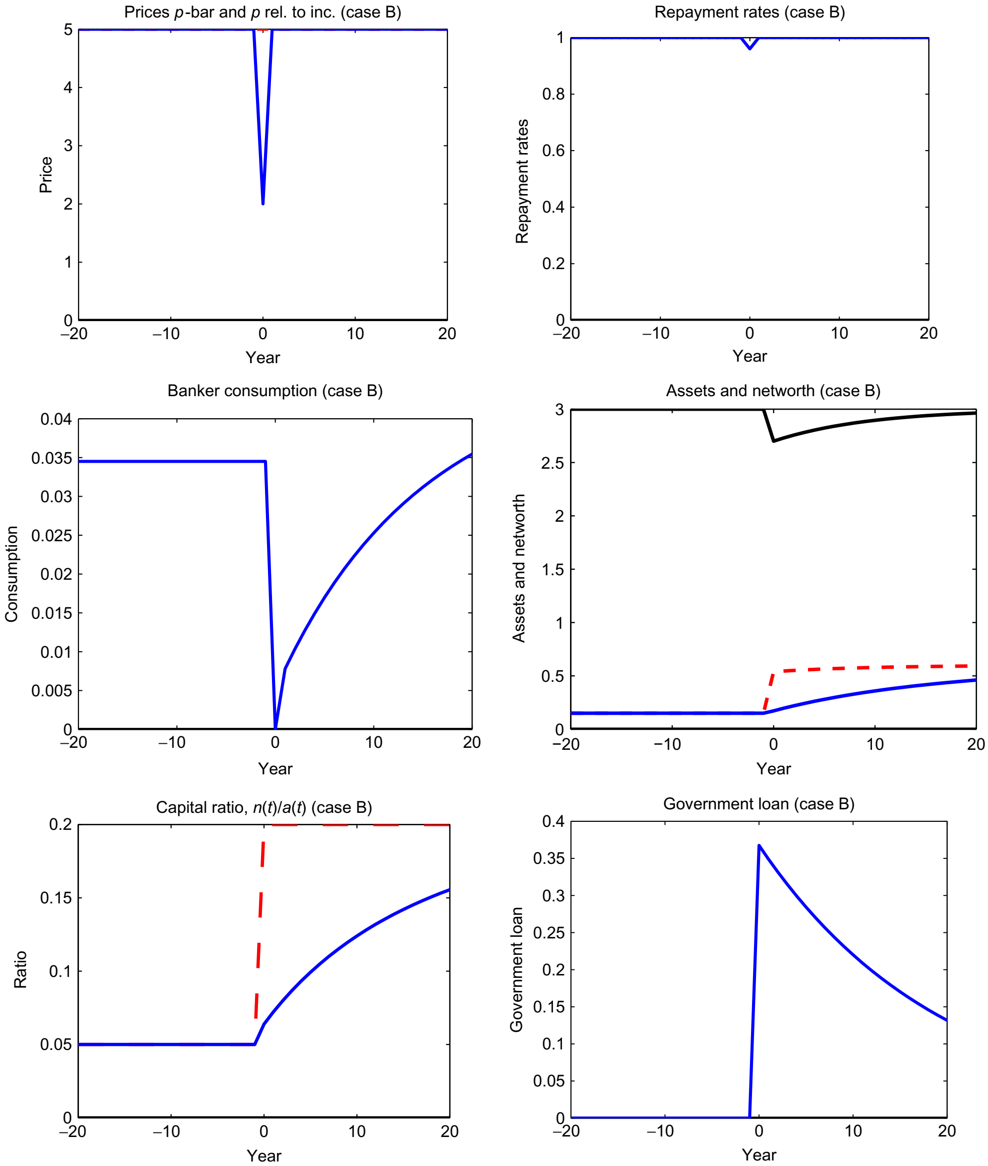


Fig. 4 An exogenous crash in leverage without infusion of bank equity. Prices crash endogenously, as can be seen in the top left panel. There is a fairly brief period of default, as shown in the top right panel. For the parameterization here, the banks gradually rebuild their net worth to the required new ratio, as indicated by the red (gray in the print version)-dashed line in the left panel of the third row.

willing to spend, ie, $p_t = \bar{p}_t$. This will also be true in the alternative specification of the model, provided that banks are able to build up net worth fast enough to finance the new loans, without necessitating the injection of further equity. This imposes some constraints on the parameters, which we shall illuminate.

Using (26) and $p_t = \bar{p}_t$ for all $t < 0$, assets a_t per Eq. (11) can be rewritten as

$$a_t = p^* - \theta + \frac{\lambda\gamma}{\gamma + \lambda - 1}(p_t - p^*)$$

for $t < 0$.

As a useful benchmark, assume $p^* = 0$ and $\theta = 0$. The asset-to-price ratio then is

$$\frac{a_t}{p_t} = \frac{\lambda\gamma}{\gamma + \lambda - 1}.$$

This relationship between current price and the stock of outstanding assets for $t < 0$ is plotted in Fig. 5. As one can see, higher house price growth makes assets look small compared to current house prices. One may interpret this as a reason, why financial institutions are less concerned about default risks during price booms. With (27) and $a_t = \gamma a_{t-1}$ for $t < 0$, Eq. (13) implies

$$\frac{m_t}{a_t} = \frac{1}{\gamma}(r - r_D(1 - \kappa) - \kappa(\gamma - 1)) \tag{27}$$

The balance sheets are growing for $t < 0$. If the banks finance this growth exactly out of earnings, so that $c_{b,t} = m_t = 0$, one obtains

$$\gamma = 1 + \frac{1}{\kappa}(r - r_D(1 - \kappa)) \tag{28}$$

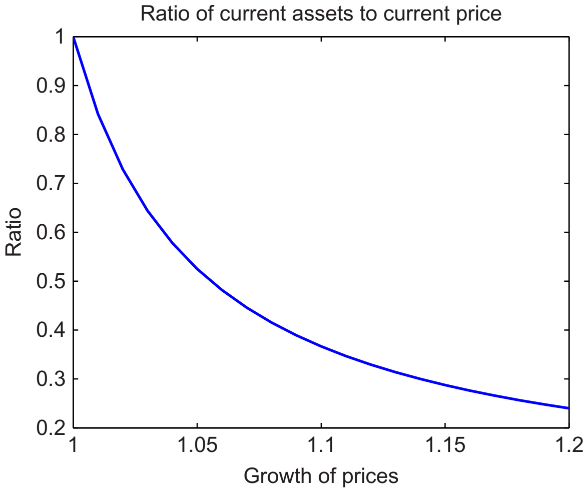


Fig. 5 Ratio of assets to prices, for various values of γ, when $\lambda = 0.05$.

which is intuitive. In particular, if $r = r_D$, then

$$\gamma = 1 + r \tag{29}$$

so that the interest paid on net worth must exactly finance its growth. Put differently, the values calculated for γ in (28) or (29) are the upper bounds for the price growth rates γ to avoid that banker consumption falls into negative territory during the price growth phase, when holding leverage constant, and when assuming that $p^* = 0$ and $\theta = 0$.

With the other parameters as listed in Table 1, Eq. (28) implies that $\gamma = 1.13$ or a 13% appreciation of maximal house prices, during the run-up phase. While the numerical experiments are intended as illustrations only, this number strikes us as perhaps a high, but not entirely unreasonable value during a house boom phase. Indeed, during the pre-2008 years, house prices grew even faster, towards the end, according to the Case-Shiller index. One may also wish to read this as a reasonable upper bound of long-time house price growth, when banks are constrained from raising new equity for financing new mortgages. Consider then the implications for banker consumption in Figs. 6 and 7. For this parameter choice, they are almost flat in the precrash phase, since the rise in higher interest payments on old mortgages is now nearly offset by the rise in resources needed for paying for new mortgages. For other choices for γ, one should not expect nearly flat banker consumption during this run-up phase.

Per Eq. (2), we must be careful in letting prices grow too large. Examining the restriction at the last precrash price $\bar{p}_{-1}$, this equation implies that

$$\alpha \leq \frac{1}{\gamma}\left(\frac{\gamma}{r} + \theta - p^*\right) \approx 19.5\,\gamma \tag{30}$$

We set $\alpha = 19\,\gamma$, so that prices crash in the last possible period. These values for γ and α are listed in Table 1. Arguably, these are pretty much at the extreme end, and chosen to provide the most dramatic numerical experiment.

Consider now the postcrash phase, $t \geq 0$. Here, numerical calculations are required. The results are in Figs. 6 and 7. Fig. 6 shows what happens in the benchmark specification "case A" of the model, when bank equity can be injected, ie, when banker consumption can become negative. House prices trade at the exogenously given levels $\bar{p}_t = p^*$. There is a temporary dip in repayments, but they recover gradually, as the top right panel shows. No government loans are necessary or provided in this case.

"Case B" is the alternative specification of the model, when no fresh injection of bank equity is available. The results are now more dramatic, and shown in Fig. 7. Now, when prices crash, they crash to the down-payment level θ and stay there, for the chosen parameter configuration. Banker consumption never recovers. There is continued

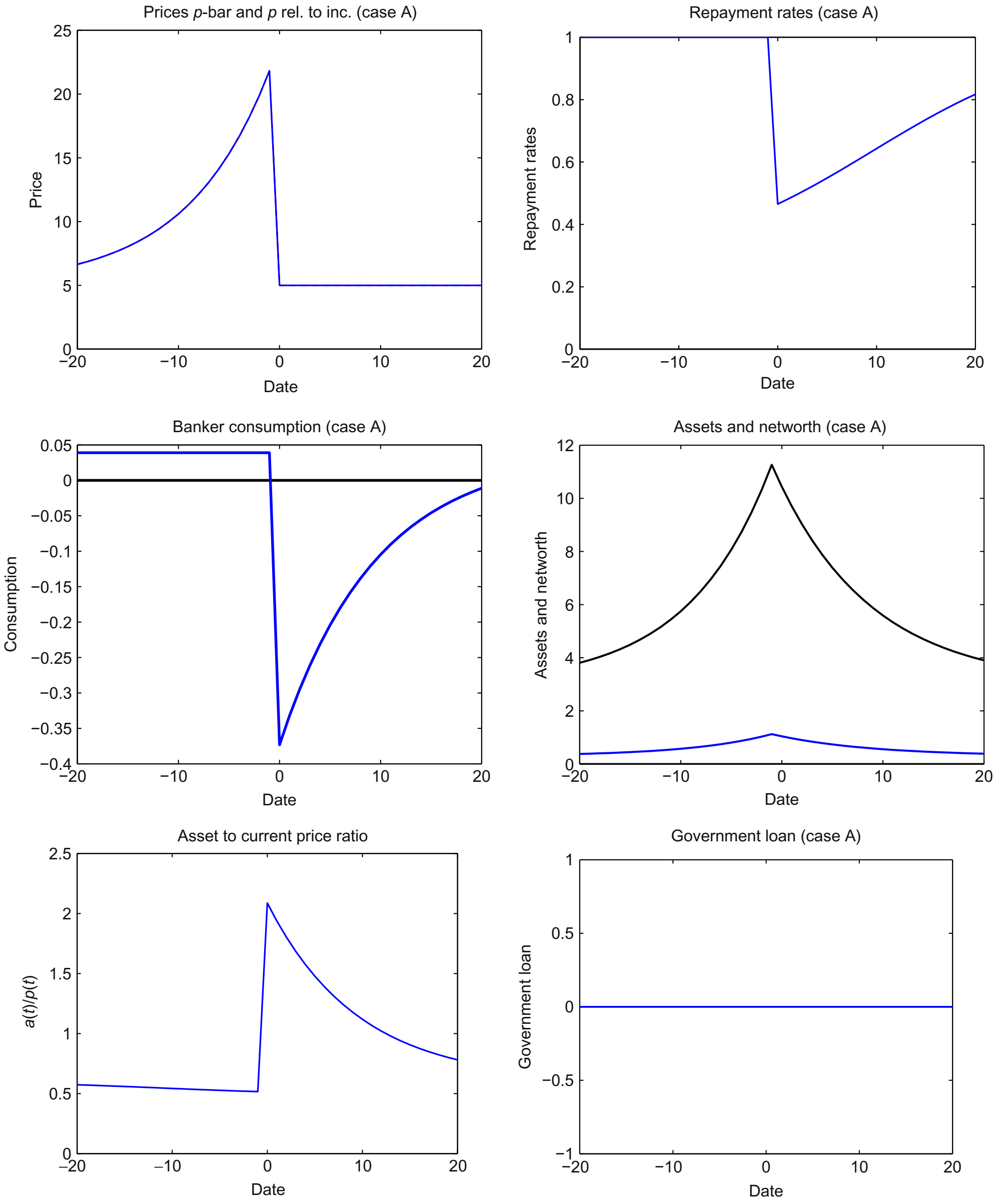


Fig. 6 House price boom and crash: Implications in the benchmark specification, when bankers inject fresh equity to cover shortfalls of funds.

default on banker assets. These are the legacy assets of precrash assets, which gradually disappear over time: households with precrash loans continue to have difficulties repaying these loans. Eventually, the government holds a bond position offset by a negative amount of net worth of bankers, without any corresponding assets.

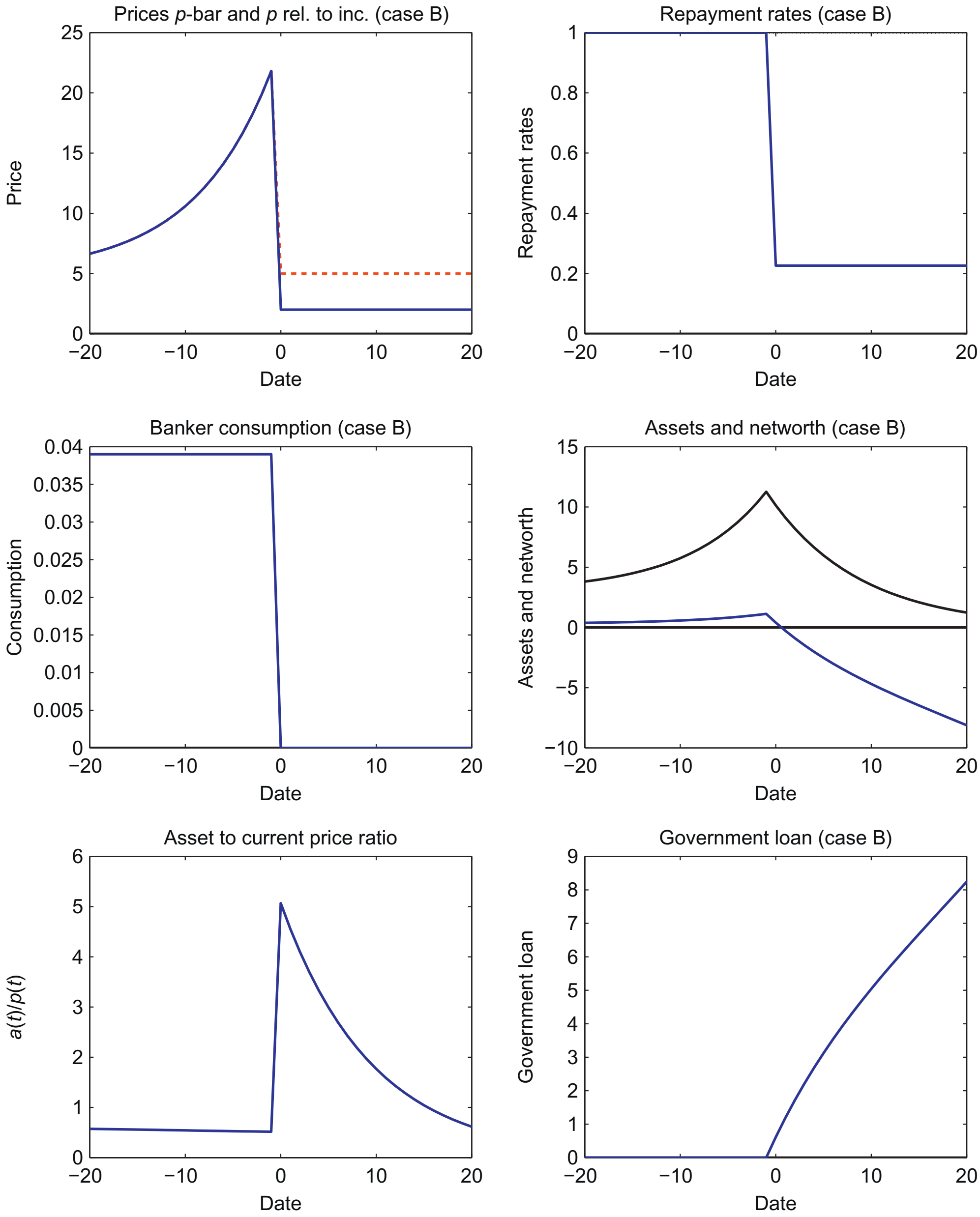


Fig. 7 House price boom and crash: Implications in the alternative specification, when bankers do not inject equity and the government provides the minimal loan to keep banks from defaulting.

2.4 Remarks

The previous numerical experiments are useful to highlight how a boom and bust in housing prices and financial markets can be qualitatively driven either by anything that affects directly house prices or by anything that affects directly banks' leverage. One important ingredient of the model that generates an interesting interaction between house prices and banks' leverage is the presence of long-term loans.

In the experiments above, a permanent decrease in leverage has more modest effects overall than an exogenous crash in house prices. However, as we repeatedly mentioned, these are only illustrative example, and the size of the exogenous crash that we imposed on leverage in the first exercise is difficult to compare to the size of the price crash that we imposed in the second exercise. It would be interesting in future work to calibrate a more realistic version of the model and try to do a horse race between the two types of shocks.

It is also interesting to highlight that in both types of numerical experiments, events unfold always more dramatically in "case B," without the fresh injection of bank equity, than in "case A." This indicates that a quick recapitalization of the banking system may be important in getting things back on track and avoiding long and persistent slumps. There is a self-feeding crisis here: without such an equity infusion, banks cannot fund new mortgages, house prices may remain low, leading to further defaults and leading to further impairments on bank balance sheets. A generous government loan program (not shown here), which supports house prices at the maximum willingness that households are willing to pay, will likewise insulate the housing markets from the drop in bank equity, but may result in keeping the government involved in the banking sector for a long time to come. Clearly, we do not model the costs of a possible government intervention, so conclusive policy recommendations are beyond the scope of this section. This is another interesting avenue for future work.

3. RELATED LITERATURE: HOUSEHOLDS' LEVERAGE

The simple model we introduced in the previous section highlights the interplay between credit cycles and boom–bust cycles in housing prices. One key ingredient in that interplay is that households borrow to become homeowners and are subject to financial constraints. Indeed one defining feature of the recent US experience is a dramatic increase in the households' gross debt to GDP ratio, which reached roughly 128% by 2008, and then sharply dropped. This drew a lot of attention on the effects of households' leveraging and deleveraging not only on housing markets, but, more generally, on aggregate activity.

3.1 Financial Frictions in Macro Models

There is a large growing literature that embeds financial frictions in macro model. Brunnermeier et al. (2011) is a comprehensive survey on this matter. We will just refer here to the seminal papers in this literature and focus next on the specific link between credit markets and house prices. One of the first papers that started a literature of macroeconomics models with financial frictions is Bernanke and Gertler (1989). They focus on long-lasting effects of temporary shocks through the feedback effect of a tightening of the financial frictions. In this model, as well as in Carlstrom and Fuerst (1997) and Bernanke et al. (1999), the key friction is the assumption of costly verification of the entrepreneur's type. Another seminal paper that had a huge impact on the macroeconomic literature is Kiyotaki and Moore (1997) who model financial frictions with a collateral constraint on borrowing rather than with a costly state verification framework. They propose a dynamic economy where durable assets play the dual role of factor of production and collateral for producers' loans. In their model, credit limits are endogenously determined and the interaction between them and asset prices generates a powerful transmission mechanism that allow temporary shocks to technology and income distribution to have large and persistent effects on asset prices and output. The mechanism is the following: after a temporary shock to productivity that reduces firms' net worth, constrained firms have to cut back their investment, hence reducing land value, and this hurts their future borrowing capacity, and reduces investment further down. This mechanism has been largely incorporated in macro models to study the real effects of financial shocks and the amplification of other types of shocks. Another influential paper on financial frictions and macro is Geanakoplos (2009) who focus on the role of leverage in boom and bust episodes. The key idea is that some investors are more optimistic than others and in good times they will lever up and drive asset prices up. However, if bad states realized, they may loose their wealth and the assets may shift in more pessimistic hands, and leverage and prices go down. This is what a leverage cycle is. Another related paper is Myerson (2012) who propose a model of credit cycles generated by moral hazard in financial intermediation.

There is a large recent literature that builds on these models to think about the role of firms' balance sheets in the macroeconomy. See for example Lorenzoni (2008), Mendoza and Quadrini (2010), Geanakoplos (2011), Brunnermeier and Sannikov (2010), He and Krishnamurthy (2013), and Bocola (2014). Gilchrist and Zakrajšek (2012) construct a new credit spread index and show that indeed a reduction in credit availability can have adverse macroeconomic consequences. However, in this chapter we focus more on the households' side and hence we tilt also the discussion of the literature in this direction.

3.2 The Effect of Credit Constraints on House Prices

There is a large strand of literature building models of the housing market where households' credit constraints play a crucial role in affecting house prices. Davis and VanNieuwerburgh (2015) also offers a nice overview of part of this literature. To the best of our knowledge, Stein (1995) is the first paper to explore the effects of down-payment requirements on house price volatility, as well as on the correlation between prices and trading volume. In particular, the paper highlights the self-reinforcing effect that runs from house prices to down payments and housing demand, back to house prices: if house prices decline, the value of households' collateral declines, depressing housing demand and hence pushing house prices further down. This multiplier effect can generate multiple equilibria and account for house price boom–bust episodes. This self-reinforcing effect is in the same spirit of the transmission mechanism in the seminal paper of Kiyotaki and Moore (1997).

In a related paper, Ortalo-Magne and Rady (2006) also explore the key role of down-payment requirements to explain house price volatility, although they focus on a different mechanism. They propose a life-cycle model of the housing market with credit constraints where there are two types of homes, "starter homes" and "trade-up homes." This allows them to focus on the key role of first-time buyers and show that income volatility of young households or relaxation of their credit constraints can explain excess volatility of house prices. Their model also delivers positive correlation between house prices and transaction volume.

More recently, Kiyotaki et al. (2011) develop a quantitative general equilibrium life-cycle model where land is a limited factor of production and is used as collateral for firms' loans. They show that, the more important is land relative to capital in the production of tangible assets, the more housing prices are sensitive to fundamental shocks as productivity growth rate or the world interest rate. Moreover, these type of shocks affect wealth and welfare of different households differently, typically making net house buyers the winners and net house sellers the losers during a housing boom. In contrast, financial innovation that relaxes collateral constraints turn out to have small effects on house prices. Similarly, Sommer et al. (2013) develop a quantitative general equilibrium model with housing and financial constraints and argue that a relaxation of financial constraints has only small effects on house prices, while movements in interest rates have large effects.

In related work, Favilukis et al. (2016) also develop a quantitative general equilibrium model with housing and collateral constraints to explore what drives fluctuations in house prices to rent ratio, but draw very different conclusions. Relative to previous quantitative papers, this model has two new features: aggregate business cycle risk and bequest heterogeneity to generate a realistic wealth distribution. In contrast to the previous literature, the authors find that a relaxation of collateral requirements can generate a large housing boom, while lower interest rates, due to an inflow of foreign capital in the domestic bond market, cannot. In particular, they show that the mechanism through which financial liberalization can generate a house price boom is by reducing the housing risk premium. In a similar spirit, Kermani (2016) propose a model to emphasize the importance of financial liberalization and its reversal to explain the housing boom and bust. He et al. (2015) also propose a model where housing collateralizes loans and house price boom and bust can be generated by financial innovation because the liquidity premium on housing is nonmonotone in the loan-to-equity ratio. In their paper, even without a change in fundamentals, house prices can be cyclical because of self-fulfilling beliefs. In a related paper, Huo and Ríos-Rull (2014) propose a model with heterogenous households, housing and credit constraints, and also show that financial shocks can generate large drops in housing prices.

In a more recent paper, Justiniano et al. (2014) ask what is the best way of formalizing the "credit easing" shock behind the recent US housing boom. Their objective is to model the shock in a way to be able to match a number of stylized facts about the housing and mortgage markets: not only the rise in house prices and households' debt, but also the fairly stable loan-to-value ratio and the decline in mortgage rates. In particular, they distinguish between a loosening of "lending constraints," ie, an increase in the availability of funds to be borrowed for the purpose of home mortgages, and a loosening of "borrowing constraints," ie, the lessening of collateral requirements. They argue that a loosening of the collateral requirements alone cannot explain the recent housing boom in the United States, but there must have been an expansion in the credit supply.

3.3 The Effect of Credit Constraints and Housing Prices on Macro

The impact of changes in credit conditions in the housing market on the overall economy and on economic policy is obviously an important question and the focus on a significant portion of the literature. Iacoviello (2005) has become a work horse model in this literature, embedding nominal households' debt and collateral constraints tied to real estate values, as in Kiyotaki and Moore (1997), into a new Keynesian model. The paper shows that demand shocks move housing and consumer prices in the same direction and hence are amplified. When demand rises due to some exogenous shock, consumer and asset prices increase. The rise in asset prices increases the borrowing capacity of the debtors, allowing them to spend and invest more. The rise in consumer prices reduces the real value of their outstanding debt obligations, positively affecting their net worth. Given that borrowers have a higher propensity to spend than lenders, the net effect on demand is positive. Thus the demand shock is amplified. Guerrieri and Iacoviello (2014) emphasize that collateral constraints drive an asymmetry in the relationship between house prices and economic activity. Brzoza-Brzezina et al. (2014) examine a DSGE model with housing and financial intermediaries. They evaluate the impact of having multiperiod vs one-period contracts on monetary and macroprudential policy, and the role of fixed-rate vs variable-rate mortgages. Garriga et al. (2016) also explore the interaction among long-term mortgages, nominal contracts and monetary policy in a similar general equilibrium model. Benes et al. (2014a) offers a richer structure yet for studying the interplay between the housing market and economic performance and its implications for macroprudential policies. Applications and extensions are in Benes et al. (2014b) and Clancy and Merola (2015).

Corbae and Quintin (2014) are interested in assessing the role of high-leverage mortgages to explain the foreclosure crisis. They propose a model with heterogenous agents who can choose between a mortgage contract with a 20% down payment and one with no down payment and can choose to default. The model show that the increase in number of high-leverage loans can explain more than 60% of the increase in foreclosure rates.

There is another strand of literature that focuses on macroeconomic models with a housing sector and collateral constraints, but takes house prices as given. Among them, Campbell and Hercowitz (2006) explores the macroeconomic consequences of the relaxation of households' collateral constraints that followed the US financial reforms in the early 1980s. They propose a general equilibrium model with heterogenous households who have access to loan contracts that require a down payment and rapid amortization. House prices are taken as given.[b] Reducing the down-payment rate or extending the term of the loans reduces macroeconomic volatility. In particular, they show that the reforms of the early 1980s can explain a large fraction of the volatility decline in hours worked, output, households debt and durables' consumption. In a similar spirit, Iacoviello and Pavan (2013) embed housing in a life-cycle general equilibrium business cycle model where households face collateral constraints. They show that higher income risk and lower down payments can explain the reduced volatility of housing investment, the procyclicality of debt and part of the reduced output volatility during the Great Moderation. They also show that looser credit conditions can make housing and debt more stable in response to small shocks but more fragile in response to large negative shock, as it happened in the Great Recession.

Since the recent boom and bust in housing prices and subsequent long recession, there has been a new wave of macro models that take households' leveraging and deleveraging as the fundamental shock affecting economic activity, even without explicitly modeling the housing market. In his 2011 Presidential Address, Hall (2011) emphasized that the "long slump" that recently hit the United States was driven by a severe decline in aggregate demand, which he attribute to the large deleveraging wave that on the onset of the 2007–08 financial crisis followed a large buildup of consumer debt at the beginning of 2000. On the empirical side, Mian and Sufi (2014) use US zip code data to argue that demand shocks were the main source of the employment decline in the recent recession. In the same spirit, there has been a growing body of work that considers a credit crunch as the fundamental shock of the economy and explores how the subsequent deleveraging affects the overall economy, and the housing market in particular. Together with

[b] In their work, house prices are constant, as in the early literature that included housing in one-sector real business cycle models in the form of capital used for home production, following the seminal papers by Benhabib et al. (1991) and Greenwood and Hercowitz (1991). More recently, Fisher (2007) extends these models by making household capital complimentary to business capital and labor in market production to reconcile the fact that household investment leads nonresidential capital over the business cycle.

Hall (2011), the first papers that develop macro models where the fundamental shock is a credit crunch type of shock (instead for example of a productivity shock) are Eggertsson and Krugman (2012) and Guerrieri and Lorenzoni (2011). Both papers propose an incomplete market model with households facing a borrowing constraint and represent a credit crunch as an unexpected tightening in the borrowing limit. In order to focus on households' gross debt positions, both papers need to introduce some form of households' heterogeneity into the model: Eggertsson and Krugman (2012) use a Keynesian model with two types of agents, borrowers and lenders, while Guerrieri and Lorenzoni (2011) use a Bewley type of model with uninsurable idiosyncratic income risk, so that households delever not only when they hit the borrowing limit, but also for precautionary reason when they are close enough to it. Both paper show that a credit crunch type of shock can have large (also persistent in Guerrieri and Lorenzoni, 2011) effects on the real economy, especially in the presence of sizeable nominal rigidities.

There has been a growing group of papers working on related incomplete market models with heterogenous households and focusing on a similar "credit tightening" shock. On a more quantitative side, Justiniano et al. (2015) and Del Negro et al. (2011) quantify the real effects of this type of shock, using different general equilibrium models and reaching different conclusions. On the one hand, Justiniano et al. (2015) builds on Iacoviello (2005) and Campbell and Hercowitz (2006) and propose a model with two types of households who can borrow using their house as collateral. They show that the leveraging and deleveraging cycle recently experienced by the United States did not have significant real effects. On the other hand, Del Negro et al. (2011) introduce liquidity frictions in an otherwise standard DSGE model and show that the effects of a liquidity shock can be large.

There are number of papers exploring the aggregate effects of a similar shock, focusing on transmission mechanisms that do not rely on nominal rigidities. Huo and Ríos-Rull (2014) study an incomplete market economy where heterogeneous households face a borrowing constraint and the fundamental shock is a tightening in the borrowing limit. The new ingredient in the model that makes the financial shock having real effects is the introduction of search frictions in some consumption markets.[c] That is, households need to engage in costly search to purchase some type of goods and hence, when the borrowing constraint tightens and households want to save more, they will also search less intensively. This will reduce demand and hence generate a recession. Moreover, there is an amplifying effect coming from the fact that consumption tilts more towards the wealthier households who are farther away from the constraint and who are the ones who exert less search effort. Another related paper is Kehoe et al. (2014) who propose a search and matching model a la Diamond–Mortensen–Pissarides with upward-sloping wage profiles

[c] The introduction of search frictions in consumption markets builds on Bai et al. Similar frictions are key in the transmission of financial shocks in Huo and Ríos-Rull (2013) who focus on a small open economy.

and risk-averse consumers who face borrowing constraints. In their model, a tightening in the borrowing limit raise workers' and firm' discount rates, hence reducing vacancy creation and employment, with a similar mechanism as in Hall (2014). This effect is amplified by the presence of on-the-job human capital accumulation and workers' debt constraints. Macera (2015) studies a model with both heterogenous households and heterogenous producers and explores the aggregate effects of a tightening in the borrowing capacity of both types of agents.

Another important related paper is Midrigan and Philippon (2011) who study a cash-in-advance economy with housing, where transactions can be conducted not only with money but also with home equity borrowing. In their economy, there is a continuum of islands that are subject to different collateral constraints. The authors parameterize the model to match the empirical evidence from Mian and Sufi (2011) at the MSA level. When house prices decline in one island, the cash-in-advance constraint tightens reducing aggregate demand in that island. This leads to a recession, thanks to nominal wage rigidities and frictions for the reallocation of labor from different sectors, which prevent households to work harder or to move to tradable sectors. The authors also consider an extension of the model with two types of households, patient and impatient, so that patient households lend to impatient households who can use housing as collateral. The authors distinguish between "liquidity shocks," ie, a tightening in the cash-in-advance constraint which affect all households, and "credit shocks," ie, a tightening the borrowing constraint which affect only impatient households, and show that liquidity shocks are very powerful. The distinction between the two types of constraints is useful to capture the empirical evidence in Johnson et al. (2006), Parker et al. (2013), and Kaplan and Violante (2014) showing that there is a large fraction of wealthy households who are liquidity constrined. In many macro models, as the ones described earlier, there is only one collateral constraint that typically captures both types of shocks.

Incomplete markets models have also been used to emphasize the effect of house prices on consumption, which is sizable according to Mian et al. (2013). There is a large empirical literature that has tried to estimate the effect of house price changes on consumption, using different data samples and different identification strategy, such as Campbell and Cocco (2007), Attanasio et al. (2009), Carroll et al. (2011), Case et al. (2013), and Ströbel and Vavra (2015) (see Iacoviello, 2012 for a more comprehensive survey on this topic). A standard permanent income hypothesis model typically delivers small consumption responses to house prices, as house prices affect households' wealth but also households' implicit rental rates. Berger et al. (2015) show that a simple incomplete market model with heterogenous agents, housing and collateral constraints, can deliver sizable consumption elasticity to house prices consistent with the empirical evidence. They show that the size of such an elasticity is determined by the correlation of marginal propensity to consume out of temporary income shocks and housing values, by deriving a simple sufficient-statistic formula for the individual elasticity. It follows that more levered economy are typically more responsive. They also analyze a boom–bust episode in the house prices similar to the one recently experience by the United States and show that a shock to expected house price appreciation can generate a large boom and bust in consumption and in residential investment at the same time. Kaplan et al. (2015) use a general equilibrium incomplete markets model with heterogeneous agents also to look at the recent boom and bust in house prices and consumption. They allow for different types of shocks: productivity shocks, taste shocks, shocks to the credit markets, and shocks to beliefs about future price appreciation. They show that this last type of shock is the most important to explain the movements in house prices, while shocks to credit conditions are important to explain homeownership, leverage and foreclosure.

Finally, there is another strand of literature that is more interested in understanding fluctuations in residential investment. Most of this literature takes house prices as exogenous. One of the seminal papers in this area is Davis and Heathcote (2005) who actually feature both endogenous housing investment and endogenous house prices. They build a neoclassical multisector stochastic growth model where one sector produces residential structures that, together with land, are used to produce houses. The model does not feature credit constraints, but already capture many facts about dynamics of residential investment. Iacoviello and Neri (2010) extend the multisector structure of Davis and Heathcote (2005) by adding, in particular, nominal rigidities and borrowing constraints. They show that demand shocks, such as housing preference shocks, are important in accounting for fluctuations in house prices. In a more recent paper, Rognlie et al. (2015) propose a model where a house price boom generates overbuilding of residential capital that would require a reallocation of resources among sectors. The authors use this model to think about the Great Recession and argue that, in the presence of a liquidity trap, this "investment hangover" can generate a recession. They show that their model is consistent with an asymmetric recovery where the residential sector has been left behind. In a related paper, Boldrin et al. (2001) use input–output tables to recover the linkages between the construction sector and the other sectors of the economy and evaluating the contribution of the construction sector to the Great Recession. This review has not included the large literature, examining the housing market in the absence of financial frictions. For example, Magnus (2011) has argued that search frictions may well be key to understanding many of the housing market phenomena such as liquidity, prices and vacancies.

4. A SIMPLE MODEL OF CATASTROPHES

In this section, we focus on the boom and bust in the credit cycle, abstracting from the dynamics of house prices. The main idea is that, if credit markets are affected by private information about the quality of the borrowers, a credit cycle can arise endogenously simply because of an increase in credit availability, which can be interpreted as a "saving glut." The idea is that when banks have easier access to credit, for example because the interest rate they face is lower, at first they will offer cheaper loans and increase their lending. However, due to the presence of adverse selection, when borrowing is cheaper worse borrowers will take loans and this can endogenously generate a crash of the credit market.

This idea is inspired by Boissay et al. (2016), but the model that we present here is quite different. The main mechanism in our model is based on adverse selection in the mortgage market, while Boissay et al. (2016) relies on a model of the interbank market affected by moral hazard. In their paper, banks are heterogeneous for their intermediation efficiency and their quality is private information. At the same time, borrowing banks can divert some of the funds to low return assets that cannot be recovered by the lending banks. This mechanism also generates endogenous credit cycles as a result of an increase in credit availability: as interest rates go down, the more efficient banks increase their activity, generating a boom of the banking sector, but as interest rates keep decreasing, worse banks have a higher incentive to divert their funds, increasing counterparty risk and possibly generating an interbank market freeze. Moreover, Boissay et al. (2016) embed their basic interbank market model in a standard DSGE model. Instead, we reduce the dynamics to 2 periods only and leave richer dynamic settings to future research.

4.1 Model

There are two periods $t=1$, 2. The economy is populated by a continuum of three types of agents: households, lenders, and banks. Lenders and banks are homogenous, while households are heterogenous.

Households enjoy utility $u(c, h)$ in period 2, where c is consumption of a nondurable good and h is housing consumption. For simplicity, let us assume that utility is linear, that is,

$$u(c,h)=c+\gamma h.$$

Houses come in fixed size equal to $\bar{h}$, so that $h\in\{0,\bar{h}\}$, and their price is fixed to 1. Households have no endowment in period 1 but receive an income draw y in period 2. They have to decide whether to buy a house or not in period 1, so if they decide to buy a house they have to borrow the full amount.

Households are heterogenous with respect to their income process. Let $\nu \in [0, 1]$ be the household's type. Assume that ν is distributed according to some distribution function $G(\nu)$ and affects the distribution of the household's income $F_\nu(y)$. Throughout, we shall assume

Assumption 1 $F_{\nu_B}(y)$ first order stochastic dominates $F_{\nu_A}(y)$ whenever $\nu_B > \nu_A$.

Thus, higher household types have a "better" income distribution.

To buy a house in period 1, a household has to borrow 1 unit of funds from the banks at some mortgage "price" p and promise to repay $1/p$ in period 2. Let us note that the label "price" (and notation p) might be a bit confusing. It does not refer to the price of the house (which remains fixed at 1), but is period-1 price for one unit of period-2 resources.

Alternatively, one may wish to think of p as the "payment" the household receives in period $t = 1$ for each unit of promised repayment at date $t = 2$. We shall continue to refer to it as the mortgage price.

At the beginning of period 2, the household's income y is realized and then the household decides whether to repay its debt or not. If it does default on its debt it does not pay anything back to the lender, but it suffers a penalty $\delta > 0$.[d] This implies that households with higher ν are better potential borrowers, in the sense that they have a lower probability of bad income realizations. Let us denote by $\chi \in\{0, 1\}$ the repayment decision, with $\chi = 1$ denoting repayment.

The household's type ν is private information of the household. However, at the beginning of period 1, households have the option to verify their type at a utility cost $\kappa > 0$.[e] Let $v(\nu) \in \{0, 1\}$ be the decision of verifying their type ($v = 1$) or not ($v = 0$). If a household verifies its type, banks can make the lending terms type-contingent, so that the mortgage price p is going to be equal to $\widetilde{p}(\nu)$. If instead a household does not verify its type, banks do not know the type of the borrower and will offer a pooling mortgage price p^P. Note that we assume that banks are restricted to offer only one mortgage price to all that have not verified. It would be interesting to extend the model allowing banks to offer more general contracts.

To sum up, households have three options: (1) do not verify their type and borrow accepting a pooling contract, that is, $h(\nu) = 1$ and $v(\nu) = 0$ and borrow at the pooling mortgage price $p = p^P$; (2) verify their type and access type-contingent contracts, that is, $h(\nu) = v(\nu) = 1$ and borrow at the type-contingent mortgage price $p = \widetilde{p}(\nu)$; and (3) do not borrow at all, that is, $h(\nu) = v(\nu) = 0$.

Let us proceed backward and consider the repayment decision, conditional on borrowing in the first period, that is, on $h(\nu) = 1$. Recall that the household suffers a penalty δ, if it defaults, and that the lender does not get anything back. Then, a household with realized income y who borrows at the mortgage price p would like to repay if

$$y - 1/p + \gamma\bar{h} \geq y - \delta.$$

We assume throughout that δ is large enough so that in equilibrium any household would like to pay back its debt if it can.[f] However, it may not be able to repay because its realized

[d] The assumption that a defaulting household does not pay anything back to the lender and only suffers a penalty is stark, but simplifies the analysis. The idea is that the household could run away, at a cost which is summarized by δ. One could relax that and assume that the lender can seize only part of the income of the borrower, as it is reasonable to assume that part of it must be lost in legal fees.

[e] The household decides whether to verify its type or not before banks make their offers. Also, we assume that the verification costs is in terms of utility, because for simplicity we assume that the households start period 1 with no endowment. However, it would be easy to extend the model to make it a monetary cost.

[f] We will show below that this is a necessary condition to have a nonempty set of borrowers at the pooling mortgage price.

income is not high enough, so that $\chi(y, p) = 1$ iff $y \geq 1/p$. Let $\pi(\nu, p)$ be the ex-ante repayment probability of a household of type ν who borrows at the mortgage price p, that is,

$$\pi(\nu,p) = E[\chi(y,p) = 1|\nu,p] = 1 - F_\nu\left(\frac{1}{p}\right)$$

Then we can show the following proposition.

Proposition 2 *The repayment probability* $\pi(\nu$, p*) is increasing in* ν *and increasing in* p.

Proof First, $\pi(\nu, p)$ is increasing in ν, since F_ν are ordered by first-order stochastic dominance. Second, $\pi(\nu, p)$ is increasing in p, since $1 - F_\nu(y)$ is decreasing in y for any ν. □

This means that a household with higher type has a higher repayment probability, for any given mortgage price p.

Let us now focus on the lending market. Let us assume that the banks can borrow from the lenders at some rate R, which is exogenously given.[g] Also, they trade the loans, which we will refer to as assets from now on, on the secondary market.[h] Each asset is characterized by the type of the associated borrower ν and has a different repayment probability $\pi(\nu, p) \in [0, 1]$. This implies that the pooling mortgage price is determined by the no-arbitrage condition

$$p^P = \frac{E\left[1 - F_\nu\left(\frac{1}{p^P}\right)|\nu \in \mathcal{S}\right]}{R}, \tag{31}$$

where $\mathcal{S} \equiv \{\nu|v(\nu) = 0\}$ is the set of households who decide not to verify their type. Likewise, the type-contingent mortgage price is determined by the no-arbitrage equation

$$\tilde{p}(\nu) = \frac{1 - F_\nu\left(\frac{1}{\tilde{p}(\nu)}\right)}{R}. \tag{32}$$

In principle, there may be none, one or several solutions to this equation. Borrowing the logic in Mankiw (1986), we assume that the highest of these prevails in equilibrium: at a lower mortgage price and thus higher promised return to all other banks, a bank could profitably deviate by offering a higher mortgage price and a better deal to the household, under mild conditions. Define ν_L as the lowest type, beyond which a type-contingent mortgage price exists for some types,

[g] The interest rate R can be interpreted as the rate at which lenders can borrow in the international market.

[h] One can potentially generalize the model to create MBS that pool different loans and a similar mechanism would go through as long as there is a constraint on the measure of types that can be pooled together.

$$\nu_L = \inf_{\nu}\{\nu \mid \text{ there is a solution to Eq. (32)}\}$$

We can then show that a type-contingent mortgage price exists for all types better than ν_L.

Proposition 3 *There exist a type-contingent mortgage price $\tilde{p}(\nu)$ for any $\nu > \nu_{\mathrm{L}}$.*

Proof Consider some ν. Let $\nu' \in [\nu_L, \nu]$ be such that there is a solution $\tilde{p}(\nu')$ to Eq. (32). Define

$$\mathcal{P}(\nu) = \{p \geq \tilde{p}(\nu') \mid p \leq \frac{1 - F_\nu\left(\frac{1}{p}\right)}{R}\}$$

Since F_ν second-order stochastically dominates $F_{\nu'}$, $\tilde{p}(\nu') \in \mathcal{P}(\nu)$ and $\mathcal{P}(\nu)$ is therefore nonempty. Let $\bar{p} = \sup\mathcal{P}(\nu)$, the supremum of $\mathcal{P}(\nu)$. Note that $\bar{p} \leq 1/R < \infty$. Consider an increasing sequence $p_j \to \bar{p}$ with $p_j \in \mathcal{P}(\nu)$. Calculate that

$$\bar{p} = \lim_{j\to\infty} p_j \leq \lim_{j\to\infty} \frac{1 - F_\nu\left(\frac{1}{p_j}\right)}{R} \leq \frac{1 - F_\nu\left(\frac{1}{\bar{p}}\right)}{R}$$

and thus $\bar{p} \in \mathcal{P}$. This shows that $\bar{p}$ is a maximum and that there is therefore at least one solution to Eq. (32). □

Obviously, households who verify their type will be able to borrow at terms that are more favourable the better their type is. That is, we can prove the following proposition.

Proposition 4 *For $\nu > \nu_{\mathrm{L}}$, the type-contingent mortgage price $\tilde{p}(\nu)$ is increasing in ν and decreasing in* R.

Proof First, let us make a change of variable and define $\tilde{R}(\nu) \equiv 1/\tilde{p}(\nu)$. Rewrite Eq. (32) as

$$1 - \frac{R}{x} = F_\nu(x). \tag{33}$$

which we seek to solve for $\tilde{R}(\nu) = x$, for a given ν. Assume then that there is at least one solution for (32), ie, that the curves defined by the left-hand side and right-hand side of that equation cross at least once. As assumed above per the logic in Mankiw (1986), pick the lowest solution to (33) or, equivalently, the highest of the solutions for (32), if there are several. This pins down a unique $\tilde{R}(\nu)$ and unique $\tilde{p}(\nu)$ for each ν. Fix some ν and its solution $\tilde{R}(\nu) = x$. Note that the left-hand side of (33) diverges to $-\infty$, as $x \to 0$, while the right-hand side converges to a nonnegative number. Thus, at the lowest solution and as a function of x, the right-hand side of (33) approaches and then either crosses or touches the left-hand side from above, as x approaches the solution from below. Per the definition of second-order stochastic dominance, the right-hand side shifts to the right, as ν is increased. Therefore a solution continues to exist for higher ν and they are to the left of the solution fixed at the beginning of this argument. As ν is decreased, the right-hand side function shifts to the left. By the similar logic, the solution either moves to the right, when the intersection between the two sides moves locally, or will jump to a solution at a higher value, if the current intersection disappears or a solution will cease to exist altogether. In sum, if a solution $\tilde{R}(\nu)$ exists, it is decreasing in ν. Equivalently, if a solution $p(\nu)$ exists, it is increasing in ν. Likewise, consider now an decrease in R. This shifts the left-hand side of (33) upward, a solution will continue to exist and will be lower than the previously fixed solution. If R increases, the current intersection may move locally or disappear: in either case, if a solution continues to exist, it will be higher than the previously fixed solution. This shows that $\tilde{R}(\nu)$ is increasing and $p(\nu)$ therefore decreasing, as a function of R. □

Consider now the household's problem. Define

$$U^B(\nu,p)=\int_{\frac{1}{p}}^{\infty}(y-\frac{1}{p}+y\bar{h})dF_\nu(y)+\int_0^{\frac{1}{p}}(y-\delta)dF_\nu(y), \tag{34}$$

which is the expected utility of a household of type ν who decides to buy a house at the mortgage price p and does not verify its type ($h(\nu)=1$ and $v(\nu)=0$). For $\nu>\nu_L$, define

$$U^V(\nu)=U^B(\nu,\tilde{p}(\nu))-\kappa, \tag{35}$$

which is the expected utility of a household who decides to borrow at mortgage price $p=\tilde{p}(\nu)$ and verify its type ($h(\nu)=v(\nu)=1$). For $\nu<\nu_L$, no such type-contingent mortgage price exists: thus define

$$U^V(\nu)=-\infty \tag{36}$$

in that case. For $\nu=\nu_L$, use (35), if there is a solution to (32), and (36), if not. Finally, define

$$U^N(\nu)=\int ydF_\nu(y)=E[y\,|\,\nu], \tag{37}$$

which is the expected utility of a household of type ν who decides not to buy a house ($h(\nu)=v(\nu)=0$), and equal to the expected income, given our assumption of linear utility. For a given pooling mortgage price p^P, the utility of the household of type ν and its maximization problem is now

$$\bar{U}(\nu,p^P)=\max\left\{U^B(\nu,p^P),U^V(\nu),U^N(\nu)\right\},$$

To make more progress, we need an assumption regarding the income uncertainty as expressed by F_ν.

Assumption 2 There is some $x^*\in\mathbb{R}$ so that $F_\nu(x)$ has nondecreasing slopes above x^*: for all x_1 and x_2 with $x^*\leq x_1\leq x_2$ and all ν_A and ν_B with $\nu_L\leq\nu_A\leq\nu_B$, we have

$$F_{\nu_A}(x_2)-F_{\nu_A}(x_1)\leq F_{\nu_B}(x_2)-F_{\nu_B}(x_1) \tag{38}$$

The assumption is trivially satisfied at some x^*, where $F_1(x^*) = 1$, if such a value x^* exists, ie, if the income distribution is bounded. Obviously then, this assumption is only useful, if x^* is fairly small and smaller than some upper bound on income. Indeed, it will be particularly convenient to assume that $x^* = R$, the safe return.

The following lemma is a bit technical, and useful for an intermediate step in the proof of the next proposition.

Lemma 1

1. *Define*

$$Z(\nu,p) \equiv \left(1 - F_\nu\left(\frac{1}{p}\right)\right)\left(\gamma\bar{h} - \frac{1}{p} + \delta\right). \tag{39}$$

and suppose that $\gamma\bar{h} - (1/p) + \delta > 0$. *Then* Z($\nu$, p) *is increasing in both* ν *and* p.

2. *For* $\nu > \nu_L$, *define*

$$g(\nu,p) = Z(\nu, \tilde{p}\,(\nu)) - Z(\nu,p) \tag{40}$$

For $\nu < \nu_L$, *define* $g(\nu,p) = -\infty$. *For* $\nu = \nu_L$, *define* $g(\nu_L, p)$ *as in Eq.* (40), *if there is a solution to Eq.* (32) *and* $g(\nu_L,p) = -\infty$ *otherwise. Suppose that* $p \leq \tilde{p}\,(\nu) \leq 1/x^*$ *for all* $\nu > \nu_L$ *and that* $\gamma\bar{h} - (1/p) + \delta > 0$. *Impose the Assumption 2 of nondecreasing slopes. Then g is increasing in* ν *and decreasing in p.*

Proof It is easy to see that $Z(\nu, p)$ is increasing in both ν and p. It follows that $g(\nu, p)$ is decreasing in p. It remains to show that g is increasing in ν for $\nu \geq \nu_L$. Suppose that $\nu_A \leq \nu_B$. We need to show that

$$g(\nu_A,p) \leq g(\nu_B,p) \tag{41}$$

Excluding the trivial case of $\nu_B = \nu_L$ with $g(\nu_L,p) = -\infty$, calculate

$$\begin{aligned} g(\nu_B,p) - g(\nu_A,p) &= Z(\nu_B, \tilde{p}(\nu_B)) - Z(\nu_A, \tilde{p}(\nu_A)) - (Z(\nu_B,p) - Z(\nu_A,p)) \\ &\geq Z(\nu_B, \tilde{p}(\nu_A)) - Z(\nu_A, \tilde{p}(\nu_A)) - (Z(\nu_B,p) - Z(\nu_A,p)) \end{aligned}$$

where we have exploited that $\tilde{p}(\nu_A) \leq \tilde{p}(\nu_B)$ per Proposition 4, and that $Z(\nu, p)$ is increasing in p. Define $x_1 = 1/\tilde{p}(\nu_A)$ and $x_2 = 1/p$ and note that $x^* \leq x_1 \leq x_2$. With that, rewrite the right-hand side of the last equation as

$$\begin{aligned} &g(\nu_B,p) - g(\nu_A,p) \\ &\geq (F_{\nu_A}(x_1) - F_{\nu_B}(x_1))(\gamma\bar{h} - x_1 + \delta) - (F_{\nu_A}(x_2) - F_{\nu_B}(x_2))(\gamma\bar{h} - x_2 + \delta) \\ &= ((F_{\nu_A}(x_1) - F_{\nu_B}(x_1)) - (F_{\nu_A}(x_2) - F_{\nu_B}(x_2)))(\gamma\bar{h} - x_1 + \delta) \\ &\ + (F_{\nu_A}(x_2) - F_{\nu_B}(x_2))(x_2 - x_1) \\ &\geq 0 \end{aligned}$$

with Assumption 2. □

The following proposition shows that the households' optimal behavior can be characterized by two cutoffs values, such that, households with low types do not buy a house,

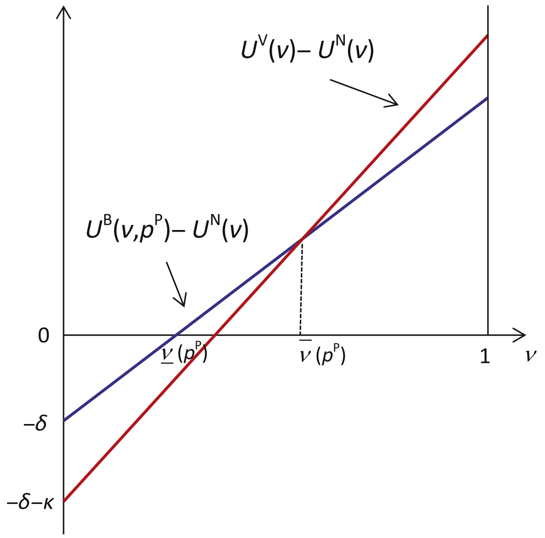


Fig. 8 Households' problem (curves are linear just for illustration).

households with high types buy a house and may verify their type, and households in the middle range, buy a house but choose to borrow at the pooling mortgage price. The logic is illustrated in Fig. 8.

Proposition 5 *Impose the Assumption 2 of nondecreasing slopes. Assume that $x^* = R$, the safe return. Then, there exists a value $\underline{p}$ so that no household buys a house at a mortgage price $p < \underline{p}$, whereas for all $p^P > \underline{p}$ there are two cutoffs $\underline{\nu}(p^P) \leq \overline{\nu}(p^P)$ such that*

1. *$h(\nu) = 0$, ie, no house purchase, if $\nu < \underline{\nu}(p^P)$.*

2. *$h(\nu) = 1$, $v(\nu) = 0$, ie, house purchase without verification at the pooling mortgage price p^P, if $\underline{\nu}(p^P) < \nu < \overline{\nu}(p^P)$.*

3. *$h(\nu) = 1$, $v(\nu) = 1$, ie, house purchase with verification at the type-contingent mortgage price $\tilde{p}\,(\nu)$, if $\nu > \overline{\nu}(p^P)$.*

4. *$\mathcal{S} = [\underline{\nu},\overline{\nu}]$ or $\mathcal{S} = (\underline{\nu},\overline{\nu}]$ or $\mathcal{S} = [\underline{\nu},\overline{\nu})$ or $\mathcal{S} = (\underline{\nu},\overline{\nu})$, for $\underline{\nu} = \underline{\nu}(p^P)$ and $\overline{\nu} = \overline{\nu}(p^P)$.*

Moreover, $\underline{\nu}(p^P)$ and $\overline{\nu}(p^P)$ are respectively decreasing and increasing in p^{P}.

Proof Let us rewrite

$$U^B(\nu,p) = E[y\,|\,\nu] + Z(\nu,p) - \delta,$$

$$U^V(\nu) = E[y\,|\,\nu] + Z(\nu,\tilde{p}\,(\nu)) - \delta - \kappa,$$

and

$$U^N(\nu) = E[y\,|\,\nu],$$

where $Z(\nu, p)$ is defined in Eq. (39).

Note that a type ν-household will choose to buy a house at mortgage price p without verification, iff $U^B(\nu, p) \geq U^N(\nu)$, that is, iff

$$Z(\nu,p)\geq\delta. \tag{42}$$

Let $\underline{p}$ be the infimum of all p, so that there exists a ν, satisfying Eq. (42). Therefore, no household of any type will purchase a house at a mortgage price $p<\underline{p}$ without verification and certainly not either, when paying the verification cost.

Consider now any $p>\underline{p}$. Since $Z(\nu, p) > 0$ for some ν, it follows that $\gamma\bar{h}-(1/p)-\delta>0$. Since $Z(\nu, p)$ is increasing in ν per Lemma 1, there is a unique cut-off $\underline{\nu}(p)$ such that such that $U^B(\nu, p)\geq U^N(\nu)$ if $\nu>\underline{\nu}(p)$. If $Z(1, p)>\delta$ and $Z(0, p) < \delta$, the cutoff $\underline{\nu}$ is implicitly defined by the infimum of all $\nu\in[0, 1]$ such that

$$Z(\nu,p)\geq\delta. \tag{43}$$

If $Z(1, p) < \delta$, then $\underline{\nu}(p)=1$ and if $Z(0, p) > \delta$, then $\underline{\nu}(p)=0$. Since $Z(\nu, p)$ is increasing in p, it follows that $\underline{\nu}(p)$ is decreasing (more precisely: nonincreasing) in p.

Note that a type ν-household will choose to verify and buy a house at the type-contingent mortgage price $\tilde{p}(\nu)$ rather than a mortgage pooling price p, iff $U^V(\nu)\geq U^B(\nu, p)$, that is, iff

$$Z(\nu,\tilde{p}(\nu))-\kappa\geq Z(\nu,p) \tag{44}$$

provided a type-contingent mortgage price $\tilde{p}\,(\nu)$ exists, or, equivalently, iff

$$g(\nu,p)\geq\kappa \tag{45}$$

where g is defined in Eq. (40). Let $\overline{\nu}(p)$ be the infimum over all $\nu\in[0, 1]$, for which (45) holds, with the convention that $\overline{\nu}(p)=1$, if no such ν exists. Consider some $\nu_A>\overline{\nu}(p)$ such that (45) holds. Since $Z(\nu, p)$ is increasing in p, it follows that $p\leq\tilde{p}(\nu)$. Eq. (32) implies that $\tilde{p}(\nu)\leq 1/x^*$ for all ν. Let $\nu_B>\nu_A$. Lemma 1 now implies that $g(\nu_B, p)\geq g(\nu_A, p)\geq\kappa$, ie, Eq. (45) also holds at ν_B. This proves that Eq. (45) holds for all $\nu>\overline{\nu}(p)$.

Finally, recall that $g(\nu, p)$ is decreasing in p, per Lemma 1. Therefore, if (45) holds at some ν and p, it continues to hold at some $p'<p$. It follows that $\overline{\nu}(p)\geq\overline{\nu}(p')$, ie, that $\overline{\nu}(p)$ is increasing in p. □

We have been careful to allow for discontinuities in all equations, and expressing solutions as infima or suprema for variables appearing in inequalities. In practice, it may be simpler to proceed with enough continuity and to assume that these equations hold with equality at the limiting points. Furthermore, it may be best to impose that $G(\nu)$ has no mass points. With that and in sum, an equilibrium can be represented by a separating mortgage price schedule $\tilde{p}\,(\nu)$ solving Eq. (32) for $\nu\geq\nu_L$, a pooling mortgage price p^P, which satisfies

$$p^P=\frac{\int_{\underline{\nu}}^{\overline{\nu}} 1-F_\nu\left(\frac{1}{p^P}\right)G(d\nu)}{R}, \tag{46}$$

and two cutoffs $\underline{\nu}$ and $\overline{\nu}$ satisfying the two conditions

$$\left[1-F_{\underline{\nu}}\left(\frac{1}{p^P}\right)\right]\left(\gamma\bar{h}-\frac{1}{p^P}+\delta\right)=\delta, \tag{47}$$

and

$$\left[1-F_{\overline{\nu}}\left(\frac{1}{p^P}\right)\right]\left(\gamma\bar{h}-\frac{1}{p}+\delta\right)=R\,\widetilde{p}(\overline{\nu})(\gamma\bar{h}+\delta)-R-\kappa. \tag{48}$$

We next want to show that multiple equilibria can arise in our model performing some simple numerical exercises.

4.2 Multiple Equilibria

In our model, good households may decide to costly verify their type to signal that they are good and so not to be pooled with bad households. This feature of the model is key to generate multiple equilibria. For some parameters, we can have two equilibria: a good equilibrium where good households do not verify their type, the mortgage price is high and hence it is indeed optimal not to suffer the verification cost; and a bad equilibrium, where good households do verify their type, hence lowering the pooling mortgage price and making it indeed optimal to costly verify their type.

The possibility of multiple equilibria can generate an endogenous credit cycle, driven by a simple increase in credit availability, that is, a decrease in R. As we highlighted earlier, this can be thought as an episode of "saving glut," using Ben Bernanke language: *"I will argue that over the past decade a combination of diverse forces has created a significant increase in the global supply of saving—a global saving glut—which helps to explain both the increase in the U.S. current account deficit and the relatively low level of long-term real interest rates in the world today. The prospect of dramatic increases in the ratio of retirees to workers in a number of major industrial economies is one important reason for the high level of global saving. However, as I will discuss, a particularly interesting aspect of the global saving glut has been a remarkable reversal in the flows of credit to developing and emerging-market economies, a shift that has transformed those economies from borrowers on international capital markets to large net lenders."*

In the next section we will show a numerical example where this is the case. Let us first describe the mechanics behind such an endogenous cycle.

1. Let us imagine that we start in an equilibrium where R is relatively high so that the pool of borrowers is relatively good, that is, $\underline{\nu}$ is large, and all borrowers are pooled together, that is, $\overline{\nu}=1$.
2. Then, assume that R declines, pushing both p and $\widetilde{p}\,(\nu)$ up and hence increasing both U^B and U^V. This implies that $\underline{\nu}$ decreases and more bad households become borrowers. However, let us assume that it is still the case that $\overline{\nu}=1$. The change in the composition of the loans tends to depress mortgage prices.

However, mortgage prices have to go up on net, so the interest rate effect has to dominate.[i]

3. If R decreases further, at some point $\overline{\nu}$ will become smaller than 1 and some borrowers will decide to verify that they are good types, hence worsening the pool of households who borrow at the pooling mortgage price, p^P. There are two possibilities:
   (a) p^P increases, then both U^B and U^V shift further up and both $\underline{\nu}$ and $\overline{\nu}$ decline, dampening the increase in p^P.
   (b) p^P declines, then U^B has to shift down and $\underline{\nu}$ increases. In this case, it must be that the decline in $\overline{\nu}$ is strong enough to more than compensate the pressure upward on p^P played by the increase in R and in $\underline{\nu}$.

   Let us imagine that the second case arises.
4. If R decreases even further, the economy is now stuck in a bad equilibrium with some separation.

The shift from a good equilibrium to a bad equilibrium can be interpreted as a market crash, as mortgage prices suddenly drop or, equivalently, required interest payments on mortgages suddenly increase.

4.3 Some Numerical Examples

In this section, we show some simple numerical examples to illustrate how our model can generate an endogenous credit cycle.

For simplicity, let us assume that the income process follows a binary distribution with $y \in \{0, \overline{y}\}$, where $\overline{y} > R$ is sufficiently high that repayment is guaranteed. Let ν be the probability for the high outcome, that is, $\nu = Pr(y = \overline{y})$. The income distribution F_ν is then given by

$$F_\nu(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - \nu, & \text{if } 0 \leq x < \overline{y} \\ 1, & \text{if } x \geq \overline{y} \end{cases}$$

Let x^* be some small, but positive real number, $0 < x^* < R$. Let $\nu_A < \nu_B$. For all x_1 and x_2 with $x^* \leq x_1 \leq x_2 < \overline{y}$ or with $\overline{y} \leq x_1 \leq x_2$, we have

$$F_{\nu_A}(x_2) - F_{\nu_A}(x_1) = 0 = F_{\nu_B}(x_2) - F_{\nu_B}(x_1) \tag{49}$$

Suppose then that $x^* \leq x_1 < \overline{y} \leq x_2$. Now,

$$F_{\nu_A}(x_2) - F_{\nu_A}(x_1) = \nu_A \leq \nu_B = F_{\nu_B}(x_2) - F_{\nu_B}(x_1) \tag{50}$$

Thus, Assumption 2 is satisfied and Proposition 5 applies.

[i] Imagine, by contradiction that p declines, then $\underline{\nu}$ has to increase, but then p has to increase, generating a contradiction.

Eq. (31) now reduces to

$$\tilde{p}(\nu) = \frac{\nu}{R}.$$

The two cut-off $\underline{\nu}$ and $\overline{\nu}$ are given by

$$\underline{\nu} = \frac{\delta}{\gamma\bar{h} + \delta - p^P - 1}, \tag{51}$$

and

$$\overline{\nu} = (R - \kappa)p^P, \tag{52}$$

where the pooling mortgage price p^P is given by condition (46), which can be rewritten as

$$p^P = \frac{E[\nu | \nu \in [\underline{\nu}, \overline{\nu}]]}{R}.$$

For all the numerical examples, we set $\gamma\bar{h} = 2$ and $\delta = 0.1$. We then experiment with different distributions H for ν.

We start by a baseline example where we assume that ν is uniformly distributed on [0, 1]. In this case, the pooling mortgage price can be solved for in closed form and is equal to

$$p^P = \left(\frac{\delta}{R - \kappa} + 1\right)\frac{1}{\gamma\bar{h} + \delta}.$$

This implies that in this simple benchmark the pooling mortgage price is monotonically decreasing in R, and hence a reduction in R is always going to increase p^P and decrease the two cutoffs $\underline{\nu}$ and $\overline{\nu}$, so there is no possibility of multiple equilibria.

Fig. 9 shows the results for this numerical case. The top panel on the left shows the equilibrium manifold for mortgage prices as a function of the exogenous interest rate R and clearly shows that in this case multiple equilibria never arise. The top panel on the right just shows the distribution for ν. The two panels in the middle illustrate that there is a unique equilibrium for any level of R, showing in particular the case of $R = 1.1$, $R = 1.4$, and $R = 1.7$. Finally the bottom panel on the left shows the two cutoffs, $\underline{\nu}$, in red, and $\overline{\nu}$, in blue, as a function of the pooling mortgage price for a specific R. Finally, the bottom right panel shows the volume of loans offered in equilibrium, again for $R = 1.1$, $R = 1.4$, and $R = 1.7$, and shows that, as expected, it is increasing both in the pooling mortgage price and in R.

We then explore the following two alternative distributions:

1. a mixture of two exponential densities,

$$h(\nu) = \omega\frac{\lambda_1 e^{-\lambda_1 \nu}}{1 - e^{-\lambda_1}} + (1 - \omega)\frac{\lambda_2 e^{-\lambda_2 \nu}}{1 - e^{-\lambda_2}};$$

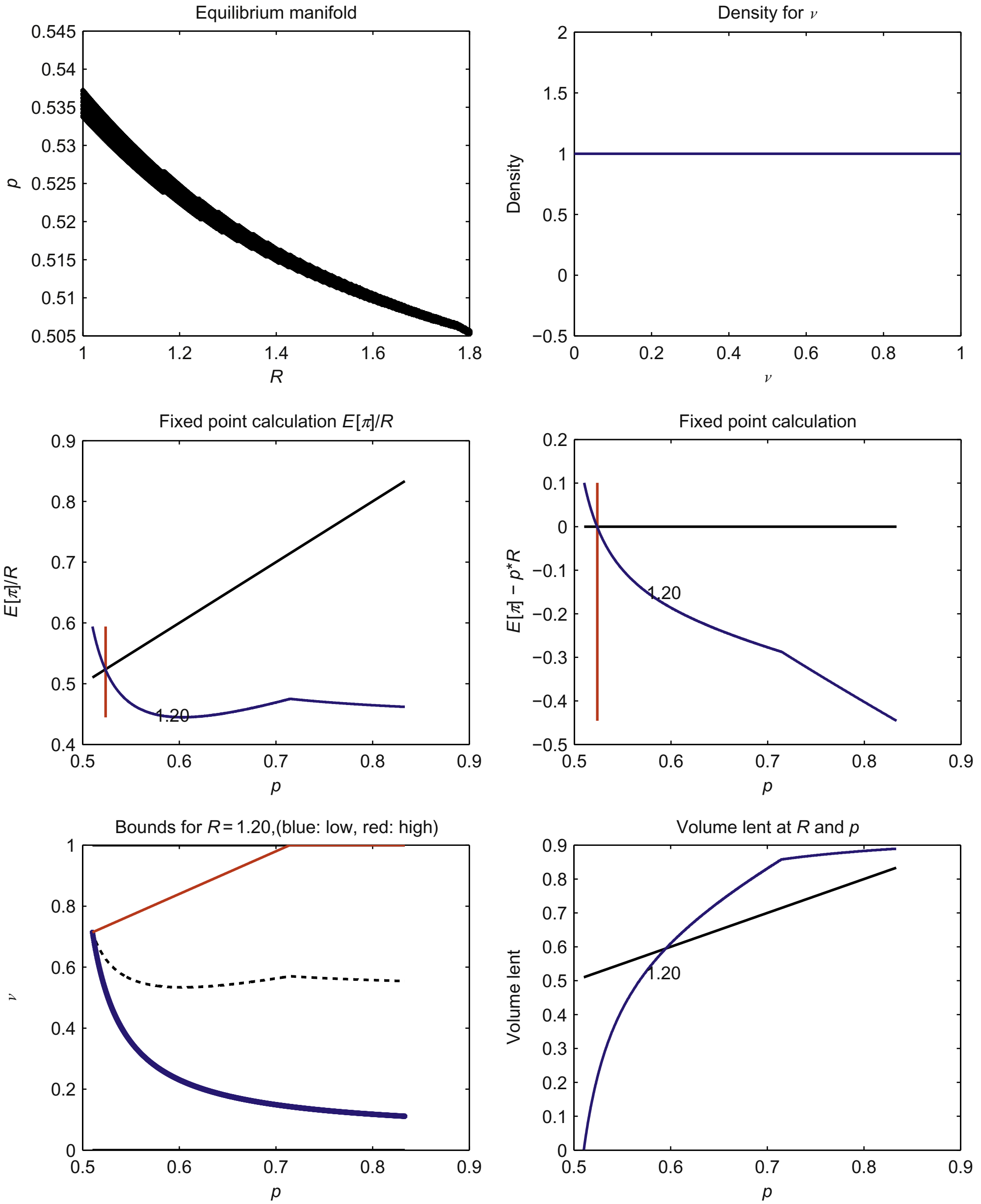


Fig. 9 Uniform distribution for ν on [0, 1]. No multiplicity.

2. a mixture of an exponential density and a truncated normal (where, in terms of the parameterization, we have not normalized the latter to integrate to unity, just the density $h(\nu)$ as a whole),

$$h(\nu) \propto \omega \frac{\lambda e^{-\lambda\nu}}{1-e^{-\lambda}} + (1-\omega)\frac{e^{-(\nu-\nu^e)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma}}.$$

The first example we consider assumes that H is a mixture of an exponential density and a truncated normal, where $\kappa = 0.25$, $\lambda = -20$, $\nu^e = 0.1$, $\sigma = 0.2$, and $\omega = 0.6$. Fig. 10 shows the results for this case. The top panel on the left shows again the equilibrium manifold for mortgage prices as a function of the exogenous interest rate R and shows that for middle-range levels of R, multiple equilibria can arise. The top panel on the right just shows again the distribution for ν. The two panels in the middle show that the number of equilibrium pooling mortgage prices depends on the level of R. For example, we obtain a unique pooling equilibrium when $R = 1.5$, multiple equilibria when $R = 1.4$, and a unique separating equilibrium when $R = 1.3$.[j] In the case of multiple pooling mortgage prices, we have two stable ones and one unstable in the middle. Finally the bottom panel on the left shows again the two cutoffs, $\underline{\nu}$, in red, and $\overline{\nu}$, in blue, as a function of the pooling mortgage price, for the case $R = 1.4$. The bottom right panel shows the volume of loans offered in equilibrium as a function of the pooling mortgage price, for the three different levels of R considered above. This illustrates that if the economy starts at $R = 1.5$ and then R declines, the pooling mortgage price and the loan volume can first increase and then drop as a result of a shift from a good to a bad equilibrium.

For the second case, a mixture of two exponentials, we set the verification cost $\kappa = 0.15$, and the parameters of the H distribution to $\lambda_1 = -20$, $\lambda_2 = 5$, and $\omega = 0.8$. Fig. 11 shows the results for this numerical case. The plots are analogous to the one described earlier. In this case, we show that a unique pooling equilibrium arises when $R = 1.65$, multiple equilibria arise when $R = 1.58$ and a unique separating equilibrium arises when $R = 1.4$. This implies that, also in this case, a decline in R can generate an endogenous boom and bust in the credit markets, represented by an initial increase and a following decline in the pooling mortgage price and in the loan volume. Again, the bust is originated by a shift from a good to a bad equilibrium.

5. RELATED LITERATURE: SENTIMENTS AND BUBBLES

One story about run-ups in house prices and subsequent crashes, which appears to be popular in journalistic descriptions of financial crises, runs as follows: as prices rise,

[j] This is clearly a numerical example, not a calibration. In any case, high interest rates might be justified by the fact that mortgages are long-period contracts.

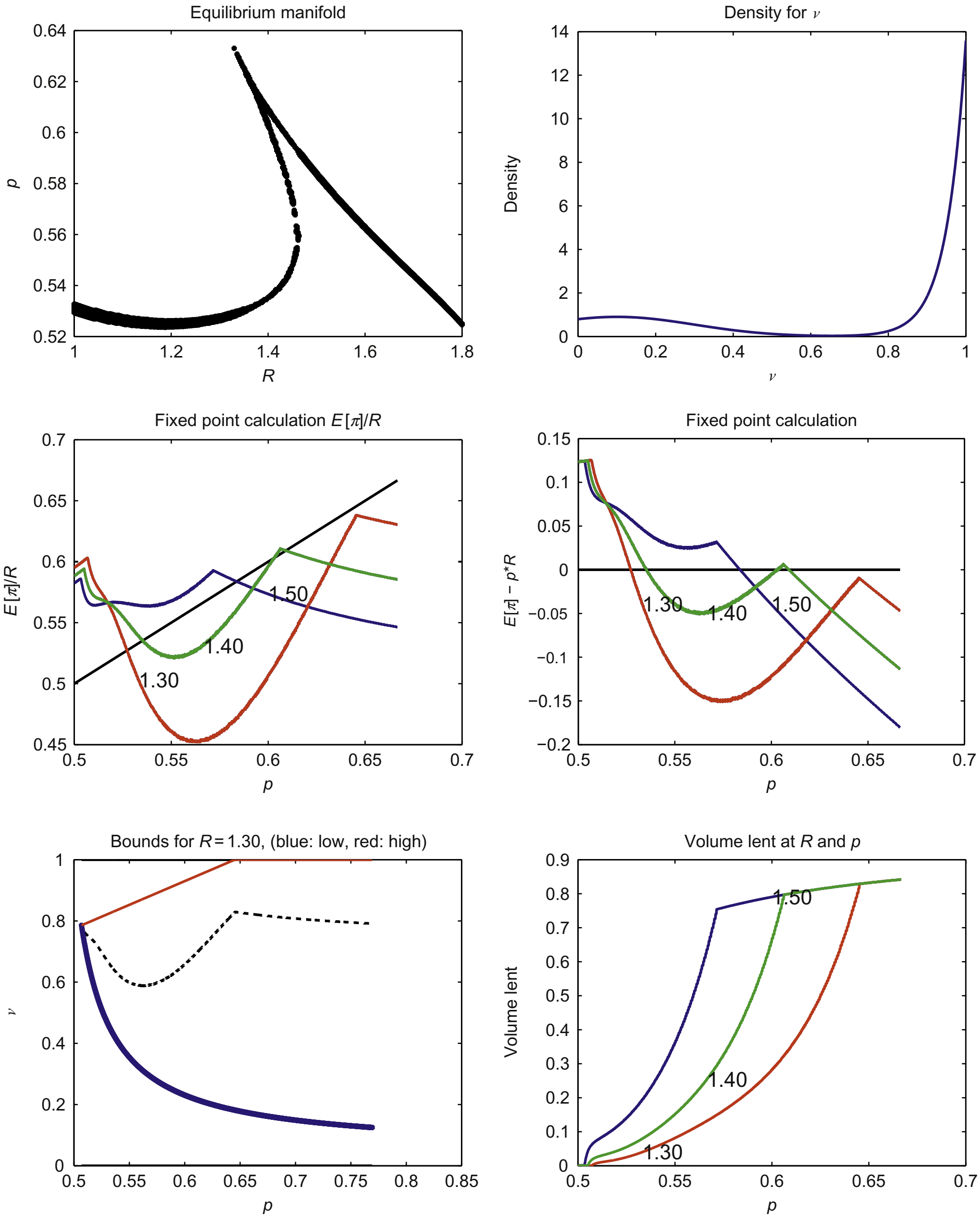
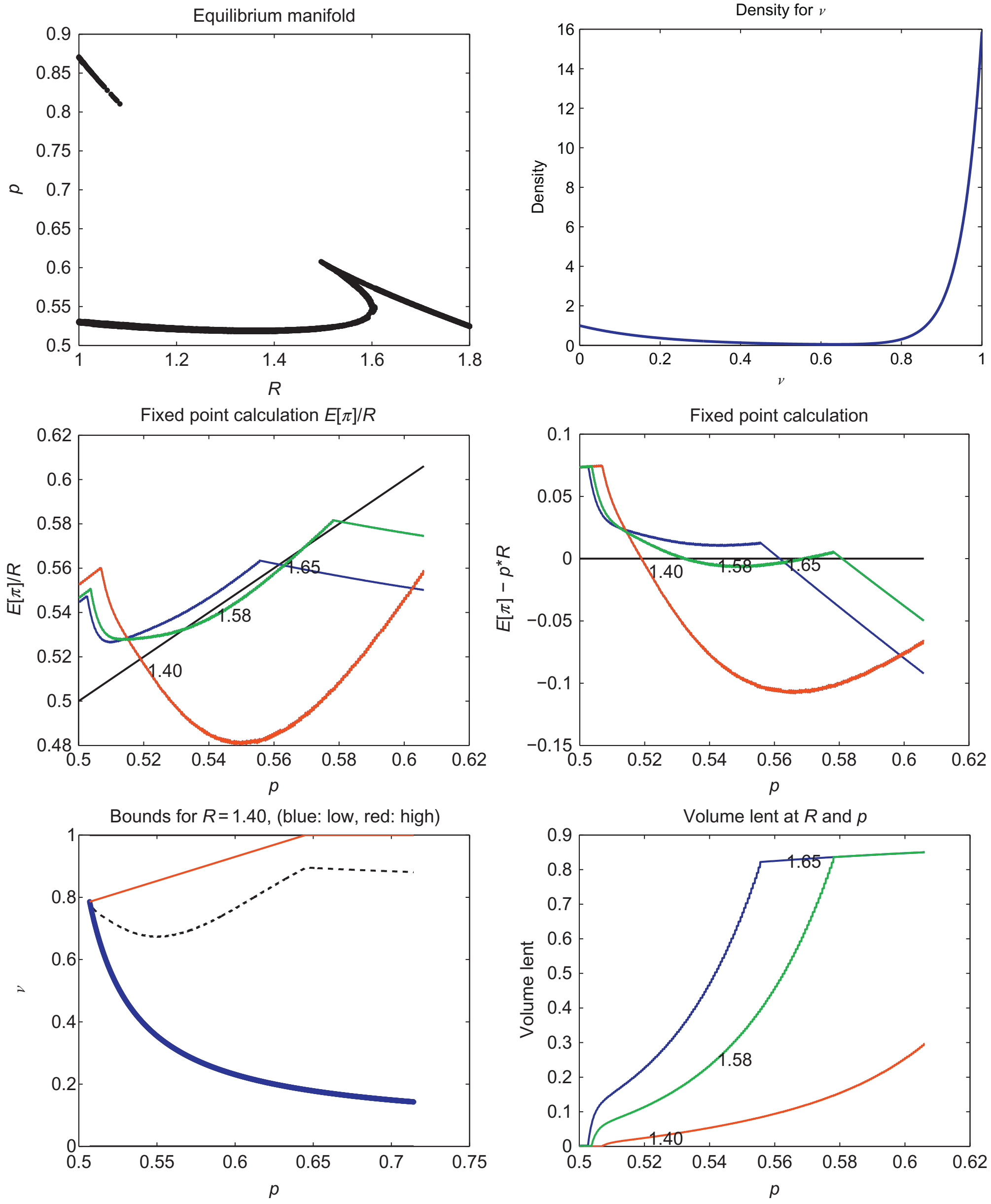


Fig. 10 For the mixture of an exponential density and a truncated normal, where $\gamma\bar{h} = 2$, $\delta = 0.1$, $\kappa = 0.25$, $\lambda = -20$, $v^e = 0.1$, $\sigma = 0.2$, and $\omega = 0.6$, we obtained a unique separating equilibrium for $R = 1.3$, multiple equilibria for $R = 1.4$ and a unique pooling equilibrium for $R = 1.5$.


Equilibrium manifold

Density for ν

Fixed point calculation $E[\pi]/R$

Fixed point calculation

Bounds for $R = 1.40$, (blue: low, red: high)

Volume lent at R and p

Fig. 11 For the mixture of two exponentials, where $\gamma\bar{h}=2$, $\delta=0.1$, $\kappa=0.15$, $\lambda_1=-20$, $\lambda_2=5$, and $\omega=0.8$ (with κ the cost for verification), we obtained a unique separating equilibrium for $R=1.4$, multiple equilibria for $R=1.58$ and a unique pooling equilibrium for $R=1.65$.

speculators are drawn to the market, hoping to sell at a higher price tomorrow. Eventually, this comes to an end, the speculators withdraw from the market and prices come crashing down. While appealing, formalizing this intuition tends to run into the obstacle that if prices are expected to be high tomorrow, then demand for credit and thus houses should be high today, and that should drive up prices today, making it less likely that prices will increase. Or, put differently, as prices rise, they eventually must get to a near maximum at some date (unless for some reason agents can buy on credit against the future resale): call that date "today." At that point, prices are expected to decline in the future. But if so, banks and speculating households are less likely to buy today and prices should not be high today to start with. These types of bubbles are typically ruled out by thinking about a "last fool" who is willing to buy at the highest price, when prices can only go down from there: such fools should not exist in rational expectations equilibria. Agents should realize that prices cannot outgrow the economy forever: using backward induction, the bubble then gets stopped dead in its tracks before it can get going at all. That is why formalizing this popular story indeed presents a challenge.

Two strands of the literature have evolved to address this challenge. One strand of the literature, that we refer to as "bubbles," keeps the expected growth rate of bubbles bounded by the growth rate of the economy, thus circumventing the backward induction logic. These models may additionally invoke irrational beliefs or differences in sentiments, but many do not. Another strand of literature, that we define "sentiments," allows agents to believe that bubbles will grow faster than the economy, but then also needs to invoke irrational beliefs for at least some portion of the agents in order to disable the backward induction logic described earlier.

Related to the literature on bubbles, there is another strand of literature that focus on generating momentum in house price changes. Among the papers in this literature, Case and Shiller (1989), Barberis et al. (1998), Hong and Stein (1999), Capozza et al. (2004), Frazzini (2006), Glaeser et al. (2014), Anenberg (2014), Head et al. (2014), Glaeser and Nathanson (2016), and Guren (2016).

5.1 Bubbles

Perhaps the most prominent and earliest example of a model with bubbles is the celebrated overlapping generations model of money by Samuelson (1958). If other means of savings do not produce a rate of return higher than the growth rate of the economy, then an intrinsically worthless asset (fiat money) can have a nonzero price in terms of goods, as it gets sold by the currently old agents to the next generation of currently young agents. Such economies need to be dynamically inefficient, satisfying the Cass criterion or Balasko–Shell criterion, see Cass (1972) and Balasko and Shell (1980). In such economies, the first welfare theorem might not hold, competitive equilibria might not be Pareto optimal. One may achieve a Pareto improvement by giving resources to the current old from the current young, who in turn receive resources, when they are old from the next young generation, ad infinitum. There are various ways to implement or interpret such a transfer scheme. Samuelson interpreted the scheme as fiat money, issued perhaps by the initially old generation. Others have interpreted it as government debt, to be rolled over forever, or as an unfunded pension system.

Another stream of literature that features rational bubbles is the search literatures with fiat money. The seminal paper in this literature is Kiyotaki and Wright (1989) that proposes a model of decentralized trade where agents meet randomly and fiat money can arise as general medium of exchange.

The literature of greatest interest to us here has interpreted this transfer scheme as a bubble, and has discussed how such bubbles might get introduced by various generations. In some of these papers, the transfer scheme is stochastic, and may end in any given period with some probability, generating a crash. For example Carvalho et al. (2012), Martin and Ventura (2010), Martin and Ventura (2012), Martin and Ventura (2014), and Martin and Ventura (2010) employ various versions of OLG models, in which, ideally, resources should be funneled from inefficient investors or savers to efficient investors or entrepreneurs. In particular, Martin and Ventura (2010) have used this framework to think about the financial crisis of 2008. There may, for example, be some lending friction, where entrepreneurs cannot promise repayment. They may be limited in how much paper they can issue against future cash flow from the project, or perhaps they need more financing than can be achieved by issuing such paper. They can additionally issue intrinsically worthless "bubble" securities, valued only because the buyer hopes that someone else buys them in the future. The issuance of such bubble paper starts another sequence of the intergenerational transfer scheme described earlier. The existing bubble paper in the hands of old agents as well as those created by newborn entrepreneurs get sold to savers. Savers find investing in these bubbles more attractive than investing in their own, inefficient technologies. This technology needs to be inefficient enough so that its return is on average below the growth rate of the economy, creating the dynamic inefficiency for bubbles to arise.

He et al. (2015) likewise focus on houses to facilitate intertemporal transactions when credit markets are imperfect, and the resulting liquidity premium for house prices. They obtain deterministic cyclic and chaotic dynamics as self-fulfilling prophecies, though their equilibria do not display an extended price run-up followed by collapse. Differences in beliefs are at the heart of trading in Scheinkman and Xiong (2003): the belief differences create a bubbly, but (on average) nongrowing component of asset prices.

Another paper that generate bubbles with fully rational agents and perfect foresight is Wright and Wong (2014). Here, bubbles arise in a model of bilateral exchange that involve chains of intermediaries in markets with search frictions and bargaining problems.

5.2 Sentiments

The other strand of literature we are going to focus in this section, is what we called "sentiment literature." In Section 6, we are going to propose a simple model that captures the "sentiment" idea that we mentioned in the introduction: asset prices may be above fundamental value because agents "irrationally" believe that there is always a "greater fool" who is going to be willing to buy at an even higher price. and discuss the related literature there. Here, we summarize a number of variants of this story formalized in different ways in the literature, where assets are trading above fundamental values due to diverse beliefs between optimists and pessimists, creating an "add-on" above the fundamental value, possibly requiring short-sale constraints on the pessimists. Static versions can be found in Geanakoplos (2002) or as in Simsek (2013). Dynamic versions are in Harrison–Kreps (1978) or in Scheinkman and Xiong (2003): in the latter paper, agents do understand, however, that the bubble will not grow faster than the economy, on average. Glaeser et al. (2014) study bubbles and their role for the housing market. They claim that rational bubbles can obtain, if there is no new construction. They do not provide a full general equilibrium formulation or description of the underlying credit market for this claim: one interpretation may be that agents can buy on credit against the future resale or have unlimited "deep pockets" to buy at any price. They rule out bubbles with elastic housing supply, and then proceed to use a model of irrational, exuberant buyers to study housing bubbles, relating the length and frequency of bubbles and their welfare consequences to the elasticity of housing supply.

Should asset price movements be taken into account in the central-bank interest rate setting? And if so, how? To answer this question, Adam and Woodford (2012) study the optimal monetary policy in a new Keynesian model with a housing sector, using near-rational equilibria, as developed in Woodford (2010) and allowing for a set of possible and internally coherent probability beliefs, that are not too different from the benchmark.

The papers incorporating diverse beliefs probably come closest to our model in Section 6. However, these models typically focus on the case where agents are either pessimistic or optimistic: by construction then, the optimists must be the "greatest fools." We instead wish to incorporate the idea that the more optimistic buyers are typically not yet the greatest fools themselves, but simply betting on even greater fools out there. At the extreme end, the "greatest fool" must be someone willing to pay for something that is intrinsically worthless (to all others), without ever being able to sell to someone at an even higher price, and there may be quite foolish people out there with overly strong optimism of being able to sell to such "greatest fools." So, at that end, the model may require substantial irrationality. The key here is, however, how this suspected strong irrationality at the upper end of the potential price distribution trickles down to the price and sales dynamic among the less foolish or even rational part of the population.

The model shares many elements with Golosov et al. (2014). There, assets are traded in a sequence of bilateral meetings between agents having different information regarding the fundamental value of the asset. By contrast here, everyone understands the asset to be intrinsically valueless: the differences arise in beliefs regarding the optimism of others. As such, our chapter is more closely related to Abreu and Brunnermeier (2003). A benchmark example for a dynamic model exploiting heterogeneous beliefs and changing sentiments, is the "disease" bubble model of Burnside et al. (2013). There is some intrinsically worthless bubble component, which could be part of the price of a house. An initially pessimistic population may gradually become infected to be "optimistic" and believe the bubble component actually has some intrinsic value: once, everyone is optimistic (forever, let's say), there is some constant price that everyone is willing to pay. However, "truth" may be revealed at some probability every period, and clarify that the bubble component is worthless indeed. Then, during the pessimistically dominated population epoch, prices rise during the nonrevelation phase, since the rise in prices there compensates the pessimistic investor for the risk of ending up with a worthless bubble piece, in case the truth gets revealed. The price will rise until the marginal investor is optimistic: at that point, the maximum price may be reached.

Another related recent paper is Bordalo et al. (2016), where credit cycles arise from "diagnostic expectations," that is, from the assumption that when they form expectations agents overweight future outcomes that seems more likely in light of the recent data. This can generate excess volatility, overreaction to news and predictable reversals.

6. A SIMPLE MODEL OF SENTIMENTS

In this section, we are going to propose a simple model to formalize and examine the following and often-told story about buyers and sellers in asset markets. We wish to use it in particular for thinking about the housing market, but it may apply more generally to the stock market or any other market in which assets get retraded.

The story we have in mind is as follows. Prices for assets sometimes bubble above their fundamental value and then come crashing down. They do so due to buyers betting on greater fools. More precisely, when a buyer buys an asset, she may realize that the price is above its fundamental value, but is betting on being able to sell the asset at a future date at an even higher price to a greater fool. What matters to the buyer is not, how foolish it is to keep the asset itself, but how foolish other participants are.

There are variants of this story formalized in various way in the literature, as we discussed in the literature review in Section 5.2. For our showcase model below and in contrast to the models of rational bubbles discussed in the literature review in Section 5.1, we do not assume that the economy is, effectively, dynamically inefficient. That is, we do not wish to assume that bubbles can be traded forever, because the return to be earned on these assets, as perceived by the agents, does not exceed the growth rate of the economy.

It may be important, however, to examine models in which the required rate of return is higher than the growth rate. It is then clear from the start that the price eventually must hit a ceiling: say, when the value of the asset exceeds all resources in the hands of the buyers. Usual backward induction arguments then rule out such bubbles in the first place, see Tirole (1985). The purpose of this section is to tweak the rationality argument per introducing a mythical "greatest fool," thwarting that backward induction. This "greatest fool" can alternatively be interpreted as a rational "collector," who just happens to value an asset at high price, while nobody else does. We will consider environments where this person is a myth indeed. Agents falsely believe, however, that this mythical collector is out there. Some particularly optimistic believers will buy the assets and hold it, in the hope of ultimately selling to the collector, but more importantly, some traders will buy the assets in the hope of selling to an agent who has an even more optimistic beliefs about the existence of a collector. This is what we mean by a sentiment-driven bubble. Note that it does not actually matter whether such collector agents are present: all that matters is the beliefs by the various agents in the presence of such agents. We allow for the belief in such collectors to suddenly disappear: if that happens, the price crashes.

It should be clear that one can construct higher-layer type theories too. For a second-layer theory, all agents may agree that there are no collectors. However, they may all believe that a certain fraction of "first-layer" agents out there does believe such collectors to be there, and the more optimistic agents may believe that fraction to be higher. Agents in such an economy will then not per se wait to sell to a collector (they know they cannot), but wait to sell to an agent who believes such collectors to be present. Furthermore, the agents that are less optimistic regarding the existence of such believers will sell to agents who are more optimistic regarding such first-layer believers. Once again, a bubble can arise, this time even if actually neither the collector nor first-layer believers are present in the economy. A third-layer theory would be about agents differing in their beliefs of meeting agents who believe that they can meet an original believer, etc. We feel that it would be fascinating to explore the ramifications and variations of the simple model below a lot further than we do. It is just meant as an inspiration and starting point.

6.1 The Model

Time is continuous, $t \geq 0$. There is initially a continuum of agents of total mass one. There is distribution of agent types $\theta \in [0, 1]$ in the economy, characterized by the distribution function $H(\theta)$. We assume that H has a density $h(\theta)$. We call agents of type $\theta = 1$ "collectors." We shall assume that $H(\theta^{\max}) = 1$ for some $\theta^{\max} < 1$, and thus, the distribution H assigns no weight to collector types.

Agents differ in their beliefs about the distribution of beliefs in the population, with θ parameterizing that belief. Specifically, we shall assume that, initially, an agent of type θ believs that other agents' type x is drawn from

$$H_\theta(x) = (1-\theta)H(x) + \theta 1_{x=1}. \tag{53}$$

In other words, agent θ uses a weighted average between the true distribution and a point mass at the collector type. Most of the analysis below carries over to a more general formulation: we leave these extensions to future research. Agents are aware of their differing beliefs, but they individually nonetheless insist on the beliefs they hold. We assume that an aggregate revelation event may arrive at the arrival rate α (or instantaneous probability αdt), at which point all agents suddenly understand that there are no collector types and their beliefs switch to the true distribution H. One might wish to assume that agents are unaware that this revelation event could occur (MIT shock), but it turns out that the mathematics is not much different if they do: so we shall assume that. In the latter case, the better interpretation is that agents believe that, with some probability αdt, the distribution of population types changes from H_θ to H, interpreting this as a taste and belief shift for other agents.

There is a single and indivisible asset (coconut), initially in the hand of an agent of type $\theta = 0$. There are random pairwise meetings between agents: due to our assumption of a single asset available for trading, it suffices to describe the meetings between the agent that currently has the asset and some other agent. If the agent currently holding the asset is of type θ and if the revelation event has not yet happened, then she will believe that she meets an agent drawn from the distribution H_θ at rate λ. The asset-holding agent (who we shall call the "seller") posts a take-it-or-leave-it price q_θ (the posted contract can be generalized, and we leave this to future research). The other agent (who we shall call the "buyer") decides to accept or reject the trade. If the trade is rejected, the seller keeps the asset and keeps on waiting for the next pairwise meeting. If the trade takes place, the buyer produces q_θ units of a consumption good or "cash," at instantaneous disutility q_θ, which the seller consumes, experiencing instantaneous utility q_θ. The future is discounted at rate ρ. The buyer receives the asset and then in turn waits for the next pairwise meeting. If the buyer turns out to be a collector, he will be willing to buy the asset at any price at or below some exogenously fixed value $v(1)$. The asset may provide some intrinsic value to the collector or the collector may simply be the "last fool," failing to understand that he can sell the asset at an even higher value in the future. In any case, the asset offers no intrinsic benefit to any agent who is not a collector. In other words, we assume that noncollector agents have preferences given by

$$U = E\left[\int_0^\infty e^{-\rho t} c_t dt\right] \tag{54}$$

where we allow c_t to be negative, and where c_t is the consumption flow resulting from these trades. We assume that the discount rate is strictly positive, $\rho > 0$.

A few brief remarks may be in order. We have not used time subscripts for q_θ, though there may be equilibria, in which these prices do depend on time. Here, we shall concentrate on time-invariant solutions, for simplicity. More importantly, it may seem odd to consider only a single asset, given that we have a whole continuum of agents at our disposal. This assumption considerably simplifies the analysis, though, as it allows us not to distinguish between meetings, where the potential buyer already owns an asset or not. Furthermore, over time, agents would need to keep track of the distribution of asset-owning types. It is plausible that these distributions shift to the right over time, ie, that it is the higher types holding assets, as time progresses. In the decision problem to be analyzed, selling agents would then need to forecast these evolutions, creating potentially intricate interactions and complications that go beyond the scope of this chapter. These would be good topics to pursue in future research.

Note that, in essence, the bubbly economies described in Section 5.1 can be understood as featuring $\rho \leq 0$ with $\alpha = 0$ (and finite lives), so that agents are willing to agree to a trade, in which they give up more today than they receive later, or at least do not insist on getting more later on. Here, we rule out this channel. Note also that the search theory models of money like Kiyotaki and Wright (1993) and their successors assume that the total sum of consumption is larger than zero, ie, that the seller benefits more from the sale than the buyer is hurt. If trades can only take place, using the intrinsically worthless asset, the asset helps in achieving a better outcome than autarky. Related modeling devices are used in Harrison–Kreps (1978) or Scheinkman and Xiong (2003). Here, by contrast, we shut down any benefits from the trade per se.

6.2 Analysis

We formulate the strategies of buyers as threshold strategies. A buyer of type θ picks some value v_θ, and purchases the asset, if the take-it-or-leave-it price is at or below that value, provided the revelation event has not yet taken place. For the collector type, $v_1 > 0$ is a parameter. A seller of type θ picks a take-it-or-leave-it price q_θ before the revelation event. After the revelation event has taken place, the asset is valued at zero by all and traded at zero price. A Nash equilibrium is then given by two functions $(v_\theta, q_\theta)_{\theta \in [0,1]}$, so that the strategies of agent θ maximize the utility function (54), given the strategies of all other agents. We shall additionally impose that v_θ is measurable. A seller can only hope to sell the asset in the future before the revelation event takes place. Put differently, we can assume that the seller discounts the future at rate $\alpha + \rho$, and that any before-revelation value of the asset to the seller in the future is discounted at that rate too. We could introduce a new symbol for $\alpha + \rho$. In slight abuse of notation (or appealing to the “MIT shock logic”), we shall continue to use ρ for that discount rate.

Consider now the before-revelation phase. We seek to characterize the Nash equilibrium or Nash equilibria. It is straightforward to see that a buyer of type θ will choose to buy at any price not bigger than v_θ, where v_θ is his continuation value of holding the asset. Consider then a seller of type θ, contemplating a sale price $0 \leq q \leq v(1)$

(obviously, it does not make sense to post a price above $v(1)$ or below zero). He assumes that his buyer's type x is drawn from the distribution H_θ, and that buyers follow their equilibrium strategy v_x and buy only if $q \leq v_x$. Hence, conditional on meeting a buyer, a seller of type θ who posts price q expects to sell with probability

$$\phi_\theta(q) = (1-\theta)\int 1_{v_x \geq q} h(x)dx + \theta \tag{55}$$

Proposition 6 *The probability of a sale $\phi_\theta(q)$ is decreasing in q and increasing in θ.*
Proof The proof is immediate, once one rewrite Eq. (55) as

$$\phi_\theta(q) = (1-\theta)\phi_0(q) + \theta. \qquad \square$$

If the trade takes place, the seller receives q. Trading possibilities arrive at rate λ, so in a time interval dt, the sale takes place with probability $\lambda\phi_\theta(q)$. Otherwise, the seller will remain owner of the asset at time $t + dt$, still valuing the asset at $V_\theta(q)$ then (provided the aggregate revelation event has not taken place: remember, that we implicitly took care of that via our discount factor ρ). Therefore, the continuation value of a seller of type θ who chooses a sale strategy q, $V_\theta(q)$, is equal to

$$V_\theta(q) = \lambda\phi_\theta(q)q\,dt + (1-\lambda\phi_\theta(q)\,dt)(1-\rho\,dt)V_\theta(q), \tag{56}$$

or, canceling higher order terms,

$$V_\theta(q) = \frac{q}{\dfrac{\rho}{\lambda\phi_\theta(q)}+1}. \tag{57}$$

The optimal selling strategy $q = q_\theta$ is the one maximizing $V_\theta(q)$, that is,

$$q_\theta \in \operatorname{argmax} V_\theta(q) \tag{58}$$

delivering $v_\theta = V_\theta(q_\theta)$.

It is easy to construct two bounds for the optimal continuation value. On the one hand, consider the suboptimal strategy that agents will only attempt to sell to the collector, per posting the price $q = 1$. This strategy would give

$$\underline{v}_\theta = \frac{v_1}{\dfrac{\rho}{\lambda\theta}+1}. \tag{59}$$

Clearly the optimal value function cannot be lower than $\underline{v}$, as in equilibrium there would be more trade for speculative reasons. On the other hand, consider the widely optimistic assumption, that any potential buyer is willing to purchase the asset at $q = v_1$. The value function would then be given by

$$\bar{v}_\theta = \frac{v_1}{\dfrac{\rho}{\lambda}+1} \tag{60}$$

where the omission of θ is the difference to (60). It is straightforward to show that the equilibrium value function in between these two bounds.

Proposition 7 *Suppose that the function $v : x \mapsto v_x$ used for calculating $\phi_\theta(q)$ in Eq. (55) is measurable and satisfies $\underline{v} \leq v_x \leq \bar{v}$. Then V_θ has a maximum.*

Proof Note that $V_\theta(q)$ is bounded by $\bar{v}_\theta$. Let $q^{(j)}$, $j = 1, 2, \cdot$ be a sequence, so that $V_\theta(q^{(j)})$ is increasing, converging against $\sup V_\theta(q)$. Since $q^{(j)} \in [0, v(1)]$, we can find a convergent subsequence, which we can furthermore assume to be monotone. Wlog, assume that $q^{(j)} \rightarrow q^*$ for some q^* and is monotonically increasing or decreasing. If the sequence $q^{(j)}$ is monotonically increasing,

$$\bigcap_j \{x \mid v_x \geq q^{(j)}\} = \{x \mid v_x \geq q^*\}$$

Therefore $\phi_\theta(q^*) = \lim_{j\rightarrow\infty} \phi_\theta(q^{(j)})$ and hence

$$\frac{q^*}{\dfrac{\rho}{\lambda\phi_\theta(q^*)} + 1} = \lim_{j\rightarrow\infty} \frac{q^{(j)}}{\dfrac{\rho}{\lambda\phi_\theta(q^{(j)})} + 1}$$

If the sequence $q^{(j)}$ are monotonously decreasing, then

$$\bigcup_j \{x \mid v_x \geq q^{(j)}\} \subseteq \{x \mid v_x \geq q^*\}$$

and therefore $\phi_\theta(q^*) \geq \lim_{j\rightarrow\infty} \phi_\theta(q^{(j)})$. Hence

$$\frac{q^*}{\dfrac{\rho}{\lambda\phi_\theta(q^*)} + 1} \geq \lim_{j\rightarrow\infty} \frac{q^{(j)}}{\dfrac{\rho}{\lambda\phi_\theta(q^{(j)})} + 1}$$

Here, though, ">" is ruled out, since the right-hand side is the supremum of $V_\theta(q)$. We can conclude that q^* maximizes $V_\theta(q)$. □

The axiom of choice now implies that q_θ is well defined.

Proposition 8 *The value v_θ of any Nash equilibrium is increasing in θ.*

Proof Let $\tilde{\theta} > \theta$. Note that $V_{\tilde{\theta}}(q) \geq V_\theta(q)$ for all q, since $\phi_{\tilde{\theta}}(q) \geq \phi_\theta(q)$. Since this is true in particular at $q = q_{\tilde{\theta}}$, the claim now follows. □

Now we can define the set of potential value functions

$$\mathcal{V} = \{v : [0,1] \rightarrow \mathbb{R} \mid v \text{ is increasing and } \underline{v} \leq v \leq \bar{v}\}.$$

Given that increasing functions are measurable, we can consider the mapping $T : \mathcal{V} \rightarrow \mathcal{V}$, defined by the following steps:

1. map $v \in \mathcal{V}$, into a function $\phi_\theta(q)$, using Eq. (55);
2. map ϕ into a function $V_\theta(q)$ using Eq. (56);
3. map $V_\theta(q)$ into the function v_θ that maximizes $V_\theta(q)$ (this maximum exists thanks to Proposition 7).

Proposition 9 *The mapping $T:\mathcal{V}\rightarrow\mathcal{V}$ is monotone and has a fixed point in $\mathcal{V}$. Therefore, a Nash equilibrium with $v\in\mathcal{V}$ exists.*

Proof For monotonicity, check that each step of the mapping is monotone. That is, if $\widetilde{v}\geq v$, then $\widetilde{\phi}\geq\phi$ for the first step, and so forth, where the inequalities are understood to hold pointwise for all arguments. Note that $\mathcal{V}$ is a complete lattice, with the usual order structure. Tarski's fixed point theorem now delivers the result that the set of fixed points of T forms a nonempty complete sublattice of $\mathcal{V}$. □

Next proposition shows that the equilibrium exhibits a threshold property.

Proposition 10 *For each sale price q, there is a threshold buyer type $\underline{x}(q)$ such that all buyers of type $x\geq\underline{x}(q)$ will buy the asset and all buyers of type $x<\underline{x}(q)$ will not, ie,*

$$x\geq\underline{x}(q)\quad\Leftrightarrow v_x\geq q. \tag{61}$$

The function $\underline{x}(q)$ is increasing in q. Furthermore, for $q\leq v_1$,

$$\phi_\theta(q)=(1-\theta)(1-H(\underline{x}(q)))+\theta. \tag{62}$$

Proof The proof follows immediately from Proposition 8. □

To obtain a bit more analytic insight, consider a price q, where $\underline{x}(q)$ is differentiable.

Proposition 11 *Suppose $\underline{x}(q)$ is differentiable at $q=q_\theta$. Then, the optimal q_θ satisfies the first-order condition*

$$0=1+\frac{\lambda}{\rho}\phi_\theta(q)-\eta_\theta(q) \tag{63}$$

where $\eta_\theta(q)$ is the elasticity of the sale probability,

$$\eta_\theta(q)=-\frac{\phi_\theta'(q)q}{\phi_\theta(q)}=\frac{h(\underline{x}(q))\underline{x}(q)'q}{\frac{1}{1-\theta}-H(\underline{x}(q))} \tag{64}$$

Proof Differentiate $V_\theta(q)$ with respect to q, and note that $V_\theta'(q)=0$ at $q=q_\theta$. □

One can rewrite the sales probability elasticity a bit further. Let

$$\psi_\theta(x)=(1-\theta)(1-H(x))+\theta$$

be the probability of meeting a buyer of type x or better (including the collector), from the perspective of a type-θ seller. Define its elasticity

$$\eta_{\theta,\psi}(x)=\frac{\psi_\theta'(x)x}{\psi_\theta(x)}=-\frac{h(x)x}{\frac{1}{1-\theta}-H(x)}$$

Define the elasticity of the threshold buyer type,

$$\eta_{\underline{x}}(q)=\frac{\underline{x}(q)'q}{\underline{x}(q)}$$

Then

$$\eta_\theta(q) = \eta_{\theta,\psi}(\underline{x}(q))\eta_{\bar{x}}(q).$$

This is the usual chain rule for elasticities, of course, applied to $\phi_\theta(q) = \psi_\theta(\underline{x}(q))$.

The results earlier suggest a strategy for characterizing an equilibrium. Suppose, one has some conjectured threshold buyer type function $\underline{x}(q)$, which is increasing and differentiable in q. With that, solve (63) for the optimal strategy q_θ and thereby for the value $v_\theta = V_\theta(q_\theta)$. With the value, calculate the resulting buyer threshold type function

$$\underline{x}^*(x) = \text{argmin}_x v_x \geq q$$

If $\underline{x}^*(q) = \underline{x}(q)$, one has obtained an equilibrium.

It may be possible to obtain analytical examples, for smart choices for H, exploiting this strategy. We leave this to future research to pursue. Here instead, we shall provide a numerical example.

6.3 Numerical Example

Rather than employing the first-order conditions above, we compute equilibria, using a rather brute-force grid-maximization algorithm. We create a suitable grid in q and θ. We start the iteration at the lower bound $v^{(0)} = \underline{v}$, defined over a grid in θ. We iterate on the mapping $T : \mathcal{V} \to \mathcal{V}$ described earlier. Specifically for step j, calculate $\phi_\theta^{(j)}(q)$ on the q-grid, using $\underline{x}^{(j-1)}$ on the right-hand side of Eq. (62). Now, calculate $V_\theta^{(j)}(q)$ per (57) for all grid values θ and q. For each grid value θ, find $v_\theta^{(j)}$ as the maximum of $V_\theta^{(j)}(q)$ over the grid values q. For each grid value q, find the smallest x, so that $v_x^{(j)} \geq q$, exploiting (61). This is the new $\underline{x}^{(j)}(q)$ and the next iteration step can commence. Iterate sufficiently often to obtain a reasonable degree of accuracy with the last solution.

As parameters, we chose $\lambda = 1$, $\rho = 0.1$ and let H be a uniform distribution on [0, 0.25]. The "collector price" was normalized at $v_1 = 1$. We used an evenly spaced grid of 500,001 points for q and 1001 points for $\theta \in [0, 1]$ or 251 points in the relevant range [0, 0.25]. As a starting point, we set $v^{(0)} = \underline{v}$, as defined in (60). For each grid value q, we then find the smallest $x = \underline{x}^{(0)}(q)$, so that $v_x^{(0)} \geq q$, exploiting (61).

The results are in Figs. 12 and 13. As one can see, agents with low θ pursue a strategy of seeking to sell to higher-θ agents, but beyond (approximately) $\theta = 0.1$, agents now only wait for the collector to make the sale. This can also be seen from the probability of sales. The black-dashed horizontal line shows the price chosen by the $\theta = 0$ types. This indicates that this market proceeds in two stages only, starting from the asset initially in the hands of a $\theta = 0$ agent (or an agent with a low θ). That agent will charge a price q_θ such that a sale only takes place, when meeting an agent with a fairly high $\tilde{\theta}$, who in turn hopes to sell to the collector, at $q_{\tilde{\theta}} = v_1 = 1$. It would be interesting to find examples, in

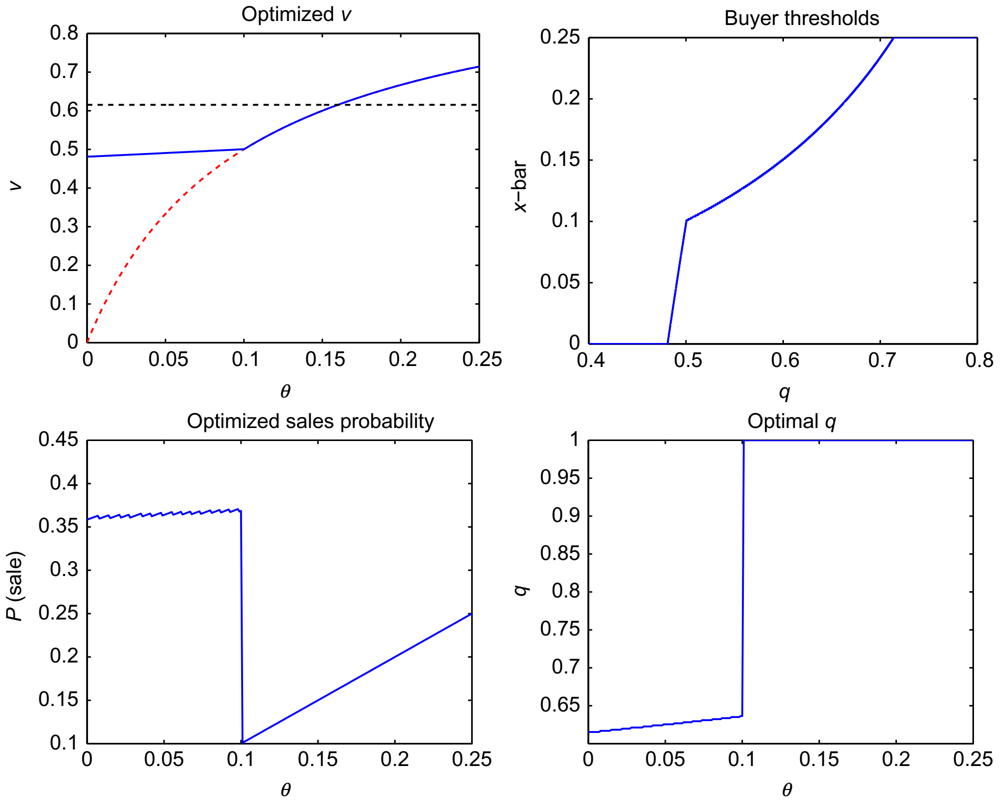


Fig. 12 Results from a numerical example. As parameters, we chose $\lambda = 1$, $\rho = 0.1$, and H to be a uniform distribution on $[0, 0.25]$. The "collector price" was normalized at $v_1 = 1$. In the top left panel, we compare the optimal value function to the value function $\underline{v}$ obtained per only selling to the collector. As one can see, agents with low θ pursue a strategy of seeking to sell to higher-θ agents, but beyond $\theta = 0.1$, agents now only wait for the collector to make the sale. This can also be seen from the probability of sales. The black-dashed horizontal line shows the price chosen by the $\theta = 0$ types. The top-right panel is essentially the top-left panel, flipped at the 45 degree line.

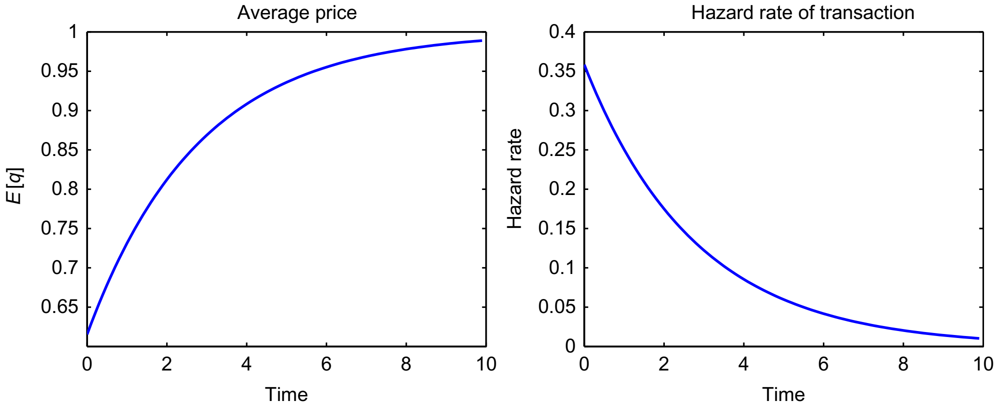


Fig. 13 Averaging over many posted price paths. The left panel shows the average price and how it is increasing over time. The right panel shows the hazard rate of a transaction and how it is decreasing over time.

which there are several stages of sale and resale to ever-more optimistic agents: we leave this to future research on this topic.

Consider now averaging across many simulations or individual markets, where the asset is initially held by the least optimistic agent $\theta = 0$. In principle, one can obtain the average price as well as the average hazard rate of a sale by simulation, using the results calculated thus far. Due to the two-stage structure of the sales process, it is easier to proceed analytically instead (and these arguments can be generalized to a multistage structure as well). If the asset is still in the hands of the initial $\theta = 0$ agent, it will be sold at the hazard rate $\xi = \phi_0(q_0)$, where we introduced the new symbol ξ for this hazard rate, to save on subsequent notation. Once the asset is sold, it will be posted at price $q_{\tilde{\theta}} = 1$ and not trade again, since there is no collector in the market. The unconditional date-t probability π_t, that the asset remains in the hands of initial $\theta = 0$ agent, solves the linear differential equation $\dot{\pi}_t = -\xi\pi_t$, with the solution given by $\pi_t = \exp(-t\xi)$. The average price is given by

$$E[q_t] = v_1 - (v_1 - q_0(q_0))\exp(-t\xi)$$

The unconditional or average hazard rate of a sale occurring is $\xi\pi_t$. Fig. 13 shows the resulting average price path and average sales hazard rate $E[\phi_\theta(q_\theta)]$. As one can see, the average posted price rises, while the average sales probability falls over time. The price path is conditional on the revelation event not occurring. Once the revelation event happens, the price crashes to zero. This captures the original story, with which this section got started.

The logic of the calculation just presented can also be used to calculate the results in Fig. 12 directly. Calculate first $v_\theta^{(0)} = \underline{v}_\theta$ for all θ per Eq. (60). Invert that function and use the distribution function H to calculate $\phi_\theta(q)$. With that, calculate $V_\theta(q)$ and find its maximum, for each q and the resulting $v_\theta^{(1)}$. The last step appears tedious, but may be solvable in closed form. This is the final result, in the situation that there are at most two stages of selling (first, sell to a more optimistic agent, second, attempt to sell to the collector), as here. One has to verify that indeed there are no further stages. Put differently, one now has to check that less optimistic agents would not now want to "change their mind" and sell to even more optimistic agents, who now value the assets higher, due to their reselling to optimistic agents. We leave the details of the calculations and the verification condition to the interested reader.

7. EVIDENCE

What caused the subprime crisis and, by extension, the financial crisis of 2008? What moved first, what moved later? What was cause and what was effect? The chapter has focused on two possible stories. One possibility is that house prices fell first for exogenous reasons, impairing bank balance sheets and leading to a financial collapse. Another possibility is that the banking system collapsed, leading to a reduction in mortgage lending and a fall in house prices. Perhaps, the fall in house prices triggered greater reluctance by banks to issue subprime loans, or perhaps and conversely, mortgage lending and, in particular, subprime borrowing, was reduced first, triggering a fall in house prices. Perhaps, subprime lending was reduced in the wake of higher delinquency rates on subprime mortgages or perhaps subprime lending was reduced, and the subsequent fall in house prices triggered delinquencies. Perhaps delinquencies rose because the pool of borrowers worsened or perhaps short-term interest rates rose, leading to higher rates for ARM mortgages and thereby higher delinquency rates. There are various ways of thinking through the interactions and tell the story. And how much did the interplay and feedback loop enhance the original shock?

Considerable research has been undertaken to seek to sort out these channels empirically: much more work still awaits to be done. We shall not attempt to give a full-fledged overview of the existing and large literature. We instead select some figures and facts from parts of the literature, and give them a somewhat impressionistic interpretation. Clearly, this is no substitute for careful empirical research on these data, but it may provide a good guide to questions and to developments to look at in the raw data. Most of the facts concern the United States. This generates a frontier of research interest and common ground for researchers to discuss, but it may miss important relationships and facts, compared to employing a world-wide perspective. We return to the latter towards the very end.

Figs. 1 and 14 show the S&P/Case-Shiller Home Price Indices. There is a run-up in house prices up to somewhere in the middle of 2006. According to the 20-city index

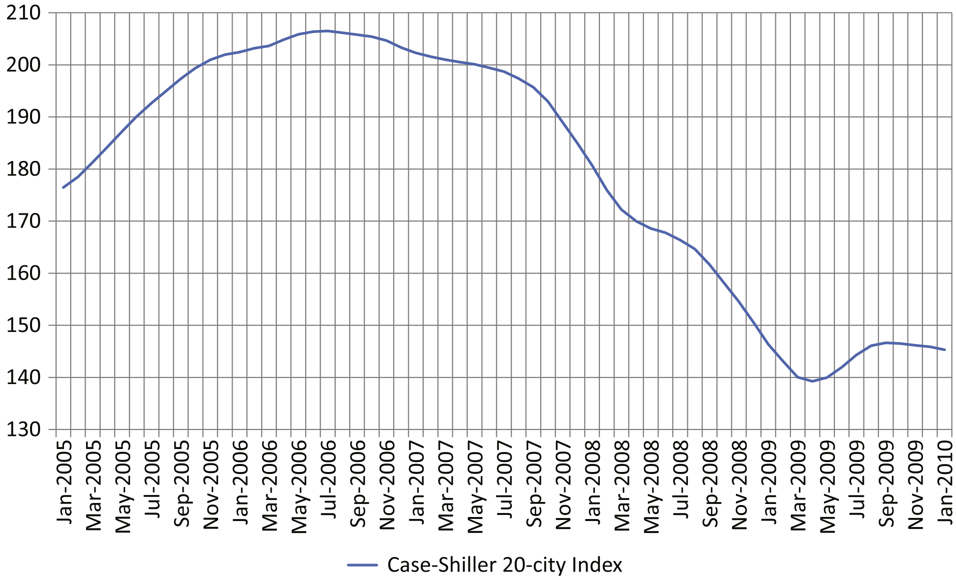


Fig. 14 The S&P/Case-Shiller home price 20 city index. *http://us.spindices.com/indices/real-estate/sp-case-shiller-20-city-composite-home-price-index*.

in Fig. 14, the peak is in July 2006. From there, prices started to drop, falling by 6.5% in October 2007, a relatively small drop. However, by October 2008, the date of the Lehman Brothers crisis and, in essence, the date of the financial collapse, house prices were at 25% below their July-2006 peak level, having fallen rather quickly and continuously from October 2007. House prices fell a bit more subsequently, reaching their bottom in April 2009, having fallen 32.6% from the original peak. From the sequence of these events, it appears plausible that house prices fell first, and the financial system collapsed subsequently.

However, the share of mortgages in the form of subprime fell quite substantially much earlier, as Fig. 2 reveals. Again, the peak subprime lending share of all mortgage originations occurred in 2006, at 23.5%, with a dramatic fall to 9.2% in 2007 and a near-zero in 2008. The peak occurred roughly at the same time as the peak of the S&P/Case-Shiller index in Figs. 1 and 14 and one might even wish to argue that the hump near the peak looks rather similar here. The decline in the peak subprime lending share from 2006 to 2007 was very sharp.

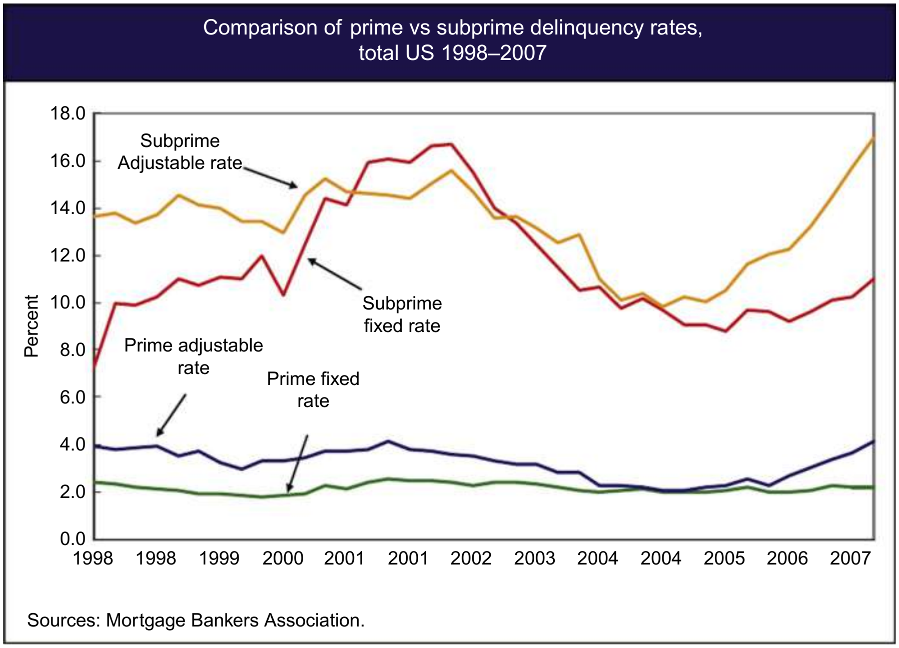
From this comparison, it appears plausible that subprime lending rose and fell together with house prices. If anything, subprime lending collapsed and decreased sharply, before house prices did. Thus, it may have been the reduction in subprime lending, causing the fall in house prices rather than vice versa.

One might wish to blame the decline in subprime lending on delinquency rates. Here, Fig. 15 is revealing. First, it shows that delinquency rates on fixed rate mortgages, be they prime or subprime, did not noticeably increase in 2006 and 2007: if anything, subprime fixed rate delinquency rates were near their all-time low of 2005.

The story is different for adjustable rate mortgages or ARMs. Here, rates did go up somewhat in 2006 and then somewhat more from their all-time low in 2004, but even there, the level in 2007 is rather comparable to the levels before 2002, both for prime adjustable rates as well as subprime adjustable rates, as Fig. 15 shows.

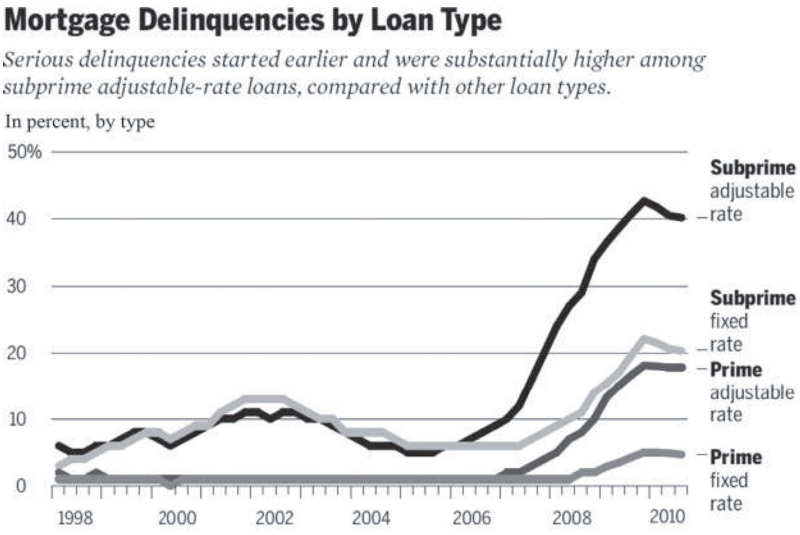
These movements are important for interpreting the course of events leading up to the crisis, but they are fairly small, compared to the subsequent development of delinquency rates shown in Fig. 16. Delinquencies later rose to unprecedented levels (at least for this time interval), peaking at rates somewhat above 40% for subprime adjustable rate mortgages around the end of 2009. In particular, delinquencies on subprime mortgages and prime adjustable rate mortgages had risen already considerably until October 2008, the date of the financial collapse. Overall, though, it does not seem plausible to argue that subprime lending was reduced in 2007, because delinquency rates had increased already at that point.

If anything, perhaps the delinquency rates in 2007 and the overall movement of delinquency rates on adjustable rate mortgages up to 2007 are linked to short-term interest rates. These rates are shown in Fig. 17. The Federal Reserve Bank increased the Federal Funds Rate in a sequence of small steps, starting at 1% in June 2004 to 5.25% in July 2006,



Sources: Mortgage Bankers Association.

Fig. 15 Before the crisis: subprime delinquency rates 1998–2007. *Senator Schumer, Rep Maloney, Report of Joint Economic Committee, 2007.*



Note: Serious delinquencies include mortgages 90 days or more past due and those in foreclosure.
Source: Mortgage Bankers Association National Delinquency Survey

Fig. 16 Including the crisis: subprime delinquency rates 1998–2011. *The Financial Crisis Inquiry Report, National Commission, January 2011.*

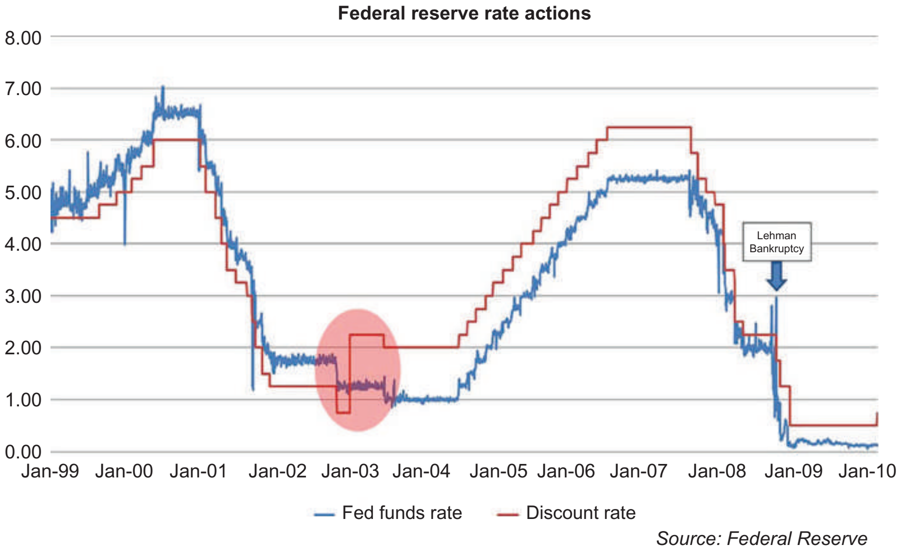


Fig. 17 Federal funds rate. *http://www.zerohedge.com/article/comparing-fed-funds-rateprimary-credit-discount-rate-over-past-decade*.

and then leveling off, before dramatically reversing course at the end of 2007. It is fairly plausible that the rise in short-term market interest rates from mid-2004 to mid-2006 resulted in the rise of delinquencies on adjustable rate mortgages from mid-2004 to mid-2006, seen in the previous figures.

Justiniano et al. (2015) have argued that house prices rose from 2000 to 2007 without an expansion of leverage, ie, at rather constant rates of mortgages to real estate, see the bottom-right panel of Fig. 18. The subsequent fall in house prices then went along with an increase in leverage and not necessarily a reduction in the volume of outstanding mortgages, see the top left panel.

These authors point to the fact that, first, there was an increase in available funds without an increase in leverage, leading to more mortgages at stable interest rates and stable leverage ratios, leading to a run-up in house prices.

For 2007 and beyond, they argue that the collateralizability of houses relative to available funds increased (or that available funds for lending decreased), leading to a rise in mortgage rates and a collapse in house prices. There certainly are some interesting comovements in Fig. 18 that deserve explanation, though not all readers may buy into the hypothesis that the collapse in house prices was caused by their relatively better collateralizability.

Most notably, Jorda, Schularick, Taylor and their coauthors have investigated the interplay between credit booms, house price booms and economic performance in a series of papers, constructing and providing new data sets along the way. Schularick

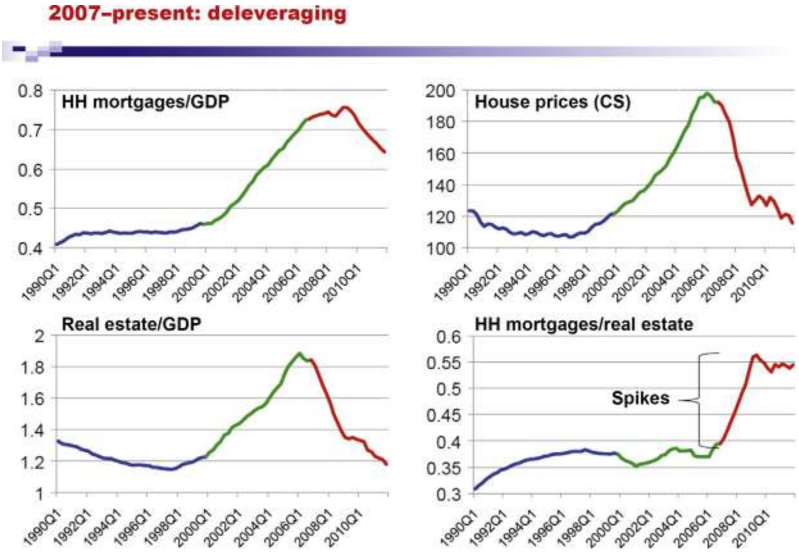


Fig. 18 The Justiniano–Primiceri–Tambalotti facts. *Justiniano, A., Primiceri, G.E., Tambalotti, A., 2015. Household leveraging and deleveraging. Rev. Econ. Dyn. 18, 3–20.*

and Taylor (2012) provide "a new long-run historical dataset for 14 developed countries over almost 140 years" and show how credit growth is a powerful predictor of financial crises. Jorda et al. (2013) subsequently argue that financial crisis recessions are costlier than typical recession. These data sets are updated and extended in Jorda et al. (2016a) for 17 advanced economies from 1870, covering disaggregated bank credit to the domestic nonfinancial private sector, with special attention to mortgage lending. They claim that "mortgage lending booms were only loosely associated with financial crisis before WWII, but …[have] become a more important predictor of impeding financial fragility" subsequently. Knoll et al. (2014) construct a house price index for 14 advanced economies from 1870 to 2012, assembling a variety of data sources. They argue that real house prices have largely followed a "hockey stick" pattern: fairly constant for a long time initially, followed a pronounced appreciation towards the end of the sample. They furthermore say that most of the price increase can be attributed to the increase in the price of land. Knoll (2016) subsequently argues that the rise in house prices coincides with a rise in the price–rent ratio. Combining data from these papers for 14 advanced economies, Jorda et al. (2015) claim that the 20th century has been an era of increasing "bets on the house." They write that "mortgage credit has risen dramatically as a share of banks' balance sheets from about one third at the beginning of the 20th century to about two thirds today." Using IV regressions, they show that "mortgage booms and house price bubbles have been closely associated with a higher likelihood of a financial crisis." Jorda et al. (2016b) provide further insights into the nature of these interactions, extending their data sets once more. They point to the fact that the build-up of leverage leads to higher tail risk. In a related paper, Mondragon et al. (2016), using a spatial IV-strategy, document empirical evidence that local credit supply shocks generate quantitatively significant boom–bust cycles in local house prices. In a similar spirit, Favara and Imbs (2015) show that an expansion in mortgage credit has significant effects on house prices, using the US branching deregulation between 1994 and 2005 as an instrument for credit. More recently, Di Maggio and Kermani (2016) show that a credit expansion can generate a boom and bust in house prices and real activity, using the change in national banks' regulation in 2004 by banning the antipredatory lending laws that a number of states adopted in 1999.

These series of papers and insights are completely in line with the fact that "a rise in household debt to GDP ratio predicts lower output growth," as shown by Mian et al. (2015).

We now show some figures taken from these papers to highlight some of these insights. Fig. 19, from Jorda et al. (2016b), show the "hockey stick" both for real house prices and mortgages.

The data can be sliced in other ways too, as Fig. 20 shows. That figure plots results both for the United States alone, their "benchmark economy," and for the sample of 17 countries investigated in Jorda et al. (2016b). House price growth and mortgage growth generally comove. In relation to real GDP, the real house price hockey stick, visible in Fig. 19, now becomes a downward trend, while the mortgage hockey stick becomes an upward trend. These figures raise intriguing, additional issues, concerning the attribution of changes in these series to their underlying causes.

Knoll et al. (2014) use Fig. 21 and additional analysis to show that house prices have risen faster than income in recent decades, while they have fallen relative to income in the first half of the 20th century and especially in the interwar period. Knoll (2016) argues that the rise in house prices coincided with a rise in the house price to rent ratio, as shown in Fig. 22. The price-to-rent ratio is similar to the often used price–dividend ratio for stocks, which has been shown to be useful for predicting stock returns. It is plausible that a similar phenomenon is at work for the housing market, as Knoll (2016) investigates.

The data sets created by these authors will be useful for further empirical investigations of the issues at hand. Luigi Bocola, in his discussion of a first draft of this chapter, combined the quarterly house price dataset from 1975 for a number of advanced countries, described by Mack and Martínez-García (2011), with the data for the 19 crisis events for advanced economies after 1975, described in Schularick and Taylor (2012). He constructed Fig. 23 to shed light on the relationship between house price growth, credit growth and GDP performance. The figure compares all crisis events (blue line) to the five events with the highest house price drop (red line) in the group: Denmark-87, Spain-08, Uk-91, Norway-88, Swe-91. This figure indicates once more the comovement of house prices and credit growth, but may suggest that the size of house price boom does not matter much for the average size of the subsequent recession.

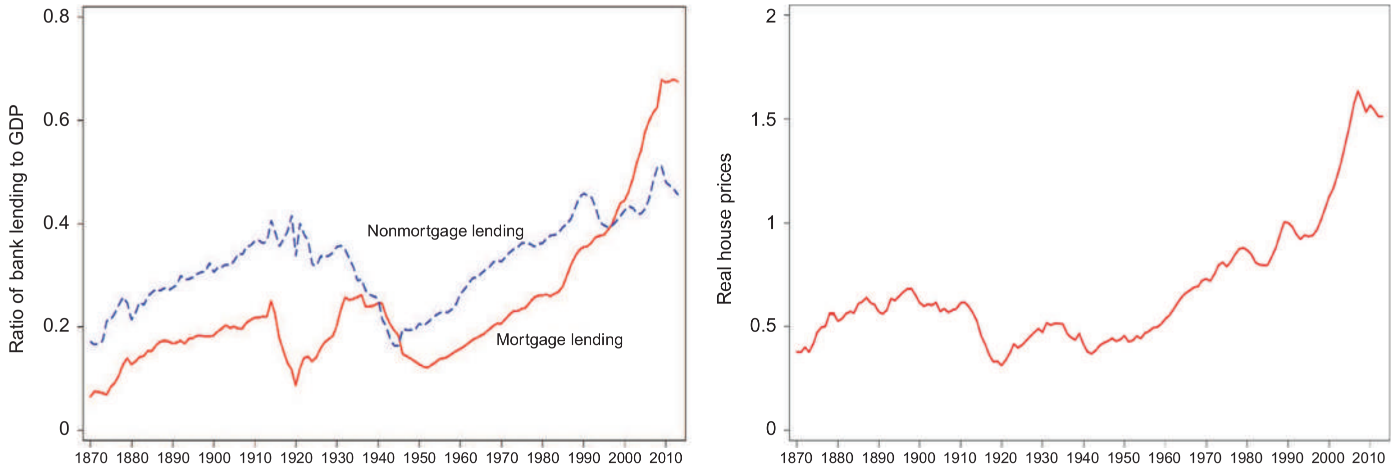


Fig. 19 Hockey sticks for real house prices and mortgages.

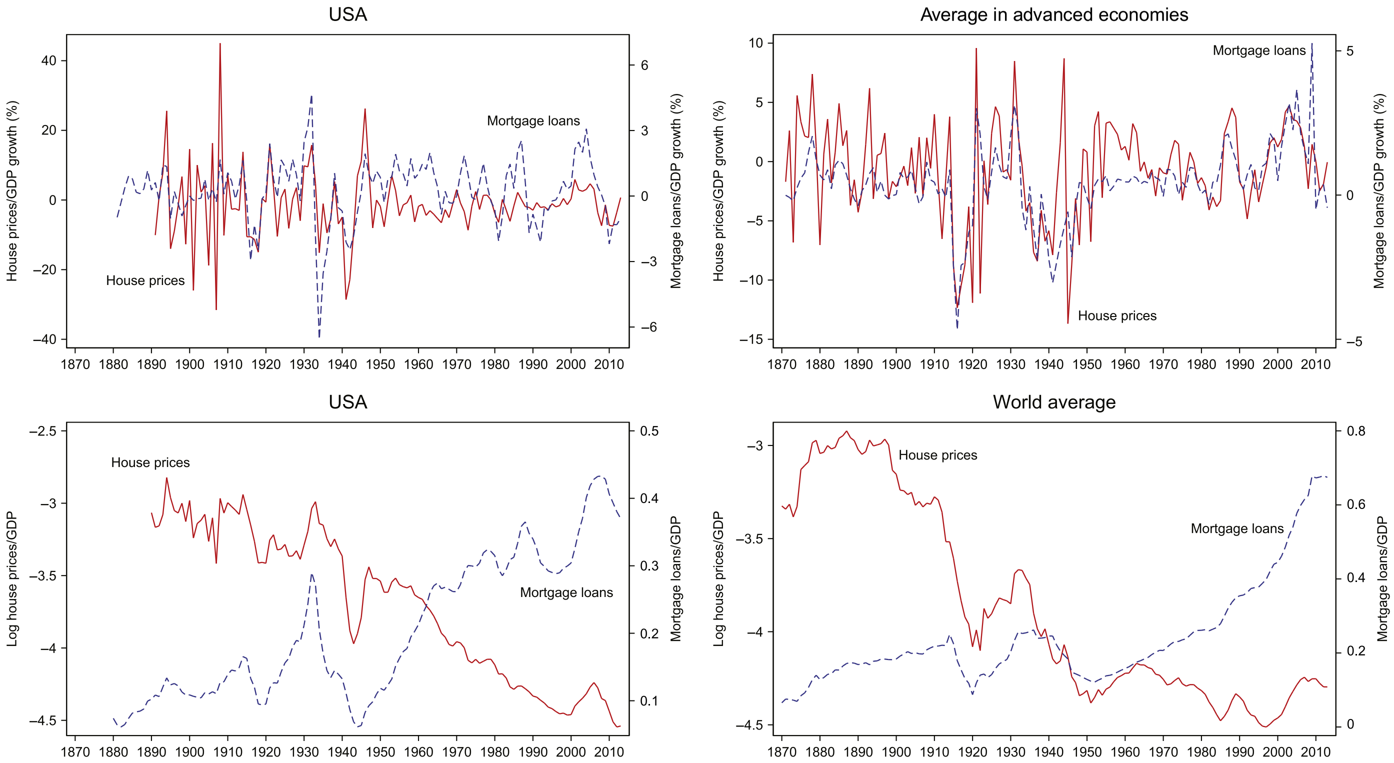


Fig. 20 Growth and trends in mortgages and house prices. *These figures have been provided by Helen Irvin and Oscar Jorda.*

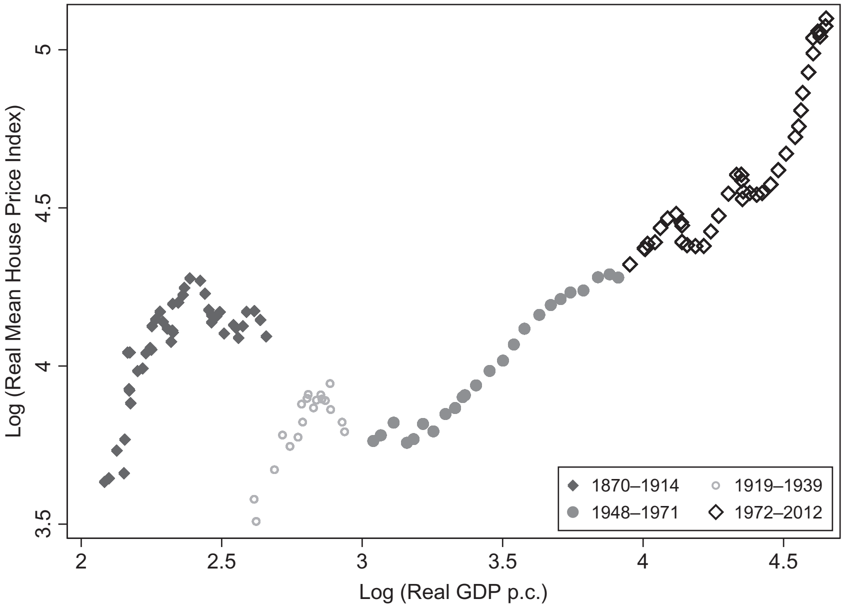


Fig. 21 House prices have risen faster than incomes in recent decades, while they have fallen relative to incomes in the first half of the 20th century and especially in the interwar period.

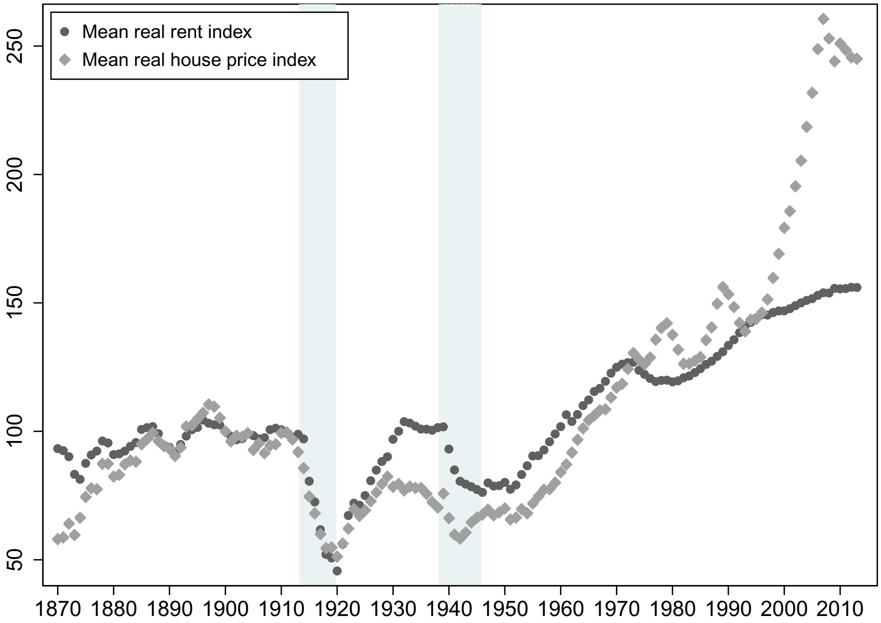


Fig. 22 The rise in house prices coincided with a rise in the house price to rent ratio.

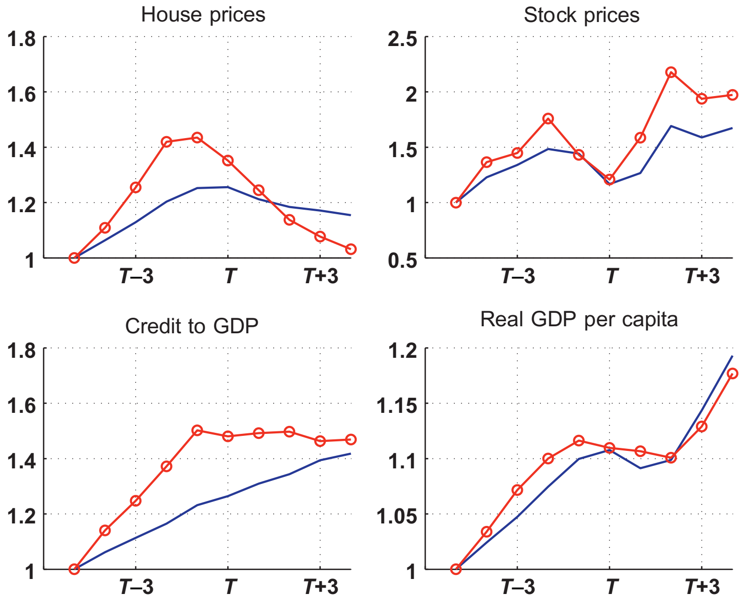


Fig. 23 Date "*T*" denotes the Schularick–Taylor crisis dates. The panels compare all crisis events (blue (dark gray in the print version) line) to the five events with the highest house price drop (red (black in the print version) line with circles) in the group: Denmark-87, Spain-08, Uk-91, Norway-88, Swe-91. The lines are cross-country averages for the four variables and normalized to equal unity at $T - 5$.

Once more, we want to stress that this section is not meant to be a comprehensive review of the empirical literature on this topic. Our objective was just to report a subsample of facts and empirical papers that relate to the theoretical literature we have focused on in this chapter. These are all suggestive figures. However, the debate on whether house prices have been the main driving source of the credit cycle or financial conditions the main driving force of house price cycle is still open and hopefully future research will sheds more light on this topic.

8. CONCLUSIONS

The purpose of this chapter was to explore a key connection between boom–bust episodes in housing markets and boom–bust episodes in credit markets and to point to their effects on macroeconomic activity. To do so, we investigated several benchmark approaches and channels, and related them to the existing literature. It is already a challenge to understand the house price boom–bust together with the credit boom–bust, without analyzing the aggregate activity repercussions. We therefore mostly focused on the interaction of the first two. In particular, there are two broad possible approaches to think about this interaction:

1. The house price boom–bust generates the credit boom–bust.
2. The credit boom–bust generates the house price boom–bust.

We started the chapter by proposing a stark mechanical model to think about this interaction. On purpose, it is designed to avoid several thorny issues that arise in a fully specified equilibrium model. Next, we explored these two main approaches in more detail.

First, we proposed a simple model of catastrophes, where we focused on the credit cycle. The idea is that an increase in credit availability can generate first a boom and then a bust in mortgage markets because of adverse selection issues. In particular, in a world where banks do not know the quality of their borrowers, that is, their expected default rate, and borrowers can either pool or pay a cost to verify their type, multiple equilibria can arise. If we start from an equilibrium with pooling, an increase in credit supply translates into a decrease in the quality of the pool of active borrowers (like "subprime borrowers"). This, in turns, can generate a switch to an equilibrium where good borrowers separate themselves and the pooling market crashes.

Second, we proposed a simple model of sentiments, where we focus on the housing price cycle. The main idea is that a house price bubble can arise when speculating households believe that there is always a "bigger fool" out there that is going to be willing to buy housing at a higher price. While appealing, formalizing this intuition tends to run into the "conundrum of the single equilibrium": if prices are expected to be high tomorrow, then demand for credit and thus houses should be high today (see above), and that should drive up prices today, making it less likely that prices will increase. Or, put differently, as prices rise, they eventually must get to a near maximum at some date: call it "today." At that price, prices are expected to decline in the future. But if so, banks and speculating households are less likely to buy today: but then, the price should not be high today. That is, in rational expectations models, there should not be anybody willing to buy at the highest price when prices can only go down from there. We break the curse of the conundrum by departing from the rational expectations framework and assuming that households always believe that with some positive probability there is a bigger fool, although he does not really exists.

In the course of the chapter, we related our simple models to the large literature on these topics. At the end, we also point to some empirical papers that propose facts related to these two theoretical approaches.

We wish this chapter is going to trigger further research and thinking on this important connection. As has become clear, the issues are far from resolved.

ACKNOWLEDGMENTS

We are grateful to Chiara Fratto, Ken Kikkawa, and Chao Ying for their excellent research assistance. For helpful comments, we are grateful to Fabrice Collard, Matteo Iacoviello, Oscar Jorda, Guido Lorenzoni, Moritz Schularick, Alan M. Taylor, and, in particular, to Luigi Bocola, Zhen Huo, and Nicolas Trachter for their thoughtful discussions as well as John Taylor for his detailed and useful comments. We are also grateful to audiences at the handbook conference at the University of Chicago, the "Multiple Equilibria and Financial Crises" conference at NYU, the "Mad Money" conference at the University of Wisconsin, and the seminar audience at the University of Toronto. This research has been supported by the NSF grant SES-1227280 and by the INET grant #INO1100049, "Understanding Macroeconomic Fragility." H.U. has an ongoing consulting relationship with a Federal Reserve Bank, the Bundesbank, and the ECB. V.G. has an ongoing consulting relationship with the Chicago Federal Reserve Bank.

REFERENCES

Abreu, D., Brunnermeier, M.K., 2003. Bubbles and crashes. Econometrica 71 (1), 173–204.

Adam, K., Woodford, M., 2012. Housing prices and robustly optimal monetary policy. J. Monet. Econ. 59, 468–487.

Anenberg, E., 2014. Information frictions and housing market dynamics. Working Paper.

Attanasio, O.P., Blow, L., Hamilton, R., Leicester, A., 2009. Booms and busts: consumption, house prices and expectations. Economica 76 (301), 20–50.

Balasko, Y., Shell, K., 1980. The overlapping generations model, I: the case of pure exchange without money. J. Econ. Theory 23 (3), 281–306.

Barberis, N., Shleifer, A., Vishny, R., 1998. A model of investor sentiment. J. Financ. Econ. 49 (3), 307–343.

Benes, J., Kumhof, M., Laxton, D., 2014a. Financial crises in DSGE models: a prototype model. IMF Working Paper Series 14/57.

Benes, J., Kumhof, M., Laxton, D., 2014b. Financial crises in DSGE models: selected applications of MAPMOD. IMF Working Paper Series 14/56.

Benhabib, J., Rogerson, R., Wright, R., 1991. Homework in macroeconomics: household production and aggregate fluctuations. J. Polit. Econ. 99, 1166–1187.

Berger, D., Guerrieri, V., Lorenzoni, G., Vavra, J., 2015. House prices and consumer spending. National Bureau of Economic Research.

Bernanke, B., Gertler, M., 1989. Agency costs, net worth, and business fluctuations. Am. Econ. Rev. 79, 14–31.

Bernanke, B.S., Gertler, M., Gilchrist, S., 1999. The financial accelerator in a quantitative business cycle framework. In: Taylor, J., Woodford, M. (Eds.), Handbook of Macroeconomics, vol. 1. Elsevier, North-Holland, pp. 1341–1393.

Bocola, L., 2014. The Pass-Through of Sovereign Risk (draft). Northwestern University.

Boissay, F., Collard, F., Smets, F., 2016. Booms and banking crises. J. Polit. Econ. 124 (2), 489–538.

Boldrin, M., Christiano, L.J., Fisher, J.D.M., 2001. Habit persistence, asset returns, and the business cycle. 91 (1), 149–166.

Bordalo, P., Gennaioli, N., Shleifer, A., 2016. Diagnostic expectations and credit cycles. Working Paper.

Brunnermeier, M.K., Sannikov, Y., 2010. A Macroeconomic Model with a Financial Sector (draft). Princeton University.

Brunnermeier, M., Eisenbach, T.M., Sannikov, Y., 2011. Macroeconomics with Financial Frictions: A Survey (draft). Princeton University.

Brzoza-Brzezina, M., Gelain, P., Kolasa, M., 2014. Monetary and Macroprudential Policy with Multi-Period Loans (draft).

Burnside, C., Eichenbaum, M., Rebelo, S., 2013. Understanding booms and busts in housing markets. Working Paper, Northwestern University.

Campbell, J.Y., Cocco, J.F., 2007. How do house prices affect consumption? Evidence from micro data. J. Monet. Econ. 54 (3), 591–621.

Campbell, J.R., Hercowitz, Z., 2006. The role of collateralized household debt in macroeconomic stabilization. Working Paper.

Capozza, D.R., Hendershott, P.H., Mack, C., 2004. An anatomy of price dynamics in illiquid markets: analysis and evidence from local housing markets. Real Estate Econ. 32 (1), 1–32.

Carlstrom, C.T., Fuerst, T.S., 1997. Agency costs, net worth, and business fluctuations: a computable general equilibrium analysis. Am. Econ. Rev. 87 (5), 893–910.

Carroll, C.D., Otsuka, M., Slacalek, J., 2011. How large are housing and financial wealth effects? A new approach. J. Money Credit Bank. 43 (1), 55–79.

Carvalho, V.M., Martin, A., Ventura, J., 2012. Understanding bubbly episodes. Am. Econ. Rev. 102 (3), 95–100.

Cass, D., 1972. On capital overaccumulation in the aggregative neoclassical model of economic growth: a complete characterization. J. Econ. Theory 4 (2), 200–223.

Case, K., Shiller, R., 1989. The efficiency of the market for single-family homes. Am. Econ. Rev. 79 (1), 125–137.

Case, K.E., Quigley, J.M., Shiller, R.J., et al., 2013. Wealth effects revisited 1975–2012. Crit. Finance Rev. 2 (1), 101–128.

Clancy, D., Merola, R., 2015. Counter-cyclical capital rules for small open economies. Working Paper.

Corbae, D., Quintin, E., 2014. Leverage and the foreclosure crisis.

Davis, M.A., Heathcote, J., 2005. Housing and the business cycle. Int. Econ. Rev. 46 (3), 751–784.

Davis, M.A., VanNieuwerburgh, S., 2015. Housing, finance and the macroeconomy. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), Handbook of Urban and Regional Economics, vol. 5. Elsevier, pp. 753–811.

Del Negro, M., Eggertsson, G., Ferrero, A., Kiyotaki, N., 2011. The great escape? A quantitative evaluation of the Fed's liquidity facilities. Staff Report 520, Federal Reserve Bank of New York.

Di Maggio, M., Kermani, A., 2016. Credit induced boom and bust. Working Paper.

Eggertsson, G.B., Krugman, P., 2012. Debt, deleveraging, and the liquidity trap: a Fisher-Minsky-Koo approach. Q. J. Econ. 127 (3), 1469–1513.

Favara, G., Imbs, J., 2015. Credit supply and the price of housing. Am. Econ. Rev. 105, 958–992.

Favilukis, J., Ludvigson, S., Nieuwerburgh, S.V., 2016. The macroeconomic effects of housing wealth, housing finance, and limited risk-sharing in general equilibrium. J. Polit. Econ. Forthcoming.

Fisher, J.D., 2007. Why does household investment lead business investment over the business cycle? J. Polit. Econ. 115 (1), 141–168.

Frazzini, A., 2006. The disposition effect and underreaction to news. J. Finance 61 (4), 2017–2046.

Garriga, C., Kydland, F.E., Šustek, R., 2016. Mortgages and monetary policy. Working Paper.

Geanakoplos, J., 2002. Liquidity, default and crashes: endogenous contracts in general equilibrium. Discussion Paper 1316RR, Cowles Foundation for Research in Economic – Yale University.

Geanakoplos, J., 2009. The leverage cycle. Discussion Paper 1715, Cowles Foundation for Research in Economic – Yale University.

Geanakoplos, J., 2011. What's missing from macroeconomics: endogenous leverage and default. Discussion Paper 1332, Cowles Foundation for Research in Economic – Yale University.

Gilchrist, S., Zakrajšek, E., 2012. Credit spreads and business cycle fluctuations. Am. Econ. Rev. 102 (4), 1692–1720.

Glaeser, E.L., Nathanson, C.G., 2016. An extrapolative model of house price dynamics.

Glaeser, E.L., Gyourko, J., Morales, E., Nathanson, C.G., 2014. Housing dynamics: an urban approach. J. Urban Econ. 81, 45–56.

Golosov, M., Lorenzoni, G., Tsyvinski, A., 2014. Decentralized trading with private information. Econometrica 82 (3), 1055–1091.

Greenwood, J., Hercowitz, Z., 1991. The allocation of capital and time over the business cycle. J. Polit. Econ. 99 (6), 1188–1214.

Guerrieri, L., Iacoviello, M., 2014. Collateral constraints and macroeconomic asymmetries.

Guerrieri, V., Lorenzoni, G., 2011. Credit crises, precautionary savings, and the liquidity trap. NBER Working Paper Series 17583.

Guren, A.M., 2016. The causes and consequences of house price momentum.

Hall, R., 2011. The long slump. Am. Econ. Rev. 101 (2), 431–469.

Hall, R.E., 2014. High discounts and high unemployment. National Bureau of Economic Research.

Harrison, J.M., Kreps, D.M., 1978. Speculative investor behavior in a stock market with heterogeneous expectations. Q. J. Econ. 92 (2), 323–336.

He, Z., Krishnamurthy, A., 2013. Intermediary asset pricing. Am. Econ. Rev. 103 (2), 732–770.

He, C., Wright, R., Zhu, Y., 2015. Housing and liquidity. Rev. Econ. Dyn. 18 (3), 435–455.

Head, A., Lloyd-Ellis, H., Sun, H., 2014. Search, liquidity, and the dynamics of house prices and construction. Am. Econ. Rev. 104 (4), 1172–1210.

Hong, H., Stein, J.C., 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. J. Finance 54 (6), 2143–2184.

Huo, Z., Ríos-Rull, J.V., 2013. Paradox of thrift recessions. National Bureau of Economic Research.

Huo, Z., Ríos-Rull, J.V., 2014. Financial frictions, asset prices, and the great recession.

Iacoviello, M., 2005. House prices, borrowing constraints, and monetary policy in the business cycle. Am. Econ. Rev. 95 (3), 739–764.

Iacoviello, M., 2012. Housing wealth and consumption. In: Smith, S. (Ed.), International Encyclopedia of Housing and Home. Elsevier, pp. 673–678.

Iacoviello, M., Neri, S., 2010. Housing market spillovers: evidence from an estimated DSGE model. Am. Econ. J. Macroecon. 2 (2), 125.

Iacoviello, M., Pavan, M., 2013. Housing and debt over the life cycle and over the business cycle. J. Monet. Econ. 60, 221–238.

Johnson, D.S., Parker, J.A., Souleles, N.S., 2006. Household expenditure and the income tax rebates of 2001. Am. Econ. Rev. 96 (5), 1589–1610.

Jorda, O., Schularick, M., Taylor, A.M., 2013. When credit bites back. J. Money Credit Bank. 45 (2), 3–28.

Jorda, O., Schularick, M., Taylor, A.M., 2015. Betting the house. J. Int. Econ. 96, S2–S18.

Jorda, O., Schularick, M., Taylor, A.M., 2016a. The great mortgaging: housing finance, crises and business cycles. Econ. Policy 31, 107–152.

Jorda, O., Schularick, M., Taylor, A.M., 2016b. Macrofinancial history and the new business cycle facts. In: Eichenbaum, M., Parker, J. (Eds.), NBER Macroeconomics Annual 2016. University of Chicago Press, Chicago, Il., U.S.A.

Justiniano, A., Primiceri, G.E., Tambalotti, A., 2014. Credit supply and the housing boom. Working Paper.

Justiniano, A., Primiceri, G.E., Tambalotti, A., 2015. Household leveraging and deleveraging. Rev. Econ. Dyn. 18 (1), 3–20.

Kaplan, G., Violante, G.L., 2014. A model of the consumption response to fiscal stimulus payments. Econometrica 82 (4), 1199–1239.

Kaplan, G., Mitman, K., Violante, G., 2015. Consumption and house prices in the great recession: model meets evidence. Working Paper.

Kehoe, P., Midrigan, V., Pastorino, E., 2014. Debt constraint and unemployment. Working Paper.

Kermani, A., 2016. Cheap credit, collateral and the boom-bust cycle. Working Paper.

Kiyotaki, N., Moore, J., 1997. Credit cycles. J. Polit. Econ. 105 (2), 211–248.

Kiyotaki, N., Wright, R., 1989. On money as a medium of exchange. J. Polit. Econ. 97, 927–954.

Kiyotaki, N., Wright, R., 1993. A search-theoretic approach to monetary economics. Am. Econ. Rev. 83 (1), 63–77.

Kiyotaki, N., Michaelides, A., Nikolov, K., 2011. Winners and losers in housing markets. J. Money Credit Bank. 43, 255–296.

Knoll, K., 2016. Return Predictability in International Housing Markets, 1870–2014 (Dissertation draft). University of Bonn.

Knoll, K., Schularick, M., Steger, T., 2014. No price like home: global house prices, 1870-2012. Working Paper, University of Bonn.

Lorenzoni, G., 2008. Inefficient credit booms. Rev. Econ. Stud. 75, 809–833.

Macera, M., 2015. Credit crises and private deleveraging. Working Paper.

Mack, A., Martínez-García, E., 2011. A cross-country quarterly database of real house prices: a methodological note. Working Paper 99, Federal Reserve Bank of Dallas Globalization and Monetary Policy Institute.

Magnus, G., 2011. The dynamics of prices, liquidity and vacancies in the housing market (Dissertation). University of Chicago.

Mankiw, N.G., 1986. The allocation of credit and financial collapse. Q. J. Econ. 101 (3), 455–470.

Martin, A., Ventura, J., 2010. Theoretical notes on bubbles and the current crisis. NBER Working Paper 16399, National Bureau of Economic Research.

Martin, A., Ventura, J., 2012. Economic growth with bubbles. Am. Econ. Rev. 102 (6), 3033–3058.

Martin, A., Ventura, J., 2014. Managing credit bubbles. NBER Working Paper 19960, National Bureau of Economic Research.

Mendoza, E.G., Quadrini, V., 2010. Financial globalization, financial crises and contagion. J. Monet. Econ. 57, 24–39.

Mian, A.R., Sufi, A., 2011. House prices, home equity-based borrowing, and the U.S. household leverage crisis. Am. Econ. Rev. 101 (5), 2132–2156.

Mian, A., Sufi, A., 2014. What explains the 2007–2009 drop in employment? Econometrica 82 (6), 2197–2223.

Mian, A.R., Rao, K., Sufi, A., 2013. Household Balance Sheets, Consumption, and the Economic Slump (draft). University of Chicago Booth School.

Mian, A., Sufi, A., Verner, E., 2015. Household debt and business cycles worldwide. NBER Working Papers 21581, National Bureau of Economic Research.

Midrigan, V., Philippon, T., 2011. Household leverage and the recession. NYU Working Paper.

Mondragon, J., Wieland, J., Yang, M.J., 2016. Credit supply shocks and house price boom-bust cycles. Working Paper.

Myerson, R., 2012. A model of moral-hazard credit cycles. J. Polit. Econ. 120 (5), 847–878.

Ortalo-Magne, F., Rady, S., 2006. Housing market dynamics: on the contribution of income shocks and credit constraints. Rev. Econ. Stud. 73 (2), 459–485.

Parker, J.A., Souleles, N.S., Johnson, D.S., McClelland, R., 2013. Consumer spending and the economic stimulus payments of 2008. Am. Econ. Rev. 103 (6), 2530–2553.

Rognlie, M., Shleifer, A., Simsek, A., 2015. Investment hangover and the great recession. Working Paper.

Samuelson, P.A., 1958. An exact consumption-loan model of interest with or without the social contrivance of money. J. Polit. Econ. 66 (6), 467–482.

Scheinkman, J.A., Xiong, W., 2003. Overconfidence and speculative bubbles. J. Polit. Econ. 111 (6), 1183–1220.

Schularick, M., Taylor, A., 2012. Credit booms gone bust: monetary policy. Leverage cycles, and financial crises, 1870-2008. Am. Econ. Rev. 102 (2), 1029–1061.

Simsek, A., 2013. Belief disagreements and collateral constraints. Econometrica 81, 1–53.

Sommer, K., Sullivan, P., Verbrugge, R., 2013. The equilibrium effect of fundamentals on house prices and rents. J. Monet. Econ. 60, 854–870.

Stein, J.C., 1995. Prices and trading volume in the housing market: a model with downpayment effects. Q. J. Econ. 110 (2), 379–406.

Ströbel, J., Vavra, J., 2015. House prices, local demand, and retail prices. Working Paper.

Tirole, J., 1985. Asset bubbles and overlapping generations. Econometrica 53, 1071–1100.

Woodford, M., 2010. Robustly optimal monetary policy with near-rational expectations. Am. Econ. Rev. 100, 274–303.

Wright, R., Wong, Y.Y., 2014. Buyers, sellers, and middlemen: variations on search-theoretic themes. Int. Econ. Rev. 55 (2), 375–397.

CHAPTER 18

Macro, Money, and Finance: A Continuous-Time Approach

M.K. Brunnermeier, Y. Sannikov
Princeton University, Princeton, NJ, United States

Contents

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.06.002

Abstract

This chapter puts forward a manual for how to setup and solve a continuous time model that allows to analyze endogenous (1) level and risk dynamics. The latter includes (2) tail risk and crisis probability as well as (3) the Volatility Paradox. Concepts such as (4) illiquidity and liquidity mismatch, (5) endogenous leverage, (6) the Paradox of Prudence, (7) undercapitalized sectors (8) time-varying risk premia, and (9) the external funding premium are part of the analysis. Financial frictions also give rise to an endogenous (10) value of money.

Keywords

Macroeconomic modeling, Monetary Economics, (Inside) Money, Endogenous Risk Dynamics, Volatility paradox, Paradox of Prudence, Financial frictions

JEL Classification Codes:

C63, E32, E41, E44, E51, G01, G11, G20

1. INTRODUCTION

The recent financial crisis in the United States and the subsequent Euro Crisis are vivid reminders of the importance of financial frictions in understanding macroeconomic trends and cycles. While financial markets are self-stabilizing in normal times, economies become vulnerable to a crisis after a run up of (debt) imbalances and (credit) bubbles. In particular, debt, leverage, maturity and liquidity mismatch tend to rise when measured volatility is low. Vulnerability risk tends to build up in the background, and only materializes when crises erupt, a phenomenon referred to as the "Volatility Paradox."

Adverse feedback loops can make the market spiral out of balance. The dynamics of an economy with financial frictions are highly nonlinear. Small shocks lead to large economic dislocations. In situations with multiple equilibria, runs on financial institutions or sudden stops on countries can occur even absent any fundamental trigger. Empirically, these phenomena show up as fat tails in the distribution of real economic variables and asset price returns.

Our research proposes a continuous time method to capture the whole *endogenous risk dynamics* and hence goes beyond studying simply the persistence and amplification of an individual adverse shock. Instead of focusing only on levels, the first moments, the second moments, and movements of risk variables are all an integral part of the analysis, as they drive agents' consumption, (precautionary) savings and investment decisions. After a negative shock, we do not assume that the economy returns to the steady state deterministically, but rather uncertainty might be heightened making the length of the slump stochastic. As agents respond to the new situation, they affect both the risk and the risk premia.

Endogenous risk is time-varying and depends on illiquidity. Liquidity comes in three flavors. Technological illiquidity refers to the irreversibility of physical investment.

Instead of undoing the initial investment, another option is to sell off the investment. This is only reasonable when market liquidity is sufficiently high. Finally, with sufficient funding liquidity one can issue claims against the payoffs of the assets. Incentive problems dictate that these claims are typically short-term debt claims. Debt comes with the drawback that risk is concentrated in the indebted sector. In addition, short-term debt leads to liquidity risk exposure. Agents may be forced to fire-sell their assets if they cannot undo the investment, market liquidity is low and funding is restricted, eg, very short term. In short, when there is a liquidity mismatch between technological and market liquidity on the asset side and funding liquidity on the liability side of the balance sheet, the economy is vulnerable to instability.

Models with financial frictions necessarily have to encompass multiple sectors. Financial frictions prevent funds from flowing to undercapitalized sectors, create debt overhang problems, and/or preclude optimal ex-ante risk sharing. This is in contrast to a world with perfect financial markets in which only aggregate risk matters, as all agents' marginal rate of substitutions are equalized in equilibrium and consequently aggregation to a single representative agent is possible. In models with financial frictions and heterogeneous agents the wealth distribution matters.

Importantly, financial frictions also give rise to the value of money. Money is a liquid store of value and safe asset. This approach provides not only a complementary perspective to New Keynesian models, in which price and wage rigidities are the primary drivers of money value, but also enables the revival of the traditional literature on "money and banking."[a]

Ultimately, economic analysis should guide policy. It is important to go beyond partial equilibrium analysis since general equilibrium effects can be subtle and counterintuitive. A model has to be tractable enough to conduct a meaningful welfare analysis to evaluate various policy instruments. A welfare analysis lends itself to study the interaction of various policy instruments.

In sum, the goal of this chapter is to put forward a manual for how to setup and solve a continuous time macrofinance model. The tractability that continuous time offers allows us to study a host of new properties of fully solved equilibria. This includes the full characterization of endogenous (1) level and risk dynamics. The latter includes (2) tail risk and crisis probability as well as (3) the Volatility Paradox. In addition, it should help us think about (4) illiquidity and liquidity mismatch, (5) endogenous leverage, (6) Paradox of Prudence, (7) undercapitalized sectors, (8) time-varying risk premia, and (9) the external funding premium. From a welfare perspective, we would like to ask normative questions about the (10) inefficiencies of financial crises and (11) the effects of policies using various instruments.

[a] See, eg, Chandler (1948).

We start with a brief history of macro and finance research since the Great Depression in the 1930s. We then put forward arguments in favor of continuous time modeling before surveying the ongoing continuous time literature. The main part of the paper builds up a step by step outline how to solve continuous time models starting with the simplest benchmark and enriching the model by adding more building blocks.

1.1 A Brief History of Macroeconomics and Finance

Macroeconomics as a field in economics was born during the great depression in the 1930s. At that time, economists like Fisher (1933), Keynes (1936), Gurley and Shaw (1955), Minsky (1957), and Kindleberger (1978) stressed the importance of the interaction between financial instability and macroeconomic aggregates. In particular, certain sectors in the economy including the financial sector can become balance sheet impaired and can drag down parts of the economy. Patinkin (1956) and Tobin (1969) also emphasized that financial stability and price stability are intertwined and hence that macroeconomics, monetary economics and finance are closely linked.

As economics became more analytical and model based, macroeconomics and finance went into different directions. See Fig. 1. Hicks' (1937) IS-LM Keynesian macro model is both static and deterministic. Macroeconomic growth models, most prominently the Solow (1956) growth model, are dynamic and many of them are in continuous time. However, they exclude stochastic elements: risk and uncertainty play no role. In contrast, the formal finance literature starting with Markowitz (1952) portfolio theory focused exclusively on risk. These models are static models and ignore the time dimension.

In the 1970s and early 1980s macroeconomists introduced stochastic elements into their dynamic models. Early "fresh water" models that included time and stochastic elements were Brock and Mirman's (1972) stochastic growth model and real business cycle models à la Kydland and Prescott (1982). The influential graduate text book of

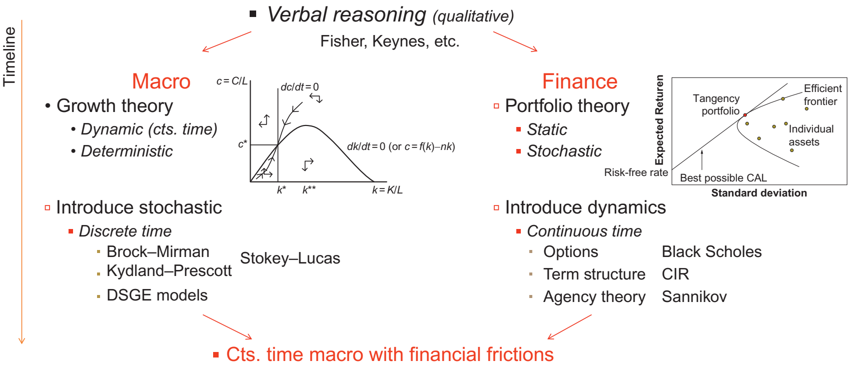


Fig. 1 Methods in macroeconomic and financial research since the great depression.

Stokey and Lucas (1989) provided the necessary toolkit for a fully microfounded dynamic and stochastic analysis. The "salt water" New Keynesian branch of macro introduced price rigidities and studied countercyclical policy in rational expectations models, Taylor (1979) and Mankiw and Romer (1991). The two branches merged and developed DSGE models which were both dynamic, the D in DSGE, and stochastic, the S in DSGE. However, unlike in many of the earlier growth models, time is discrete in real business cycle and New Keynesian DSGE models à la Woodford (2003). Most DSGE models capture only the log-linearized dynamics around the steady state. The log-linearized theoretical analysis squared nicely with its empirical counterpart, the linear Vector Autoregression Regression (VAR) estimation technique pioneered by Sims (1980).

Finance also experienced great breakthroughs in the 1970s. Stochastic Calculus (Ito calculus), which underlies the Black and Scholes (1973) option pricing model, revolutionized finance. Besides option pricing, term structure of interest rate models like Cox et al. (1985) were developed. More recently, Sannikov (2008) developed continuous time tools for financial contracting, which allow one to capture contracting frictions in a tractable way.

Our line of research is the next natural step. It essentially merges macroeconomics and finance using continuous time stochastic models. In terms of financial frictions, it builds on earlier work by Bernanke et al. (1999) (BGG), Kiyotaki and Moore (1997) (KM), Bianchi (2011), Mendoza (2010), and others. Our approach replicates two important results from the linearized versions of classic models of BGG and KM, that (1) temporary macro shocks can have a *persistent* effect on economic activity by making borrowers "undercapitalized" and (2) price movements *amplify* shocks. In KM, the leverage is limited by an always binding collateral constraint. In Bianchi (2011) and Mendoza (2010) it is occasionally binding. Our approach focuses mostly on incomplete market frictions, where the leverage of potentially undercapitalized borrowers is usually endogenous. In particular, it responds to the magnitude of fundamental (exogenous) macroeconomic shocks and the level of financial innovations that enable better risk management. Interestingly, leverage responds to a much lesser extent to the presence of endogenous tail risk. Equilibrium leverage in normal times is a key determinant of the probability of crises.

1.2 The Case for Continuous-Time Macro Models

As economists we have no hesitation in assuming a continuous action space in order to ensure nice first order optimality conditions that are free of integer problems. In the same vein, we typically assume a continuum of agents to guarantee an environment with perfect competition and (tractable) price taking behavior.

Assuming a continuous time framework has two advantages: it is often more tractable and might conceptually be a closer representation of reality. In terms of tractability, continuous time allows one to derive more analytical steps and more closed form characterizations of the equilibrium before resorting to a numerical analysis. For example, in our case one can derive explicit closed form expressions for amplification terms. The reason is that only the slope of the price function, ie, the (local) derivative w.r.t. state variables, is necessary to characterize amplification. In contrast, in discrete time settings the whole price function is needed, as the jump size may vary. Also, instantaneous returns are essentially log-normal, which makes it easy to take expectations. It is also easy to derive the portfolio choice problem and to link returns to net worth dynamics via the budget constraint. In discrete-time models, this feature can only be achieved through a (Campbell–Shiller) log-linear approximation. It is therefore not surprising that the term structure literature uses continuous time models. Admittedly, some of these features are due to the continuous nature of certain stochastic processes, like Brownian Motions and other Ito Processes. Hereby, one implicitly assumes that agents can adjust their consumption or portfolio continuously as their wealth changes. The feature that their wealth never jumps beyond a specific point, eg, the insolvency point, greatly simplifies the exposition.

Conceptually, in certain dimensions a continuous time representation might also square better with reality. People do not consume only at the end of the quarter, even though data come in quarterly. Discrete time models implicitly assume linear time aggregation within a quarter and a nonlinear one across quarters. In other words, the intertemporal elasticity of consumption within a quarter is infinite while across quarters it is given by the curvature of the utility function. Continuous time models treat every time unit the same. Similarly, it is well known that for multivariate models mixing data with different degrees of smoothness and frequency (such as consumption data and financial data) can seriously impair inference.

The biggest advantage of our continuous-time approach is that it allows a full characterization of the whole dynamical system including the risk dynamics instead of simply a log-linearized representation around the steady state. Note that impulse response functions capture only the expected path after a shock that starts at the steady state. Also, the stationary distribution can be bimodal and exhibit large swings, unlike stable normal distributions that log-linearized models imply.

1.3 The Nascent Continuous-Time Macrofinance Literature

This chapter builds on Brunnermeier and Sannikov (2014).[b] It extends this work by allowing for more general utility functions, precautionary savings and for endogenous equity issuance. Work by Basak and Cuoco (1998) and He and Krishnamurthy (2012, 2013) on intermediary asset pricing are part of the core papers in this literature.

[b] For an alternative survey on continuous time macro models, see, eg, Isohätälä et al. (2016).

Isohätälä et al. (2014) study a partial equilibrium model. DiTella (2013) introduces exogenous uncertainty shocks that can lead to balance sheet recessions even when contracting based on aggregate state variables is possible.

Phelan (2014) considers a setting in which banks issue equity and leverage can be procyclical. Adrian and Boyarchenko (2012) achieve procyclical leverage by introducing liquidity preference shocks. Adrian and Boyarchenko (2013) consider the interaction between two types of intermediaries: banks and nonbanks. Huang (2014) studies shadow banks, which circumvent regulatory constraints but are subject to an endogenous enforcement constraint. In Moreira and Savov (2016)'s macro model shadow banks issue money-like claims. In downturns they scale back their activity. This slows down the recovery and creates a scarcity in collateral. Klimenko et al. (2015) show that regulation that prohibits dividend payouts is typically superior to very tight capital requirements. In Moll (2014) capital is misallocated since productive agents are limited by collateral constraints to lever up.

Several papers also tried to calibrate continuous time macrofinance models to recent events. For example, He and Krishnamurthy (2014) do so by including housing as a second form of capital. Mittnik and Semmler (2013) employ a multi-regime vector autoregression approach to capture the nonlinearity of these models.[c]

In international economics, these methods are employed in Brunnermeier and Sannikov (2015b). In a two-good, two-country model, the overly indebted country is vulnerable to sudden stops, and hence capital controls might improve welfare. Maggiori (2013) models risk sharing across countries which are at different stages of financial development.

Models with financial frictions also open up an avenue for new models in monetary economics thereby reviving the field "money and banking." In Brunnermeier and Sannikov (2015a)'s "The I Theory of Money" money is a bubble like in Samuelson (1958) or Bewley (1977). Inside money is created endogenously by the intermediary sector, and monetary policy and macroprudential policy interact. Achdou et al. (2015) provide a solution algorithm for Bewley models with uninsurable endowment risk in a continuous time setting. In Drechsler et al. (2016) banks are less risk averse and monetary policy affects risk premia. Silva (2016) studies how unconventional monetary policy reallocates risk. Werning (2012) studies the zero lower bound problem in a tractable deterministic continuous time New Keynesian model.

Rappoport and Walsh (2012) setup a discrete-time macro model, which has similar economic results, and which converges in the continuous-time limit to the model of Brunnermeier and Sannikov (2014).

[c] Note that in the estimation of DSGE models, Fernandez-Villaverde and Rubio-Ramirez (2010) show that parameter estimates and the moments generated by the model depend quite sensitively on whether a linearized DSGE is estimated via Kalman filtering or whether the true DSGE model is estimated via particle filtering.

2. A SIMPLE REAL ECONOMY MODEL

We start first with a particularly simple model to illustrate how equilibrium conditions—utility maximization and market clearing—translate into an equilibrium characterization. This simple model trivializes most of the issues we are after, eg, the model has no price effects or endogenous risk. We do get some interesting takeaways, such as that risk premia spike in crises. After establishing the conceptual framework for what an equilibrium is, we move on to tackle more complex models.

2.1 Model Setup

This model is a variation of Basak and Cuoco (1998). The economy has a risky asset in positive net supply and a risk-free asset in zero net supply. There are two types of agents—experts and households. Only experts can hold the risky asset—households can only lend to experts at the risk-free rate r_t, determined endogenously in equilibrium. The friction is that experts can finance their holdings of the risky asset only through debt—by selling short the risk-free asset to households. That is, experts cannot issue equity. We assume that all agents are small, and behave as price-takers. That is, unlike in microstructure models with noise traders, agents have no price impact.

2.1.1 Technology

Net of investment, physical capital, k_t, generates consumption output at the rate of

$$(a-\iota_t)k_t\,dt,$$

where a is a productivity parameter and ι_t is the reinvestment rate per unit of capital. The production technology is constant returns to scale.

The productive asset (capital), k_t, evolves according to

$$\frac{dk_t}{k_t}=(\Phi(\iota_t)-\delta)dt+\sigma\,dZ_t, \tag{1}$$

where $\Phi(\iota_t)$ is an investment function with adjustment costs, such that $\Phi(0)=0$, $\Phi'>0$ and $\Phi''\leq 0$. Thus, in the absence of investment, capital simply depreciates at rate δ. The concavity of $\Phi(\cdot)$ reflects decreasing returns to scale, and for negative values of ι_t, corresponds to *technological illiquidity*—the marginal cost of capital depends on the rate of investment/disinvestment.

The aggregate amount of capital is denoted by K_t, and q_t is the price of capital. Hence, the aggregate net worth in the economy is q_tK_t. If N_t is the aggregate net worth of experts, then the aggregate net worth of households is $q_tK_t - N_t$.

Experts' wealth share is denoted by

$$\eta_t=\frac{N_t}{q_tK_t}\in[0,1].$$

2.1.2 Preferences

For *tractability*, all agents are assumed to have logarithmic utility with discount rate ρ, of the form

$$E\left[\int_0^\infty e^{-\rho t}\log c_t dt\right],$$

where c_t is consumption at time t.

2.2 A Step-By-Step Approach

Definition An equilibrium is a map from histories of macro shocks $\{Z_s, s \leq t\}$ to the price of capital q_t, risk-free rate r_t, as well as asset holdings and consumption choices of all agents, such that

1. agents behave to maximize utility and

2. markets clear.

To find an equilibrium, we need to write down equations that the processes q_t, r_t, etc., have to satisfy, and that characterize how these processes evolve with the realizations of shocks Z. It will be convenient to express these relationships using a state variable. Here the relevant state variable, which describes the distribution of wealth, is the fraction of wealth owned by the experts, η_t. When η_t drops, experts become more balance sheet constrained.

We solve the equilibrium in three steps. First, we postulate some endogenous processes. As a second step, we use the equilibrium conditions, ie, utility maximization and market clearing, to write down restrictions q_t and r_t need to satisfy. In this simple model, we will be able to express q_t and r_t as functions of η_t in closed form. Third, we need to derive the law of motion of the state variable, the wealth share η_t.

Step 1: Postulate Equilibrium Processes. The first step is to postulate certain endogenous price processes. For example, suppose that the price per unit of capital q_t follows an Ito process

$$\frac{dq_t}{q_t} = \mu_t^q dt + \sigma_t^q\, dZ_t, \tag{2}$$

which, of course, is endogenous in equilibrium.

An investment in capital generates, in addition to the dividend rate $(a - \iota)k_t dt$, the capital gains at rate

$$\frac{d(k_t q_t)}{k_t q_t}.$$

Ito's Lemma for the product of two stochastic processes can be used to derive this process.

Ito's Formula for Product

Suppose two processes X_t and Y_t follow

$$\frac{dX_t}{X_t} = \mu_t^X dt + \sigma_t^X dZ_t \quad \text{and} \quad \frac{dY_t}{Y_t} = \mu_t^Y dt + \sigma_t^Y dZ_t.$$

Then the product of two processes follows

$$\frac{d(X_t Y_t)}{X_t Y_t} = (\mu_t^X + \mu_t^Y + \sigma_t^X \sigma_t^Y)dt + (\sigma_t^X + \sigma_t^Y)dZ_t. \tag{3}$$

Using Ito's Lemma, the investment in capital generates capital gains at rate

$$\frac{d(k_t q_t)}{k_t q_t} = (\Phi(\iota_t) - \delta + \mu_t^q + \sigma\sigma_t^q)dt + (\sigma + \sigma_t^q)dZ_t.$$

Then capital earns the return of

$$dr_t^k = \underbrace{\frac{a - \iota_t}{q_t}dt}_{\text{dividend yield}} + \underbrace{(\Phi(\iota_t) - \delta + \mu_t^q + \sigma\sigma_t^q)dt + (\sigma + \sigma_t^q)dZ_t}_{\frac{d(k_t q_t)}{k_t q_t}, \text{ the capital gains rate}}. \tag{4}$$

Thus, generally a part of the risk from holding capital is fundamental, σdZ_t, and a part is endogenous, $\sigma_t^q dZ_t$.

Remarks

- For general utility functions one also has to postulate the stochastic discount factor process or equivalently a process for the marginal utility or the consumption process dc_t/c_t. For details see Section 3.1.
- Note that in monetary models like Brunnermeier and Sannikov (2015a, 2016) one also has to postulate a process p_t for the value of money which can be stochastic due to inflation risk. In Section 4 we present a simple monetary model.

Step 2: The Equilibrium Conditions.

Equilibrium conditions come in two flavors: Optimality conditions and market clearing conditions.

Optimal internal investment rate. Note that the rate of internal investment ι_t does not affect the risk of capital. The optimal investment rate that maximizes the expected return satisfies the first-order condition

$$\Phi'(\iota_t) = \frac{1}{q_t}. \tag{5}$$

Optimal consumption rate. Logarithmic utility has two convenient properties, which we derive formally for a more general case in Section 3.1. These two properties help reduce the number of equations that characterize equilibrium. First, for agents with log utility

$$\text{consumption} = \rho \cdot \text{net worth} \tag{6}$$

that is, they always consume a fixed fraction of wealth (permanent income) regardless of the risk-free rate or risky investment opportunities. The consumption Euler equation reduces to a particularly simple form.

Optimal portfolio choice. The optimal risk exposure of a log-utility agent in the optimal portfolio choice problem depends on the attractiveness of risky investment, measured by the Sharpe ratio, defined as expected excess returns divided by the standard deviation. Formally, the equilibrium condition is

$$\text{Sharpe ratio of risky investment} = \text{volatility of net worth}, \tag{7}$$

where the volatility is relative (measured as percentage change per unit of time).[d]

Goods Market clearing. We use Eqs. (6) and (7) to formalize equilibrium conditions, and characterize equilibrium. First, from condition (6), the aggregate consumption of all agents is $\rho q_t K_t$, and aggregate output is $(a - \iota(q_t))K_t$, where investment ι is an increasing function of q defined by (5). From market clearing for consumption goods, these must be equal, and so

$$\rho q_t = a - \iota(q_t). \tag{8}$$

This determines the equilibrium price of the risky capital. The aggregate consumption of experts must be $\rho N_t = \rho \eta_t q_t K_t$, and the aggregate consumption of households is $\rho(1 - \eta_t) q_t K_t$. Condition (8) alone leads to a *constant* value of the *price of capital* q. That is, $\mu_t^q = \sigma_t^q = 0$.

Example with Log Investment Function

Suppose the investment function takes the form

$$\Phi(\iota) = \frac{\log(\kappa\iota + 1)}{\kappa},$$

where κ is the adjustment cost parameter. Then $\Phi'(0) = 1$. Higher κ makes function Φ more concave, and as $\kappa \to 0, \Phi(\iota) \to \iota$, a fully elastic investment function with no adjustment costs. The optimal investment rate is $\iota = (q - 1)/\kappa$, and the market-clearing condition (8) leads to the price of

$$q = \frac{1 + \kappa a}{1 + \kappa \rho}.$$

The price converges to 1 as $\kappa \to 0$, ie, the investment technology is fully elastic. The price q converges to a/ρ as $\kappa \to \infty$.

[d] For example, if the annual volatility of S&P 500 is 15% and the risk premium is 3% (so that the Sharpe ratio is 3%/15% = 0.2), then a log utility agent wants to hold a portfolio with volatility 0.2 = 20%. This corresponds to a weight of 1.33 on S&P 500, and −0.33 on the risk-free asset.

Second, we can use condition (7) for experts to figure out the equilibrium *risk-free rate*. We first look at the return on risky and risk-free assets to compute the Sharpe ratio of risky investments. We then look at balance sheets of experts to compute the volatility of their wealth. Finally, we use Eq. (7) to get the risk-free rate.

Because q is constant, the risky asset earns a return of

$$dr_t^k = \underbrace{\frac{a-\iota}{q}dt}_{\rho,\ \text{dividend yield}} + \underbrace{(\Phi(\iota)-\delta)dt + \sigma\, dZ_t}_{\text{capital gains rate}},$$

and the risk-free asset earns r_t. Note that the dividend yield equals ρ by the goods market clearing condition. Hence, the Sharpe ratio of risky investment is

$$\frac{\rho + \Phi(\iota) - \delta - r_t}{\sigma}.$$

Note that since the price-dividend ratio is constant any change in the risk premium must come from the variation in the risk-free rate r_t.

Because experts must hold all the risky capital in the economy, with value q_tK_t (households cannot hold capital), and absorb risk through net worth N_t, the volatility of their net worth is

$$\frac{q_tK_t}{N_t}\sigma = \frac{\sigma}{\eta_t}.$$

Using (7),

$$\frac{\sigma}{\eta_t} = \frac{\rho + \Phi(\iota) - \delta - r_t}{\sigma} \quad \Rightarrow \quad r_t = \rho + \Phi(\iota) - \delta - \frac{\sigma^2}{\eta_t}. \tag{9}$$

Step 3: The Law of Motion of $\boldsymbol{\eta_t}$. To finish deriving the equilibrium, we need to describe how shocks Z_t affect the state variable $\eta_t = N_t/(q_tK_t)$. First, since η_t is a ratio, the following formula will be helpful for us:

Ito's Formula for Ratio

Suppose two processes X_t and Y_t follow

$$\frac{dX_t}{X_t} = \mu_t^X dt + \sigma_t^X dZ_t \quad \text{and} \quad \frac{dY_t}{Y_t} = \mu_t^Y dt + \sigma_t^Y dZ_t.$$

Then ratio of two processes follows

$$\frac{d(X_t/Y_t)}{X_t/Y_t} = (\mu_t^X - \mu_t^Y + (\sigma_t^Y)^2 - \sigma_t^X\sigma_t^Y)dt + (\sigma_t^X - \sigma_t^Y)dZ_t. \tag{10}$$

Second, it is convenient to express the laws of motion of the numerator and denominator of η_t in terms of total risk and the Sharpe ratio given by (9). Specifically,

$$\frac{dN_t}{N_t} = r_t dt + \underbrace{\frac{\sigma}{\eta_t}}_{\text{risk}} \underbrace{\frac{\sigma}{\eta_t}}_{\text{Sharpe}} dt + \frac{\sigma}{\eta_t} dZ_t - \underbrace{\rho dt}_{\text{consumption}} \quad \text{and}$$

$$\frac{d(q_t K_t)}{q_t K_t} = r_t dt + \underbrace{\sigma}_{\text{risk}} \underbrace{\frac{\sigma}{\eta_t}}_{\text{Sharpe}} dt + \sigma dZ_t - \underbrace{\rho dt}_{\text{dividend yield}} .$$

In the latter equation, we subtract the dividend yield from the total return on capital to obtain the capital gains rate.

Using the formula for the ratio,

$$\begin{aligned} \frac{d\eta_t}{\eta_t} &= (r_t + \sigma^2/\eta_t^2 - \rho - r_t - \sigma^2/\eta_t + \rho + \sigma^2 - \sigma^2/\eta_t) dt + (\sigma/\eta_t - \sigma) dZ_t \\ &= \frac{(1-\eta_t)^2}{\eta_t^2} \sigma^2 dt + \frac{1-\eta_t}{\eta_t} \sigma dZ_t. \end{aligned} \tag{11}$$

Step 4: Expressing $q(\eta)$ as a function of η is not necessary in this simple model, since q is a constant.

2.3 Observations

Several key observations about equilibrium characteristics are worth pointing out. Variable η_t fluctuates with macro shocks—a positive shock increases the wealth share of experts. This is because experts are levered. A negative shock erodes η_t, and experts require a higher risk premium to hold risky assets. Experts must be convinced to keep holding risky assets by the increasing Sharpe ratio

$$\frac{\sigma}{\eta_t} = \frac{\rho + \Phi(\iota) - \delta - r_t}{\sigma},$$

which goes to ∞ as η_t goes to 0. Strangely, this is achieved due to the risk-free rate $r_t = \rho + \Phi(\iota) - \delta - \sigma^2/\eta_t$ going to $-\infty$, rather than due to a depressed price of the risky asset, as illustrated in the top right panel of Fig. 2.

Because q_t is constant, as illustrated in the top left panel, there is no endogenous risk, no amplification and no volatility effects. Therefore, in this model, assumptions that allow for such a simple solution also eliminate any price effects that we are so interested in. We have to work harder to get those effects.

Besides the absence of price effects, in this model it is also the case that in the long run the expert sector becomes so large that it overwhelms the whole economy. To see this,

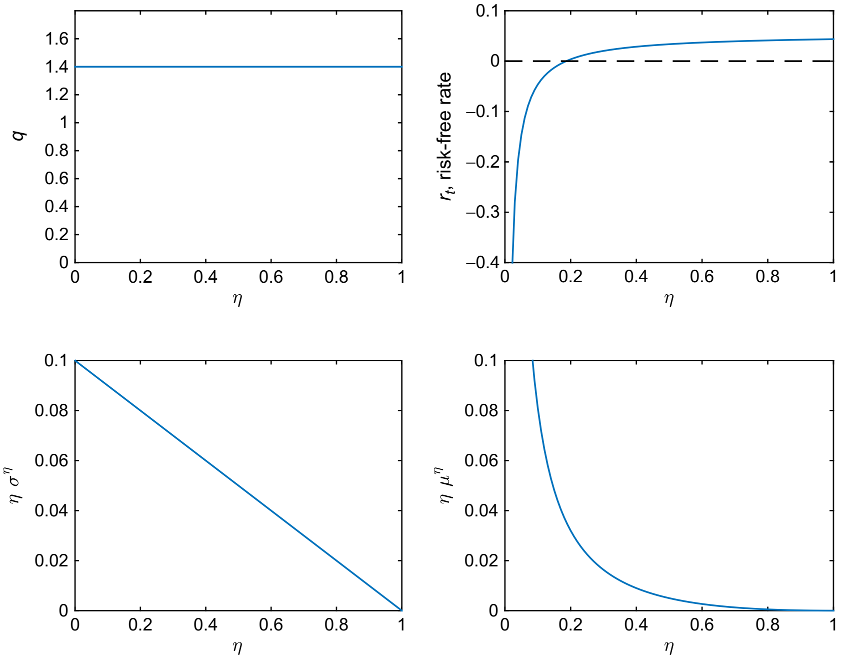


Fig. 2 Equilibrium in the simple real model, $a = 0.11$, $\rho = 5\%$, $\sigma = 0.1$, and $\Phi = \log(\kappa\iota + 1)/\kappa$ with $\kappa = 10$.

note that the drift of η_t is always positive. This feature is typical of models in which one group of agents has an advantage over another group—in this case only experts can invest in the risky asset. It is possible to prevent the expert sector from becoming too large through an additional assumption. For example, Bernanke et al. (1999) assume that experts are randomly hit by a shock that makes them households. Alternatively, if experts have a higher discount rate than households, then a greater consumption rate prevents the expert sector from becoming too large.

The main purpose of this section was to show how equilibrium conditions can be translated into formulas that describe the behavior of the economy. Next, we can consider more complicated models, in which the price of the risky asset q_t reacts to shocks. We also develop a methodology that allows for agents to have more complicated preferences and for a nontrivial distribution of assets among agents.

3. A MODEL WITH PRICE EFFECTS AND INSTABILITIES

We now illustrate how our step approach can be used to solve a more complex model, which we borrow and extend from Brunnermeier and Sannikov (2014). We will be able to get a number of important takeaways from the model:

1. Equilibrium dynamics are characterized by a relatively stable steady state, where the system spends most of the time, and a crisis regime. In the steady state, experts are adequately capitalized and risk premia fall. The experts' consumption offsets their earnings—hence the steady state is formed. Experts have the capacity to absorb most macro shocks, hence prices near the steady state are quite stable. However, an unusually long sequence of negative shocks causes experts to suffer significant losses, and pushes the equilibrium into a crisis regime. In the crisis regime, experts are undercapitalized and constrained. Shocks affect their demand for assets—market liquidity at the macro level can dry up—and thus affect prices of the assets that experts hold. This creates feedback effects, which generate fire-sales and *endogenous risk*. Volatility is endogenous and also feeds back in agents' behavior.
2. High volatility during crisis times may push the system into a very depressed region, where experts' net worth is close to 0. If that happens, it takes a long time for the economy to recover. Thus, the system spends a considerable amount of time far away from the steady state. The stationary distribution may be bimodal.
3. Endogenous risk during crises makes assets more correlated.
4. There is a "*volatility paradox*," because risk-taking is endogenous. If the aggregate risk parameter σ becomes smaller, the economy does not become more stable. The reason is that experts take on greater leverage, and pay out profits sooner, in response to lower fundamental risk. Due to greater leverage, the economy is prone to crises even when exogenous shocks are smaller. In fact, endogenous risk during crises may actually be higher when σ is lower.
5. Financial innovations, such as securitization and derivatives hedging, that allow for more efficient risk-sharing among experts, may make the system less stable in equilibrium. The reason, again, is that risk-taking is endogenous. By diversifying idiosyncratic risks, experts tend to increase leverage, amplifying systemic risks.

Before going into details of how we can extend our simple real economy model from Section 2 to display these additional features, we take a detour to discuss the classic problem of optimal consumption and portfolio choice in continuous time.

3.1 Optimal Portfolio Choice with General Utility Functions

We start with a brief description of how to extend the optimal consumption and portfolio choice conditions (such as (6) and (7)) to the case of a general utility function. The key result is that any asset, which an agent can hold, can be priced from the agent's marginal utility of wealth θ_t. The first-order condition for optimal consumption is $\theta_t = u'(c_t)$, so the marginal utility of wealth is also the marginal utility of consumption (unless the agent is "at the corner").[e]

[e] If the agent is risk-neutral, then his marginal utility of consumption is always 1, but the agent may choose to not consume if his marginal utility of wealth is greater than 1.

If the agent has discount rate ρ, then $\xi_t = e^{-\rho t}\theta_t$ is the stochastic discount factor (SDF) to price assets. We can write

$$\frac{d\xi_t}{\xi_t} = -r_t dt - \varsigma_t dZ_t, \tag{12}$$

where r_t is the (shadow) risk-free rate and ς_t is the price of risk dZ_t.

For any asset A that the agent can invest in, with return

$$dr_t^A = \mu_t^A dt + \sigma_t^A dZ_t,$$

we must have

$$\mu_t^A = r_t + \varsigma_t \sigma_t^A. \tag{13}$$

Eqs. (12) and (13) are simple, yet extremely powerful.

3.1.1 Martingale Method

To derive Eq. (13) consider a trading strategy of investing 1 dollar into asset A at time 0 and keep on reinvesting any dividends the asset might pay out. Denote the value of this strategy at time t by v_t (then $v_0 = 1$, obviously). Clearly, its capital gains rate is

$$\frac{dv_t}{v_t} = dr_t^A.$$

For an arbitrary $s \le t$ consider an investor who can only trade at s and t. That is, he faces a simple two-period portfolio problem. The Euler equation for the standard two-period portfolio problem is

$$v_s = E_s\left[\frac{\xi_t}{\xi_s} v_t\right] \Rightarrow \xi_s v_s = E_s[\xi_t v_t].$$

That is, $\xi_t v_t$ must be a martingale on the time domain $\{s, t\}$. For an investor who can trade continuously $\xi_t v_t$ must be a martingale for any t, since we picked s, t arbitrarily. Next, by Itô's formula

$$\frac{d(\xi_t v_t)}{\xi_t v_t} = (\mu_t^\xi + \mu_t^v + \sigma_t^\xi \sigma_t^v)dt + (\sigma_t^\xi + \sigma_t^v)dZ_t = (-r_t + \mu_t^A - \varsigma_t \sigma_t^A)dt + (\sigma_t^A - \varsigma_t)dZ_t.$$

This is a martingale if and only if the drift vanishes, ie, Eq. (13) holds.

3.1.2 Derivation via Stochastic Maximum Principle

One can also derive the pricing equations and consumption rule using the stochastic maximum principle. Let us consider an agent who maximizes

$$E\left[\int_0^\infty e^{-\rho t} u(c_t)dt\right],$$

and whose net worth follows

$$dn_t = n_t\left(r_t dt + \sum_A x_t^A((\mu_t^A - r_t)dt + \sigma_t^A dZ_t)\right) - c_t dt,$$

with initial wealth $n_0 > 0$ and where x_t^A are portfolio weights on various assets A. Investment opportunities are stochastic and exogenous, ie, they do not depend on the agent's strategy.

The stochastic maximum principle allows us to derive first-order conditions for maximization from the Hamiltonian. Introducing a multiplier ξ_t on n_t (ie, marginal utility of wealth) and denoting the volatility of ξ_t by $-\varsigma_t\xi_t$, the Hamiltonian is written as

$$H = e^{-\rho t}u(c) + \xi_t\{\underbrace{(r_t + \sum_A x^A(\mu_t^A - r_t))n_t - c}_{\text{drift of } n_t}\} - \varsigma_t\xi_t \underbrace{\sum_A x^A\sigma_t^A n_t}_{\text{volatility of } n_t}.$$

By differentiating the Hamiltonian with respect to controls, we get the first-order conditions, and by differentiating it with respect to the state n_t, we get the law of motion of the multiplier ξ_t.

The first-order condition with respect to c is

$$e^{-\rho t}u'(c_t) = \xi_t,$$

which implies that the multiplier on the agent's wealth is his discounted marginal utility of consumption. The first-order condition with respect to the portfolio weight x^A is

$$\xi_t(\mu_t^A - r_t) - \varsigma_t\xi_t\sigma_t^A = 0,$$

which implies (13).

In addition, the drift of ξ_t is

$$-H_n = -\xi_t r_t,$$

where we already used the first-order conditions with respect to x^A to perform cancellations. It follows that the law of motion of ξ_t is

$$d\xi_t = -\xi_t r_t\, dt - \varsigma_t\xi_t\, dZ_t,$$

which corresponds to (12).

3.1.3 Value Function Derivation for CRRA Utility

Macroeconomists are most familiar with this method. With CRRA utility, the agent's value function takes a power form

$$\frac{u(\omega_t n_t)}{\rho}. \tag{14}$$

This form comes from the fact that if the agent's wealth changes by a factor of x, then his optimal consumption at all future states changes by the same factor—hence ω_t is determined so that $u(\omega_t)/\rho$ is the value function at unit wealth. Marginal utility of consumption and marginal utility of wealth are equated if $c_t^{-\gamma} = \omega_t^{1-\gamma} n_t^{-\gamma}/\rho$, or

$$\frac{c_t}{n_t} = \rho^{1/\gamma}\omega_t^{1-1/\gamma}. \tag{15}$$

For log utility, $\gamma = 1$ and this equation implies that $c_t/n_t = \rho$ as we claimed in (6).

For $\gamma \neq 1$, by expressing ω_t as a function of the consumption rate c_t/n_t, we find that the agent's continuation utility is

$$\frac{c_t^{-\gamma} n_t}{1-\gamma}. \tag{16}$$

This remarkable expression shows that the agent's net worth and consumption rate are sufficient to compute the agent's welfare, and no additional information about the agent's stochastic investment opportunities is needed.

Given the agent's (postulated) consumption process of

$$\frac{dc_t}{c_t} = \mu_t^c dt + \sigma_t^c dZ_t,$$

by Ito's Lemma, marginal utility $c^{-\gamma}$ follows

$$\frac{d(c_t^{-\gamma})}{c_t^{-\gamma}} = \left(-\gamma\mu_t^c + \frac{\gamma(\gamma+1)}{2}(\sigma_t^c)^2\right)dt - \gamma\sigma_t^c dZ_t. \tag{17}$$

Substituting this into (13), we obtain the following relationship for the pricing of any risky asset relative to the risk-free asset:

$$\frac{\mu_t^A - r_t}{\sigma_t^A} = \gamma\sigma_t^c = \varsigma_t. \tag{18}$$

Recall that $\xi_t = e^{-\rho t}u'(c_t)$ and hence $\frac{d\xi_t}{\xi_t} = -\rho - \frac{d(c_t^{-\gamma})}{c_t^{-\gamma}}$. Minus the drift of the SDF is the risk-free rate, ie,

$$r_t = \rho + \gamma\mu_t^c - \frac{\gamma(\gamma+1)}{2}(\sigma_t^c)^2. \tag{19}$$

Two special cases with particularly nice analytical solutions deserve special attention.

Example with CRRA and Constant Investment Opportunities

With constant investment opportunities, then ω_t is a constant, hence (15) implies that $\sigma_t^c = \sigma_t^n$, just like in the logarithmic case. Hence, (18) implies that

$$\underbrace{\frac{\mu_t^A - r}{\sigma_t^A}}_{\varsigma} = \gamma\sigma_t^n,$$

ie, the volatility of net worth is the Sharpe ratio divided by the risk aversion coefficient γ. Note that this property also holds when ω_t is not a constant as long as it evolves deterministically.

Now, the agent's net worth follows

$$\frac{dn_t}{n_t} = rdt + \frac{\varsigma^2}{\gamma}dt + \frac{\varsigma}{\gamma}dZ_t - \frac{c_t}{n_t}dt,$$

and, since consumption is proportional to net worth, (19) implies that

$$r = \rho + \gamma\left(r + \frac{\varsigma^2}{\gamma} - \frac{c_t}{n_t}\right) - \frac{\gamma(\gamma+1)}{2}\frac{\varsigma^2}{\gamma^2} \quad \Rightarrow \quad \frac{c_t}{n_t} = \rho + \frac{\gamma - 1}{\gamma}\left(r - \rho + \frac{\varsigma^2}{2\gamma}\right).$$

Hence, consumption ratio increases with better investment opportunities when $\gamma > 1$ and falls otherwise.

Example with Log Utility

We can verify that the consumption and asset-pricing relationships for logarithmic utility of equation. Note from (15) follows directly (6),

$$c_t = \rho n_t.$$

Since the SDF is $\xi_t = e^{-\rho t}/c_t = e^{-\rho t}/(\rho n_t)$ (for any ω_t) it follows that $\sigma_t^n = \sigma_t^c = \varsigma_t$ (ie, minus the volatility of ξ_t). Hence, (13) implies that

$$\frac{\mu_t^A - r_t}{\sigma_t^A} = \sigma_t^n,$$

where the left hand side is the Sharpe ratio, and the right hand side is the volatility of net worth.

3.2 Model with Heterogeneous Productivity Levels and Preferences

In order to study endogenous risk, market illiquidity, fire-sales, etc., we now assume that the household sector can also hold physical capital, but households are assumed to be less productive. Specifically, their productivity parameter $\underline{a} < a$, and hence their willingness to pay for capital, is lower than that of experts. In this generalized setting, experts now have only two ways out when they become less capitalized and want to scale back their operation: fire-sell the capital to households at a possibly large price discount (market illiquidity) or "uninvest" and suffer adjustment costs (technological illiquidity).

Less productive households earn a return of

$$d\underline{r}_t^k = \underbrace{\frac{\underline{a} - \iota_t}{q_t} dt}_{\text{dividend yield}} + \underbrace{(\Phi(\iota_t) - \delta + \mu_t^q + \sigma\sigma_t^q)dt + (\sigma + \sigma_t^q)dZ_t}_{\frac{d(q_t k_t)}{q_t k_t},\ \text{the capital gains rate}} \tag{20}$$

when they manage the physical capital. The households' return differs from that of experts, (4), only in the dividend yield that they earn.

We generalize the model in several other ways. (i) We enable experts to issue some (outside) equity, even though they cannot be 100% equity financed. Specifically, we suppose that experts must retain at least a fraction $\underline{\chi} \in (0,1]$ of equity. (ii) We generalize the model by including a force that prevents experts from "saving their way out" away from the constraints. In particular, we assume that experts could have a higher discount rate ρ than that of households, $\underline{\rho}$. (iii) Equipped with the results derived in Section 3.1 we generalize experts' and households' utility functions from log to CRRA with risk aversion coefficient γ.[f]

To summarize, experts and households maximize, respectively,

$$E\left[\int_0^\infty e^{-\rho t} u(c_t)\right] dt \quad \text{and} \quad E\left[\int_0^\infty e^{-\underline{\rho} t} u(\underline{c}_t)\right] dt.$$

We denote the fraction of capital allocated to experts by $\psi_t \leq 1$ and the fraction of equity retained by experts by $\chi_t \geq \underline{\chi}$.

We want to characterize how any history of shocks $\{Z_s, s \leq t\}$ maps to equilibrium prices q_t and r_t, asset allocations ψ_t and χ_t, and consumption so that (1) all agents maximize utility through optimal consumption and portfolio choices and (2) markets clear. Agents optimize portfolios subject to constraints (no short-selling of capital and a bound on equity issuance by experts). For example, households can invest in capital, the risk-free asset, and experts' equity, and optimize over portfolio weights on these three assets (with a nonnegative weight on capital). Thus, the solution is based on a classic problem in asset pricing. Note also that because the required returns are different between households and experts, the experts' inside equity will generally earn a different return from the equity held by households—experts will earn "management fees" that households do not earn.[g]

[f] Brunnermeier and Sannikov (2014) explicitly consider the case of risk-neutral experts and households. Experts are constrained to consume nonnegative quantities, but households can consume both positive and negative amounts. This assumption leads to the simplification that the risk-free rate in the economy r_t always equals the households' discount rate $\underline{\rho}$.

[g] This is not a universal assumption in the literature. For example, He and Krishnamurthy (2013) assume that returns are equally split between experts and households, so that rationing is required to prevent households from demanding more expert equity than the total supply of expert equity.

3.3 The 4-Step Approach

We can solve for the equilibrium in four steps. First, postulate processes for prices and stochastic discount factor. Second, write down the consumption–portfolio optimization and market-clearing conditions. These conditions imply a stochastic law of motion of the price q_t, the required risk premia for experts and households ς_t and $\underline{\varsigma}_t$, together with variables ψ_t and χ_t. Third, focusing on the experts' balance sheets we write down the law of motion of expert's wealth share

$$\eta_t = \frac{N_t}{q_t K_t},$$

as a percentage of the whole wealth in the economy. As before, K_t is the total amount of capital in the economy. Fourth, we look for a Markov equilibrium, and characterize equations for q_t, ψ_t, etc., as functions of η_t. We solve these equations numerically either as a system of ordinary differential equations (using the shooting method) or as a system of partial differential equations in time, via a procedure analogous to value function iteration in discrete time.

Step 1: Postulating Equilibrium Processes. As before, we postulate the equilibrium prices process for physical capital.

$$\frac{dq_t}{q_t} = \mu_t^q dt + \sigma_t^q dZ_t.$$

Furthermore, as experts and households have different investment opportunities, we postulate two stochastic discount factor (SDF) processes, one for experts and one for households.

$$\frac{d\xi_t}{\xi_t} = -r_t dt - \varsigma_t dZ_t, \quad \text{and} \quad \frac{d\underline{\xi}_t}{\underline{\xi}_t} = -\underline{r}_t dt - \underline{\varsigma}_t dZ_t,$$

respectively.

Step 2: Equilibrium Conditions. Note that since both experts and households can trade the risk-free asset the drift of both SDF processes has to be the same, ie, $\underline{r}_t = r_t$. Moreover, (13) implies the following asset-pricing relationship for capital held by experts:

$$\frac{\frac{a-\iota_t}{q_t} + \Phi(\iota_t) - \delta + \mu_t^q + \sigma\sigma_t^q - r_t}{\sigma + \sigma_t^q} = \chi_t \varsigma_t + (1-\chi_t)\underline{\varsigma}_t, \tag{21}$$

where χ_t is the inside equity share, ie, the fraction of risk held by experts. The required return on capital held by experts depends on the equilibrium capital structure that experts use. If experts require a higher risk premium than households, then $\chi_t = \underline{\chi}$, ie, experts will issue the maximum equity they can. Thus, we have[h]

$$\chi_t = \underline{\chi} \;\text{ if }\; \varsigma_t > \underline{\varsigma}_t, \quad \text{otherwise} \quad \varsigma_t = \underline{\varsigma}_t.$$

Under this condition, we can replace χ_t with $\underline{\chi}$ in (21).

An asset-pricing relationship for capital held by households is

$$\frac{\frac{a - \iota_t}{q_t} + \Phi(\iota_t) - \delta + \mu_t^q + \sigma\sigma_t^q - r_t}{\sigma + \sigma_t^q} \leq \underline{\varsigma}_t, \tag{22}$$

with equality if $\psi_t < 1$, ie, households hold capital in positive amounts. Note that households may choose not to hold any capital, and if so, then the Sharpe ratio they would earn from capital could fall below that required by the asset-pricing relationship.

It is useful to combine (21) and (22), eliminating μ_t^q and r_t, to obtain

$$\frac{(a - \underline{a})/q_t}{\sigma + \sigma_t^q} \geq \underline{\chi}(\varsigma_t - \underline{\varsigma}_t), \tag{23}$$

with equality if $\psi_t < 1$.

The required risk premia can be tied to the agents' consumption processes via (35) in the CRRA case and to the agents' net worth processes in the special logarithmic case. Under the baseline risk-neutrality assumptions of Brunnermeier and Sannikov (2014), $\underline{\varsigma} = 0$ when households are risk-neutral and financially unconstrained—ie, they can consume negatively.

We will use these conditions to characterize q_t, ψ_t, χ_t, etc., as functions of η_t. Before we do that, though, we must derive an equation for the law of motion of $\eta_t = N_t/(q_t K_t)$.

Step 3: The Law of Motion of η_t. It is convenient to express the laws of motion of the numerator and denominator of η_t by focusing on risks and risk premia. Specifically, the experts' net worth follows

$$\frac{dN_t}{N_t} = r_t\,dt + \underbrace{\frac{\chi_t \psi_t}{\eta_t}(\sigma + \sigma_t^q)}_{\text{risk}} \big(\underbrace{\varsigma_t}_{\text{risk premium}} dt + dZ_t \big) - \frac{C_t}{N_t}dt.$$

To derive the evolution of $q_t K_t$, note that the capital gains rate is the same for both type of agents. Thus, we can just aggregate the individual laws of motion to an aggregate law of motion. After replacing the term $\Phi(\iota_t) - \delta + \mu_t^q - \sigma\sigma_t^q - r_t$ using (21), we obtain

[h] We can rule out the case that $\varsigma_t < \underline{\varsigma}_t$ and $\chi_t = 1$: experts cannot face lower risk premia than households if households hold zero risk.

$$\frac{d(q_tK_t)}{q_tK_t} = r_t dt + (\sigma + \sigma_t^q)\Big((\underline{\chi}\varsigma_t + (1-\underline{\chi})\underline{\varsigma}_t)dt + dZ_t\Big) - \frac{a-\iota_t}{q_t}dt.$$

This is the total return on capital (eg, that held by experts) minus the dividend yield.

Using the already familiar formula (10) for a ratio of two stochastic processes, we have

$$\frac{d\eta_t}{\eta_t} = \mu_t^\eta dt + \sigma_t^\eta dt \ = \left(\frac{a-\iota_t}{q_t} - \frac{C_t}{N_t}\right)dt + \frac{\chi_t\psi_t - \eta_t}{\eta_t}(\sigma + \sigma_t^q)\big((\varsigma_t - \sigma - \sigma_t^q)dt + dZ_t\big) + \\ (\sigma + \sigma_t^q)(1-\underline{\chi})(\varsigma_t - \underline{\varsigma}_t)dt. \tag{24}$$

Step 4: Converting the Equilibrium Conditions and Laws of Motion (24) into Equations for $q(\eta)$, $\theta(\eta)$, $\psi(\eta)$, $\chi(\eta)$, etc. The procedure to convert the equilibrium conditions and the law of motion of η_t into numerically solvable equations for $q(\eta)$, $\psi(\eta)$, etc., depends on the underlying assumptions on the agents' preferences. (The log-utility case is the easiest to solve.) In each case, we have to use Ito's Lemma, which allows us to replace terms such as $\sigma_t^q, \sigma_t^\theta$, μ_t^q, etc., with expressions containing the derivatives of q and θ, in order to arrive at solvable differential equations for these functions in the end.

For example, using Ito's Lemma we can tie the volatility of q_t with the first derivative of $q(\eta)$ as follows

$$\sigma_t^q q(\eta) = q'(\eta)\underbrace{(\chi_t\psi_t - \eta_t)(\sigma + \sigma_t^q)}_{\eta\sigma_t^\eta}. \tag{25}$$

Rewriting Eq. (25) yields a closed form solution for the amplification mechanism.

Amplification

$$\sigma_t^\eta = \frac{\frac{\chi_t\psi_t}{\eta_t} - 1}{1 - \left[\frac{\chi_t\psi_t}{\eta_t} - 1\right]\frac{q'(\eta_t)}{q(\eta_t)/\eta_t}}\sigma \tag{26}$$

The numerator $\frac{\chi_t\psi_t}{\eta_t} - 1$ captures the leverage ratio of the expert sector. The amplification increases with the leverage ratio, the leverage effect. The denominator captures the "loss spiral." Mathematically, it reflects an infinite geometric series. The impact of the loss spiral increases with the product of the leverage ratio and price elasticity, $\frac{q'}{q/\eta}$. The latter measures "market illiquidity," the percentage price impact due to a percentage decline in η_t. Market illiquidity arises from the technological specialization of capital, measured here by the difference $a - \underline{a}$ between the experts' and households' productivity parameters. Market illiquidity interacts with technological illiquidity, captured by the curvature of $\Phi(\cdot)$.

There are various methods to solve the equilibrium equations. Below, we discuss two methods that have been used in practice. One method involves ordinary differential equations (ODE)—we refer to it as the "shooting method" and illustrate it using the risk-neutral preferences of Brunnermeier and Sannikov (2014). The second method involves partial differential equations, and is reminiscent of value function iteration in discrete time. A third method has been used in the literature by Drechsler et al. (2016) and Moreira and Savov (2016), namely Chebyshev collocation which is a special type of projection method, see Judd (1998), chapter 11 for details. We do not present this method here, because a global polynomial approximation is less suitable for our model whose solutions may have kinks.

3.4 Method 1: The Shooting Method

This method involves converting the equations above into a system of ODEs. Before we dive into this, in order to understand how this can be done, we review a very simple and well-known model to illustrate the gist of what we have to do. The model illustrates the pricing of a perpetual American put.

Example of Perputual American Put a la Leland (1994)

Consider the problem of pricing a perpetual option to abandon an asset for an amount K. Given a risk-free rate of r and volatility σ, if the asset pays no dividends, its value follows a geometric Brownian motion

$$\frac{dV_t}{V_t} = r\,dt + \sigma\,dZ_t \tag{27}$$

under the risk-neutral measure.

Under the risk-neutral measure, the expected return of any security must be r. Thus, if the put value P_t follows $dP_t = \mu_t^P P_t\,dt + \sigma_t^P P_t\,dZ_t$, then we must have

$$r = \mu_t^P. \tag{28}$$

Suppose we would like to calculate how the put value P_t depends on the value of the assets V_t. Then we face a problem that is completely analogous to the model with financial frictions we described in this section. We have a law of motion of the state variable V_t and a relationship (28) that the stochastic evolution of P_t has to satisfy, and we would like to characterize P_t as a function of V_t.

How can we do this? Easy. Using Ito's Lemma

$$\mu_t^P P_t = rV_t P'(V_t) + \frac{1}{2}\sigma^2 V_t^2 P''(V_t),$$

and so (28) becomes

$$r = \frac{rVP'(V) + \frac{1}{2}\sigma^2 V^2 P''(V)}{P(V)}. \tag{29}$$

If function $P(V)$ satisfies this equation, then the process $P_t = P(V_t)$ will satisfy (28). We are able to go from an equation like (28) to a differential Eq. (29) by assuming that the value of the put is a *function* of the value of the asset.

We can solve the second-order ordinary differential equation (ODE) (29) if we have two boundary conditions. We have $P(V) \to 0$ as $V \to \infty$ since the put becomes worthless if it is never exercised. We also have $P(V) - (K - V) \geq 0$, since $P(V)$ must equal the intrinsic value at the point where the put is exercised.

Our problem is similar: we have an equation for the stochastic law of motion of the state variable (24), as well as the equilibrium conditions that processes $q(\eta_t)$, $\psi(\eta_t)$, etc., must satisfy. Certainly, the equations are more complicated than those of the put-pricing problem, and the law of motion of η_t is endogenous. However, the mechanics of solving these equations is the same—we have to use Ito's Lemma.

Here, we illustrate the derivation of an appropriate set of ordinary differential equations, as well as the "shooting" method for solving them, using the risk-neutral model of Brunnermeier and Sannikov (2014). Assume that experts and households are risk neutral, and while experts must consume nonnegatively, households can have both positive and negative consumption. Then the required risk premium of households is $\underline{\varsigma}_t = 0$. The required risk premium of experts is $-\sigma_t^\theta$, where θ_t is the marginal utility of the experts' wealth that follows

$$\frac{d\theta_t}{\theta_t} = \mu_t^\theta dt + \sigma_t^\theta dZ_t.$$

We would like to construct differential equations to solve for the functions $q(\eta)$, $\theta(\eta)$ and $\psi(\eta)$. The equations will be of second order in $q(\eta)$ and $\theta(\eta)$, ie, we will design a procedure to compute $q''(\eta)$ and $\theta''(\eta)$, as well as $\psi(\eta)$, from η, $q(\eta)$, $q'(\eta)$ and $\theta(\eta)$, $\theta'(\eta)$. Note also that, since households demand no risk premium, ie, $\underline{\varsigma}_t = 0$, experts will issue the maximum allowed fraction of equity to households, so $\chi_t = \underline{\chi}$ at all times.

In this case $q(\eta)$ is an increasing function that satisfies the boundary condition

$$q(0) = \max_{\iota} \frac{\underline{a} - \iota}{r - \Phi(\iota) + \delta},$$

the Gordon growth formula for the value of capital when it is permanently managed by households. Any expert can get infinite utility if he can buy capital at the price of $q(0)$, so

$$\lim_{\eta \to 0} \theta(\eta) = \infty. \tag{30}$$

Function $\theta(\eta)$ is decreasing: the marginal value of the experts' net worth is declining as η rises, and investment opportunities become less valuable. Experts refrain from consumption whenever $\theta(\eta) > 1,$ and consume only at point η^* where $\theta(\eta^*) = 1$, ie, the marginal value of the experts' net worth is exactly 1. That point becomes the reflecting boundary of the system. That is, the system does not go beyond the reflecting boundary and is rather thrown back. In addition, at the reflecting boundary η^* functions $q(\eta)$ and $\theta(\eta)$ must satisfy

$$q'(\eta^*) = \theta'(\eta^*) = 0.$$

Now to the differential equations. Eq. (25) implies that

$$\sigma + \sigma_t^q = \frac{\sigma}{1 - \frac{q'(\eta)}{q(\eta)}(\underline{\chi}\psi_t - \eta_t)}, \tag{31}$$

and by Ito's Lemma,

$$\sigma_t^\theta = \frac{\theta'(\eta)}{\theta(\eta)} \frac{(\underline{\chi}\psi_t - \eta_t)\sigma}{1 - \frac{q'(\eta)}{q(\eta)}(\underline{\chi}\psi_t - \eta_t)}. \tag{32}$$

Therefore, plugging these expressions into the asset-pricing Eq. (23), we obtain

$$\frac{a - \underline{a}}{q(\eta)} \geq -\underline{\chi}\frac{\theta'(\eta)}{\theta(\eta)} \frac{(\underline{\chi}\psi - \eta)\sigma^2}{\left(1 - \frac{q'(\eta)}{q(\eta)}(\underline{\chi}\psi - \eta)\right)^2}. \tag{33}$$

Assuming that $q'(\eta) > 0$ and $\theta'(\eta) < 0$, the right-hand side is increasing from 0 to ∞ as $\underline{\chi}\psi - \eta$ rises from 0 to $q(\eta)/q'(\eta)$. Thus, we have to set $\psi = 1$ whenever it is possible to do so (ie, $\underline{\chi} - \eta < q(\eta)/q'(\eta)$) and this is consistent with inequality (33). Otherwise we determine ψ by solving the quadratic Eq. (33), in which we replace the $\geq$ sign with equality.

After that, we can find σ_t^q from (31), σ_t^θ from (32), μ_t^η and σ_t^η from (24) (where we set $C_t = 0$ since experts consume only at the boundary η^*), μ_t^q from the asset-pricing condition

$$\frac{a - \iota_t}{q_t} + \Phi(\iota_t) - \delta + \mu_t^q + \sigma\sigma_t^q - r = \underline{\chi}(\sigma + \sigma_t^q)(-\sigma_t^\theta),$$

μ_t^θ from the pricing condition for the risk-free asset

$$\mu_t^\theta = \rho - r,$$

and $q''(\eta)$ as well as $\theta''(\eta)$ from Ito's formula,

$$\mu_t^q q(\eta) = \mu_t^\eta \eta q'(\eta) + \frac{1}{2}(\sigma_t^\eta)^2 \eta^2 q''(\eta) \quad \text{and} \quad \mu_t^\theta \theta(\eta) = \mu_t^\eta \eta \theta'(\eta) + \frac{1}{2}(\sigma_t^\eta)^2 \eta^2 \theta''(\eta).$$

3.4.1 Solving the System of ODEs Numerically

We can use an ODE solver in Matlab, such as ode45, to solve the system of equations. We need to perform a search, since our boundary conditions are defined at two endpoints of $[0, \eta^*]$, and we also need to deal with a singularity at $\eta = 0$. The following algorithm performs an appropriate search and deals with the singularity issue, effectively, by solving the system of equations with the boundary condition $\theta(0) = M$, for a large constant M, instead of (30):[i]

Algorithm Set

$$q(0) = \max_{\iota} \frac{\underline{a} - \iota}{r - \Phi(\iota) + \underline{\delta}}, \quad \theta(0) = 1 \quad \text{and} \quad \theta'(0) = -10^{10}.$$

Perform the following procedure to find an appropriate boundary condition $q'(0)$. Set $q_L = 0$ and $q_H = 10^{15}$. Repeat the following loop 50 ×. Guess $q'(0) = (q_L + q_H)/2$. Use Matlab function ode45 to solve for $q(\eta)$ and $\theta(\eta)$ on the interval [0, ?) until one of the following events is triggered, either (1) $q(\eta)$ reaches the upper bound

$$q\mathrm{max} = \max_{\iota} \frac{a - \iota}{r - \Phi(\iota) + \delta},$$

(2) the slope $\theta'(\eta)$ reaches 0 or (3) the slope $q'(\eta)$ reaches 0. If integration has terminated for reason (3), we need to increase the initial guess of $q'(0)$ by setting $q_L = q'(0)$. Otherwise, we decrease the initial guess of $q'(0)$, by setting $q_H = q'(0)$.

At the end, $\theta'(0)$ and $q'(0)$ reach 0 at about the same point, which we denote by η^*. Divide the entire function θ by $\theta(\eta^*)$.[j] Then plot the solutions.

3.4.2 Properties of the Solution

Let us interpret the solution of the risk-neutral model. Point η^* plays the role of the steady state of our system. The drift of η_t is positive everywhere on the interval $[0, \eta^*)$, because the expert sector, which is more productive than the household sector, is growing in expectation. Thus, the system is pushed toward η^* by the drift.

It turns out that the steady state is relatively stable, because volatility is low near η^*. To see this, recall that the amount of endogenous risk in asset prices, from (25), is given by

$$\sigma_t^q = \frac{q'(\eta)}{q(\eta)} \frac{(\underline{\chi}\psi_t - \eta_t)}{1 - \dfrac{q'(\eta)}{q(\eta)}(\underline{\chi}\psi_t - \eta_t)} \sigma.$$

From the boundary conditions, $q'(\eta^*) = 0$, so there is no endogenous risk near η^*.

[i] Footnote j explains why it actually does not to matter what exact value one sets for $\theta(0)$.

[j] We can do this because whenever functions θ and q satisfy our system of equation, so do functions $\Theta\theta$ and q for any constant Θ. Because of that, also, it is immaterial what we set $\theta(0)$ to 1.

However, below η^*, endogenous risk increases as $q'(\eta)$ becomes larger. As prices react to shocks, fundamental risk becomes amplified. As we see from the expression for σ_t^q, this amplification effect is nonlinear, since $q'(\eta)$ enters not only the numerator, but also the denominator. This happens due to the feedback effect: an initial shock causes η_t to drop, which leads to a drop in q_t, which hurts experts who are holding capital and leads to a further decrease in η_t, and so on.

Of course, far in the depressed region the volatility of η_t, $\sigma_t^\eta \eta_t$, becomes low again in this model. This leads to a bimodal stationary distribution of η_t in equilibrium.[k]

Volatility paradox refers to the phenomenon that systemic risk can build up in quiet environments. We can illustrate this phenomenon through comparative statics on σ or the degree of the experts' equity constraint $\underline{\chi}$. One may guess that the system becomes much more stable as σ or $\underline{\chi}$ decline.

This is not the case, as illustrated in Fig. 3 for parameters $\rho = 6\%$, $r = 5\%$, $a = 11\%$, $\underline{a} = 5\%$, $\delta = 3\%$, and an investment function of the form $\Phi(\iota) = \frac{1}{\kappa}(\sqrt{1+2\kappa\iota} - 1)$, $\kappa = 10$,

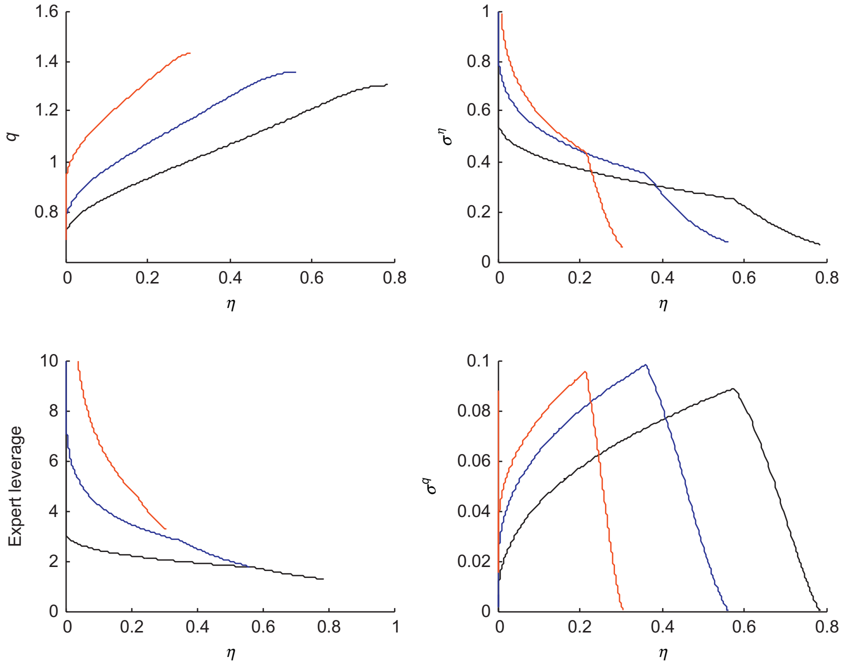


Fig. 3 Equilibrium with $\sigma = 2.5\%$ (red), 10% (blue), and 25% (black). (In the printed version red is grey and blue is dark grey.)

[k] One can prove that the stationary distribution is bimodal analytically by analyzing the asymptotic properties of the solutions near $\eta = 0$ and using the Kolmogorov forward equations that characterize the stationary density—see Brunnermeier and Sannikov (2014) for details.

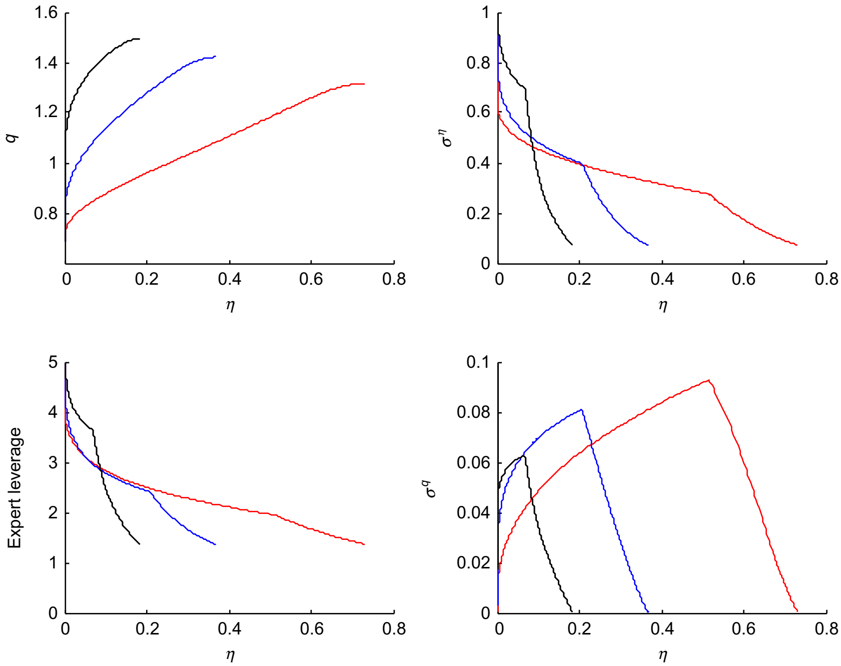


Fig. 4 Equilibrium with $\underline{\chi} = 0.25$ (black), 0.5 (blue) and 1 (red). (In the printed version blue is gray and red light gray.)

$\underline{\chi} = 1$, and various values of σ. (The investment technology in this example has quadratic adjustment costs: an investment of $\Phi + \kappa\Phi^2/2$ generates new capital at rate Φ.)

The volatility paradox shows itself in a number of metrics. As exogenous risk declines,

- maximal endogenous risk σ_t^q may increase (as σ drops from 25% to 10% in Fig. 3)
- the volatility σ_t^η near $\eta = 0$ rises (and this result can be proved analytically)
- from the steady state η^* it takes less time for volatility $\sigma + \sigma_t^q$ to double
- from the steady state, it may take less time to reach the peak of the crisis η^ψ, where experts start selling capital to households.[1]

Fig. 4 takes the same parameters and $\sigma = 20\%$, but varies $\underline{\chi}$. As $\underline{\chi}$ falls, expert net worth at the steady state η^* drops significantly, and the volatility σ_t^η in the crisis regime rises.

3.5 Method 2: The Iterative Method

Here, we describe the iterative method of finding the equilibrium, by solving a system of partial differential equations back in time away from a terminal condition. Specifically, imagine an economy that lasts for a finite time horizon $[0, T]$. Given a set of terminal conditions at time T, we would like to compute the equilibrium over the time horizon

[1] As σ declines, the system spends less time in the crisis region, so some measures of stability improve, but the amount of time spent in crisis does not converge to 0 as $\sigma \rightarrow 0$.

$[0, T]$. The iterative method is based on the premise that as we let $T \to \infty$, behavior at time 0 should converge to the equilibrium of the infinite-horizon economy. Computation uses the equilibrium conditions that express the drifts of various processes, and uses those drifts to obtain time derivatives for the corresponding functions of the state space. The iterative method is analogous to value function iteration in discrete time.

We illustrate the method here based on a model with CRRA utility

$$u(c) = \frac{c^{1-\gamma}}{1-\gamma}.$$

Equilibrium conditions (21) and (19) provide two equations that express the drift of the price q_t, as well as the drifts of aggregate consumption of experts C_t and households $\underline{C}_t$. We also have another asset-pricing condition (23), which does not contain any drift terms. In the end we have three functions but only two drift conditions. As a result, the time dimension of our computation involves only two functions—the value functions of experts and households—and the third function, the price, is found for each time point through a separate procedure.[m]

Our procedure is literally the analogue of value function iteration (but with multiple agents affecting the evolving stochastic state). It is convenient to derive directly the equations that value functions must satisfy. The value functions of experts and households can be presented in the form

$$v_t \frac{K_t^{1-\gamma}}{1-\gamma} = \frac{v_t}{(\eta_t q_t)^{1-\gamma}} \frac{N_t^{1-\gamma}}{1-\gamma} \quad \text{and} \quad \underline{v}_t \frac{K_t^{1-\gamma}}{1-\gamma}.$$

Since the marginal utilities of consumption and wealth must be the same, we have

$$C_t^{-\gamma} = \frac{v_t}{(\eta_t q_t)^{1-\gamma}} N_t^{-\gamma} = \frac{v_t}{\eta_t q_t} K_t^{-\gamma} \quad \Rightarrow \quad C_t = N_t \frac{(\eta_t q_t)^{1/\gamma - 1}}{v_t^{1/\gamma}} = K_t \frac{(\eta_t q_t)^{1/\gamma}}{v_t^{1/\gamma}}. \tag{34}$$

Hence, the risk premia of households and experts are given by

$$\varsigma_t = \gamma \sigma_t^C = -\sigma_t^v + \sigma_t^\eta + \sigma_t^q + \gamma\sigma \quad \text{and} \quad \underline{\varsigma}_t = \gamma \sigma_t^{\underline{C}} = -\sigma_t^{\underline{v}} - \frac{\eta \sigma_t^\eta}{1-\eta} + \sigma_t^q + \gamma\sigma. \tag{35}$$

Since

$$\underbrace{\int_0^t e^{-\rho s} \frac{C_s^{1-\gamma}}{1-\gamma} ds}_{\text{utility flow}} + e^{-\rho t} \underbrace{v_t \frac{K_t^{1-\gamma}}{1-\gamma}}_{\text{continuation utility}}$$

[m] If we used the shooting method to find the equilibrium with CRRA utilities, we would have a system of second-order differential equations for the value functions, and a first-order differential equation for the price.

is by standard dynamic programming arguments a martingale and

$$\frac{d(K_t^{1-\gamma})}{K_t^{1-\gamma}} = \left((1-\gamma)(\Phi(\iota_t)-\delta) - \frac{\gamma(1-\gamma)}{2}\sigma^2\right)dt + (1-\gamma)\sigma dZ_t,$$

we have

$$\frac{C_t^{1-\gamma}}{1-\gamma} - \rho v_t \frac{K_t^{1-\gamma}}{1-\gamma} + v_t \frac{K_t^{1-\gamma}}{1-\gamma}\left(\mu_t^v + (1-\gamma)(\Phi(\iota_t)-\delta) - \frac{\gamma(1-\gamma)}{2}\sigma^2 + \sigma_t^v(1-\gamma)\sigma\right) = 0.$$

Using (34), we obtain

$$\mu_t^v = \rho - \frac{(\eta_t q_t)^{1/\gamma-1}}{v_t^{1/\gamma}} - (1-\gamma)(\Phi(\iota_t)-\delta) + \frac{\gamma(1-\gamma)}{2}\sigma^2 - \sigma_t^v(1-\gamma)\sigma. \tag{36}$$

Likewise,

$$\mu_t^{\underline{v}} = \underline{\rho} - \frac{((1-\eta_t)q_t)^{1/\gamma-1}}{\underline{v}_t^{1/\gamma}} - (1-\gamma)(\Phi(\iota_t)-\delta) + \frac{\gamma(1-\gamma)}{2}\sigma^2 - \sigma_t^{\underline{v}}(1-\gamma)\sigma. \tag{37}$$

Given μ_t^v and $\mu_t^{\underline{v}}$, we obtain partial differential equations for the functions $v(\eta, t)$ and $\underline{v}(\eta, t)$ using Ito's Lemma, and they are as follows:

$$\mu_t^v v(\eta,t) = \mu_t^\eta \eta v_\eta(\eta,t) + \frac{(\sigma_t^\eta \eta)^2}{2} v_{\eta\eta}(\eta,t) + v_t(\eta,t) \quad \text{and} \tag{38}$$

$$\mu_t^{\underline{v}} \underline{v}(\eta,t) = \mu_t^\eta \eta \underline{v}_\eta(\eta,t) + \frac{(\sigma_t^\eta \eta)^2}{2} \underline{v}_{\eta\eta}(\eta,t) + \underline{v}_t(\eta,t). \tag{39}$$

3.5.1 Description of the Procedure

Below we outline the procedure of how we solve for the equilibrium using Eqs. (38) and (39).[n] There are three parts.

- The terminal conditions $v(\eta, T)$ and $\underline{v}(\eta, T)$
- The static step: finding capital price $q(\eta)$, allocations $\psi(\eta)$ and $\chi(\eta)$, volatilities and drifts at a given time point t given the value functions $v(\eta, t)$ and $\underline{v}(\eta, t)$, and
- The time step: finding $v(\eta, t - \Delta t)$ and $\underline{v}(\eta, t - \Delta t)$ from prices, allocations, volatilities and drifts at time t.

[n] For more details on the finite difference method for dynamic programming problems we refer to Candler (1999). Oberman (2006) provides sufficient conditions for a numerical scheme to converge to the solution of a general class of non-linear parabolic Partial Differential Equations.

3.5.1.1 The Terminal Conditions

Our terminal conditions specify the utilities of the representative expert and household, as functions of the experts' wealth share η_t. We have not performed a detailed theoretical study of acceptable terminal conditions, but in practice any reasonable guess works well for a wide range of parameters.

For example, if we set $q_T = 1$ and $C_T/K_T = a\eta_T$, then (34) implies that

$$v_T = \eta_T (a\eta_T)^{-\gamma} \quad \text{and} \quad \underline{v}_T = (1-\eta_T)(a(1-\eta_T))^{-\gamma}. \tag{40}$$

3.5.1.2 The Static Step

Suppose we know value functions through $v(\eta, t)$ and $\underline{v}(\eta, t)$. Let us describe how we can compute the price q_t and characterize equilibrium dynamics at time t.

There are three regions. When η is close enough to 0, then the experts' risk premia are so much higher than those of households that $\psi_t < 1$, ie, households hold capital, and Eq. (23) holds. In this region experts issue the maximal allowed equity share to households, so $\chi_t = \underline{\chi}$, since the households' risk premia are lower. In the middle region, $\psi_t = 1$, ie, only experts hold capital, but the experts' risk premia are still higher than those of households so $\chi_t = \underline{\chi}$. Finally, when $\eta \geq \underline{\chi}$, the capital is allocated efficiently to experts (ie, $\psi_t = 1$) and risk can be shared perfectly between households and experts by setting $\chi_t = \eta_t$. In the last region, (26) implies that $\sigma^\eta = 0$, so there is no endogenous risk, and risk premia of experts and households are both equal to $\varsigma_t = \underline{\varsigma}_t = \gamma\sigma$ by (35).

In the region where $\psi_t < 1$ we solve for $q(\eta)$, $\psi(\eta)$ and $\sigma + \sigma_t^q$ from a system of the following three equations, which ultimately gives us a first-order ODE in $q(\eta)$. We obtain the first by combining (23) and (35) together with evolution of η Eq. (24), we have

$$\frac{a - \underline{a}}{q_t} = \underline{\chi} \underbrace{\left(\frac{\underline{v}'(\eta)}{\underline{v}(\eta)} - \frac{v'(\eta)}{v(\eta)} + \frac{1}{\eta(1-\eta)}\right)(\underline{\chi}\psi_t - \eta)(\sigma + \sigma_t^q)^2}_{\left(\sigma_t^{\underline{v}} - \sigma_t^v + \frac{\sigma_t^\eta}{1-\eta}\right)(\sigma + \sigma_t^q)}. \tag{41}$$

The second we obtain from (25) and Ito's Lemma,

$$(\sigma + \sigma^q)\left(1 - (\underline{\chi}\psi - \eta)\frac{q'(\eta)}{q(\eta)}\right) = \sigma. \tag{42}$$

Finally, from (34) and an analogous condition for households, the market-clearing condition for output is

$$\underbrace{\frac{(\eta_t q_t)^{1/\gamma}}{v_t^{1/\gamma}} + \frac{((1-\eta_t)q_t)^{1/\gamma}}{\underline{v}_t^{1/\gamma}}}_{(C_t + \underline{C}_t)/K_t} = a\psi + \underline{a}(1-\psi) - \iota(q(\eta)). \tag{43}$$

Once ψ_t reaches 1, condition (41) is no longer relevant. From then on, we set $\psi_t = 1$, find $q(\eta)$ from (43) and $\sigma + \sigma_t^q$ from (42). Once η_t reaches $\underline{\chi}$, we enter the last region. There we set $\psi_t = 1, \chi_t = \eta_t,$ compute $q(\eta)$ from (43) and set $\sigma_t^q = 0$.

Once we know function $q(\eta)$ in all three regions, we can find the volatility of η_t from (24) and the volatilities of v_t and $\underline{v}_t$ from Ito's Lemma, ie,

$$\sigma_t^\eta = \frac{\chi_t \psi_t - \eta_t}{\eta_t}(\sigma + \sigma_t^q), \quad \sigma_t^v = \frac{v'(\eta)}{v(\eta)}\sigma_t^\eta \eta, \quad \text{and} \quad \sigma_t^{\underline{v}} = \frac{\underline{v}'(\eta)}{\underline{v}(\eta)}\sigma_t^\eta \eta. \tag{44}$$

We find the required risk premia ς_t and $\underline{\varsigma}_t$ from (35) and the drift of η_t from (24), ie,

$$\mu_t^\eta = \left(\frac{a - \iota_t}{q_t} - \frac{(\eta_t q_t)^{1/\gamma - 1}}{v_t^{1/\gamma}}\right) + \sigma_t^\eta(\varsigma_t - \sigma - \sigma_t^q) + (\sigma + \sigma_t^q)(1 - \underline{\chi})\left(\varsigma_t - \underline{\varsigma}_t\right).$$

Finally, we solve for the drifts of v_t and $\underline{v}_t$ from (36).

3.5.1.3 The Time Step

Once we have all characteristics of the equilibrium at a given time point t, we can solve for the value functions at an earlier time step $t - \Delta t$ from Eqs. (38) and (39). These are parabolic equations, which can be solved using either explicit or implicit methods.

3.5.1.4 Summary

Set terminal conditions for value functions $v(\eta, T)$ and $\underline{v}(\eta, T)$ according to (40) on a grid over η. Divide the interval $[0, T]$ into small subintervals. Going backwards in time, for each subinterval $[t - \Delta t, t]$ perform the static step and then the time step. That is, from value functions $v(\eta, t)$ and $\underline{v}(\eta, t)$ find the drift and volatility of η as well as the drifts of v and $\underline{v}$ using the following procedure (static step). Start from an initial condition near $(\eta = 0, \psi = 0)$ (perturb the condition to avoid division by 0). Solve (41), (42), and (43) (as a first-order ordinary differential equation for $q(\eta)$) until ψ reaches 1. Then set $\psi = 1$ and use (43) to find $q(\eta)$ and (42) to find σ^q. Throughout, use $\chi_t = \max(\underline{\chi}, \eta)$. With functions (of η) q, σ^q, ψ and χ obtained in this way, compute volatilities from (44), ς_t and $\underline{\varsigma}_t$ from (35), μ_t^η from (24), and the drifts of v_t and $\underline{v}_t$ from (36). Then (this is the time step) solve the partial differential Eqs. (38) and (39) for v and $\underline{v}$ backward in time over the interval $[t - \Delta t, t]$, using fixed functions $\mu_t^v, \mu_t^{\underline{v}}, \mu_t^\eta$ and σ_t^η of η computed by the static step. Continue until time 0. We get convergence when T is sufficiently large.

Remark The static step alone is sufficient to solve for the equilibrium prices, allocations and dynamics in a model with logarithmic utility (ie, $\gamma = 1$), since in this case we know that $(C_t + \underline{C}_t)/(q_t K_t) = \rho\eta + \underline{\rho}(1 - \eta)$ and expert and household risk premia are $\varsigma_t = \sigma_t^N = \chi_t \psi_t / \eta_t (\sigma + \sigma_t^q)$ and $\underline{\varsigma}_t = (1 - \chi_t \psi_t)/(1 - \eta_t)(\sigma + \sigma_t^q)$. Hence, Eqs. (41) and (43) become

$$\frac{a-\underline{a}}{q_t}=\chi\frac{\underline{\chi}\Psi_t-\eta}{\underline{\ }\eta(1-\eta)}(\sigma+\sigma_t^q)^2 \quad \text{and} \quad (\rho\eta+\underline{\rho}(1-\eta))q_t=a\psi+\underline{a}(1-\psi)-\iota(q(\eta)). \qquad (45)$$

Eq. (42) remains the same.

For logarithmic utility, however, we do not immediately obtain the agents' value functions. Those can be found using an extra step.

3.6 Examples of Solutions: CRRA Utility

In this section, we illustrate solutions generated by our code, using the iterative method, and what we learn from them. We use baseline parameters $\rho = 6\%$, $r = 5\%$, $a = 11\%$, $\underline{a}=3\%$, $\delta = 5\%$, $\sigma = 10\%$, $\underline{\chi}=0.5$, $\gamma = 2$ and an investment function of the form $\Phi(\iota)=\log(\kappa\iota+1)/\kappa$ with $\kappa = 10$. We then study how several parameters, specifically $\underline{a}$, σ, $\underline{\chi}$ and γ affect the equilibrium.

Fig. 5 illustrates the equilibrium for the baseline set of parameters. Notice that capital price q_t has a kink—the kink separates the crisis region near $\eta=0$ where $\psi_t<1$, ie, households hold some capital, and the normal region where experts hold all capital in the economy.

Here, point η^* where the drift of η_t becomes 0 plays the role of a steady state of the system. In the absence of shocks, the system stays still at the steady state and in response to small shocks, drift pushes the system back to the steady state. Moving away from the crisis regime, at η^* risk premia decline sufficiently so that the experts' earnings are exactly offset by their slightly higher consumption rates.

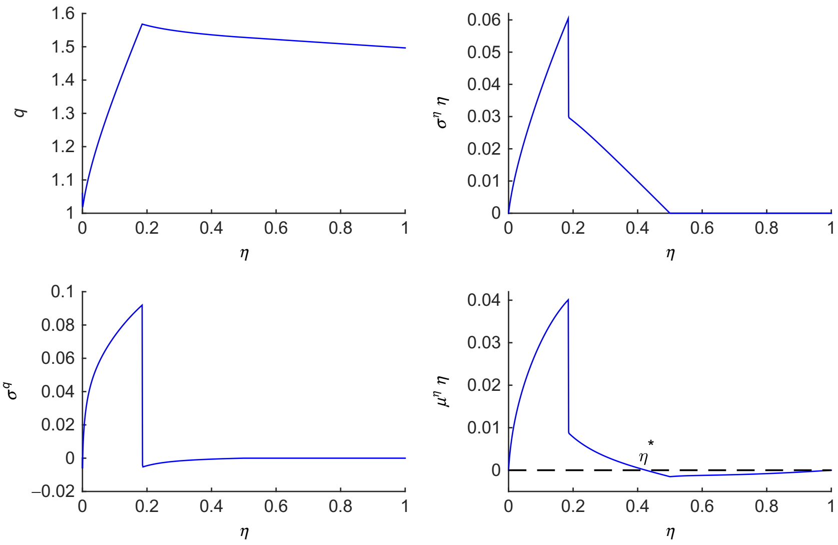


Fig. 5 Equilibrium for the baseline set of parameters.

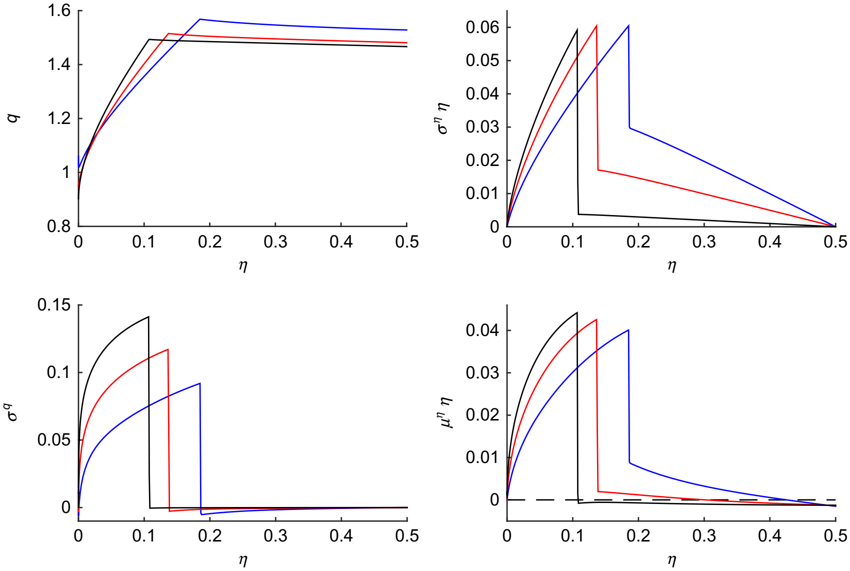


Fig. 6 Equilibrium for $\sigma = 0.01$ (black), 0.05 (red), 0.1 (blue). (In the printed version red is gray and blue is dark gray.)

Above $\eta = \underline{\chi} = 0.5$ is the region of perfect risk sharing, where the volatility of η is zero. Since the drift in that region is negative, the system never ends up there (and if the initial condition is $\eta_0 > \underline{\chi}$, then η_t drifts deterministically down to $\underline{\chi}$).

Fig. 6 shows the effect of σ on the equilibrium dynamics. We bound the horizontal axis at $\eta = \underline{\chi} = 0.5$, since the system never enters the region $\eta > \underline{\chi}$. The steady state η^* declines as σ falls, as risk premia decline in the normal regime, until η^* coincides with the boundary of the crisis region for low σ (this happens for $\sigma = 0.01$ in Fig. 6). We also observe the volatility paradox: as σ declines, endogenous risk σ_t^q does not have to fall, and may even rise.

But what happens as $\sigma \to 0$? Does endogenous risk disappear altogether, and does the solution converge to first best? The answer turns out to be no: in the limit as $\sigma \to 0$, the boundary of the crisis region η^ψ converges not to 0 but to a finite number.

Likewise, what happens if financial frictions become relaxed, and experts are able to hold capital while retaining a smaller portion of risk? It is tempting to conjecture that as financial frictions become relaxed, the system becomes more stable. Yet, as the bottom left panel of Fig. 7 demonstrates, endogenous risk σ_t^q rises sharply as $\underline{\chi}$ declines.[o]

It turns out that a crucial parameter that affects system stability is the household productivity parameter $\underline{a}$. The level of endogenous risk in crises depends strongly on the

[o] Of course, there is a discontinuity at both $\sigma = 0$ and $\underline{\chi} = 0$. As financial frictions disappear altogether, the crisis region disappears.

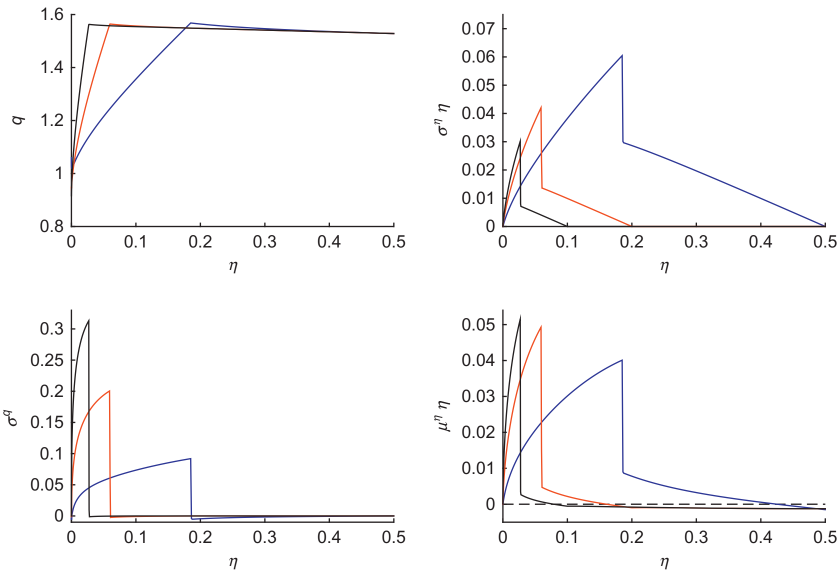


Fig. 7 Equilibrium for $\underline{\chi} = 0.1$ (black), 0.2 (red), and 0.5 (blue). (In the printed version red is gray and blue dark gray.)

market illiquidity of capital—the difference between parameter a and $\underline{a}$ that determines how much less households value capital, in the event that they have to buy it, relative to experts. Fig. 8 illustrates the equilibrium for several values of $\underline{a}$. Note that endogenous risk in crises rises sharply as $\underline{a}$ drops. However, the dynamics in the normal regime and the level of η^* have extremely low sensitivity to $\underline{a}$—only dynamics in the crisis regime are extremely sensitive. This is a surprise. While expert leverage responds endogenously to fundamental risk σ in the normal regime it does not respond strongly to endogenous tail risk. In fact, for logarithmic utility it is possible to prove analytically that the dynamics in the normal regime do not depend on $\underline{a}$ at all (but here we illustrate the dynamics for $\gamma = 2$).

Finally, let us consider risk aversion γ in Fig. 9. There are several effects. Lower risk aversion leads to a smaller crisis region (but with greater endogenous risk), and lower steady state η^* as the risk premia become lower. In this example, higher risk aversion leads to a higher price of capital, as risk creates a precautionary savings demand.

4. A SIMPLE MONETARY MODEL

So far we focused on a real model with a single risky asset, physical capital, and a risk-free asset. Now, building on Brunnermeier and Sannikov's (2015a) "I Theory of Money" we introduce instead of the (real) risk-free asset, another asset, money. In general, money has three roles: it is a unit of account, it facilitates transactions, and it serves as a store of value (safe asset). Here, we focus on its role as a store of value, which arises in our setting due to

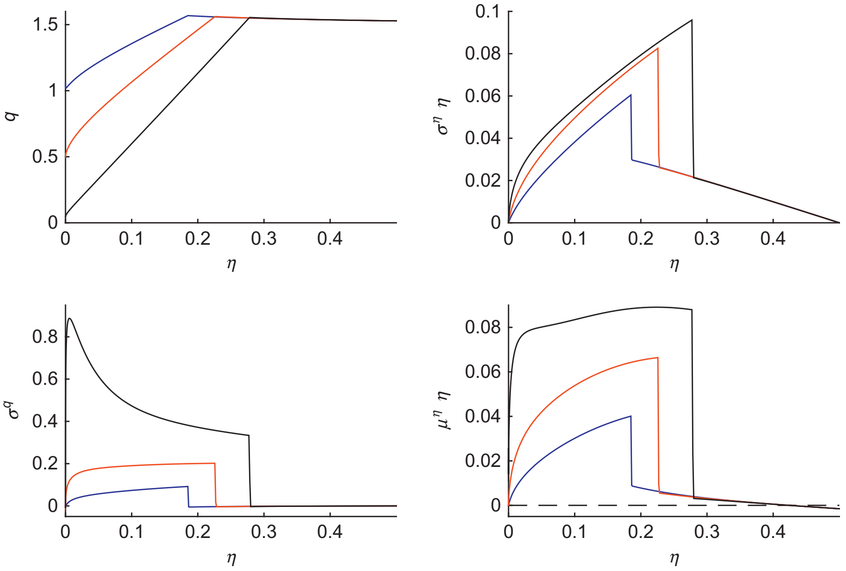


Fig. 8 Equilibrium for $\underline{a} = 0.03$ (blue), -0.03 (red), and -0.09 (black). (In the printed version blue is dark grey and red is gray.)

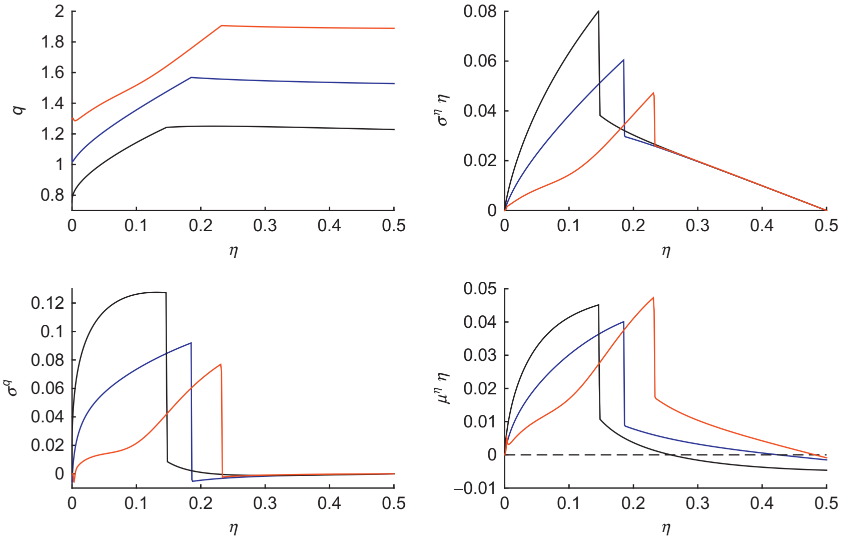


Fig. 9 Equilibrium for $\gamma = 0.5$ (black), 2 (blue), and 5 (red). (In the printed version blue is dark gray and red is gray.)

incomplete markets frictions. Unlike in New Keynesian models, which focus on the role of money as a unit of account and rely on price and wage rigidities as the key frictions, prices are fully flexible in our model.

This section focuses on the following:

1. Money can have positive value despite the fact that it never pays any dividend. That is, money is a bubble.
2. Money helps agents to share risks in an economy that is plagued by financial frictions. Hence, having a nominal store of value instead of a real short-term risk-free bond alters the equilibrium risk dynamics.
3. The "*paradox of prudence*" coined in Brunnermeier and Sannikov (2015a) arises. Experts hold money to self-insure against idiosyncratic shocks, an action which is micro-prudent but macro-*im*prudent. By selling capital to achieve a greater portfolio weight on money, experts depress aggregate investment and growth, leading to lower returns on all assets (including money). The paradox of prudence is in the risk space what Keynes' Paradox of Thrift is for the consumption-savings decision. The Paradox of Thrift describes how each person's attempt to save more paradoxically lowers overall aggregate savings.

4.1 Model with Idiosyncratic Capital Risk and Money

Let us return to the Basak–Cuoco model of Section 2 with experts holding physical capital and households who cannot, ie, $\underline{a} = -\infty$. We introduce the following two modifications: (i) Capital has in addition to aggregate risk also idiosyncratic risk. (ii) There is no risk-free asset, but there is money in fixed supply. Agents can long and short it and want to hold it to self-insure against idiosyncratic risk.

More formally, we assume as before that each expert operates a linear production technology, ak_t, with productivity a, but now they also face idiosyncratic risk $\tilde{\sigma}\, d\tilde{Z}_t$ in addition to aggregate risk σdZ_t. That is a single expert's capital k_t evolves according to

$$dk_t/k_t = (\Phi(\iota_t) - \delta)dt + \sigma dZ_t + \tilde{\sigma}\, d\tilde{Z}_t.$$

The shock dZ_t is the same for the whole economy, while the shock $d\tilde{Z}_t$ is expert-specific and orthogonal to dZ_t. Idiosyncratic shocks cancel out in the aggregate.

Since idiosyncratic risk is uninsurable due to markets incompleteness, experts also want to hold money. Money is an infinitely divisible asset in fixed supply, which can be traded without frictions. Since money does not pay off any dividends it has value in equilibrium only because agents want to self-insure against idiosyncratic shocks to their capital holdings. In other words, money is a bubble, like in Samuelson (1958) and Bewley (1980). Unlike in Bewley (1980), our idiosyncratic shocks are not endowment shocks, but investment shocks like in Angeletos (2007). We assume that idiosyncratic risk of the dividend-paying capital is large enough, $\tilde{\sigma} > \sqrt{\rho}$, so that money, which does not pay dividends, still has value in equilibrium. This is unlike Diamond (1965) who introduces physical capital in Samuelson's OLG model and Aiyagari (1994) who introduces capital in Bewley's incomplete markets setting. In those models, the presence of capital crowds out money as a store of value.[p]

Experts can invest in (outside) money and capital, while households like in Section 2 only hold money. We also assume for simplicity that all agents have logarithmic utility with time preference rate ρ.[q]

As before, let us follow our four step approach to solve the model.

4.2 The 4-Step Approach

Step 1: Postulate Price and SDF Processes. In this monetary setting we now have to postulate not only a process for the price of capital, but also for the "real price" of money. We denote (without loss of generality) the value of the total money stock in terms of the numeraire (the consumption good) by $p_t K_t$. We normalize the total value of the money stock by K_t to emphasize that, everything else being equal, the value of money should be proportional to the size of the economy.

$$\frac{dq_t}{q_t} = \mu_t^q dt + \sigma_t^q dZ_t,$$
$$\frac{dp_t}{p_t} = \mu_t^p dt + \sigma_t^p dZ_t,$$

In addition, like in Section 3 we postulate the processes for individual experts' and households' stochastic discount factors:

$$\frac{d\xi_t}{\xi_t} = -r_t dt - \varsigma_t dZ_t - \tilde{\varsigma}_t d\tilde{Z}_t \quad \text{and} \quad \frac{d\underline{\xi}_t}{\underline{\xi}_t} = -\underline{r}_t dt - \underline{\varsigma}_t dZ_t,$$

where r_t and $\underline{r}_t$ are the (real) shadow risk-free interest rates of experts and households, respectively. Note that shadow risk-free rates need not be identical, since no real risk-free asset is traded. Note also that experts require a risk premium not only for the aggregate risk ς_t but also for the idiosyncratic risk they have to bear $\tilde{\varsigma}_t$.

We will show that there exists an equilibrium in which the wealth share η_t evolves deterministically and so do the prices q_t and p_t. Hence, for simplicity we set $\sigma_t^q = \sigma_t^p = 0$. Under this conjecture the return on physical capital accruing to experts is

[p] We assume that money is intrinsically worthless, and so along with the equilibrium in which money has value, there is also an equilibrium in which money has no value. However, in a perturbation of the model, in which agents get small utility from holding money (eg, because money facilitates transactions), only the equilibrium with full value of money survives.

[q] Solving this model with CRRA models using the results on page 1511 in Section 3.1 is a worthwhile exercise.

$$dr_t^k = \frac{a - \iota_t}{q_t} dt + (\Phi(\iota_t) - \delta + \mu_t^q) dt + \sigma dZ_t + \tilde{\sigma}\, d\tilde{Z}_t$$

and world stock of money $p_t K_t$ earns the (real) return of

$$dr_t^M = \left(\Phi(\iota_t) - \delta + \mu_t^p\right) dt + \sigma dZ_t,$$

where ι_t is the investment rate in physical capital.

Step 2: Equilibrium Conditions. First, note that the optimal investment rate is determined by q_t through $\Phi'(\iota_t) = 1/q_t$. Second, the optimal consumption rate of all agents is simply ρ times their net worth, since the utility of all agents is logarithmic with time preference rate ρ. Hence, aggregate demand for the consumption good is $\rho(q_t + p_t)K_t$. Given total supply of consumption goods after investing, we have the following goods market equilibrium condition:

$$\rho(q_t + p_t)K_t = (a - \iota)K_t.$$

Next, we solve the experts' and households' portfolio problems. Notice that, given the returns dr_t^M and dr_t^k on capital and money, the only two assets traded in this economy, all agents have exposure σdZ_t to aggregate risk. At the same time, experts also have exposure $x_t \tilde{\sigma}\, d\tilde{Z}_t$ to their individual idiosyncratic shocks, where x_t is the experts' portfolio weight on capital. Hence, the required risk premia of these log-utility agents are

$$\varsigma_t = \underline{\varsigma}_t = \sigma \quad \text{and} \quad \tilde{\varsigma}_t = x_t \tilde{\sigma}\,.$$

The experts' and households' asset pricing equations for money, respectively, are

$$\frac{E_t[dr_t^M]}{dt} - r_t = \frac{E_t[dr_t^M]}{dt} - \underline{r}_t = \underbrace{\sigma^2}_{=\varsigma_t\sigma=\underline{\varsigma}_t\sigma}\,.$$

Thus, $r_t = \underline{r}_t$: even though there is no risk-free real asset in this economy, both agent types would agree on a single real risk-free real interest rate.

The experts' asset pricing equation for physical capital is

$$\frac{E_t[dr_t^k]}{dt} - r_t = \varsigma_t\sigma + \tilde{\varsigma}_t\,\tilde{\sigma}\,,$$

reflecting the fact that experts are also exposed to idiosyncratic risk for which they earn an extra risk premium. Hence,

$$\frac{E_t[dr_t^k]}{dt} - \frac{E_t[dr_t^M]}{dt} = x_t \tilde{\sigma}^2. \tag{46}$$

Capital market clearing implies that

$$x_t = \frac{q_t K_t}{\eta_t (p_t + q_t) K_t} = \frac{1}{\eta_t} \frac{q_t}{p_t + q_t}, \tag{47}$$

Step 3: Evolution of $\boldsymbol{\eta}$. Experts' aggregate net worth N_t evolves according to

$$\frac{dN_t}{N_t} = r_t + \sigma(\underbrace{\sigma}_{\varsigma_t} dt + dZ_t) + x_t \tilde{\sigma} \tilde{\varsigma}_t dt - \rho dt,$$

given their exposures to aggregate and idiosyncratic risk, and since idiosyncratic risk cancels out in the aggregate. The law of motion of aggregate wealth is

$$\frac{d((q_t + p_t)K_t)}{(q_t + p_t)K_t} = r_t + \sigma(\sigma dt + dZ_t) + \eta_t x_t \tilde{\sigma} \tilde{\varsigma}_t dt - \rho dt,$$

where $\eta_t = \dfrac{N_t}{(q_t + p_t)K_t}$ is the experts' net worth share and $\eta_t x_t = q_t/(p_t + q_t)$ is the exposure to idiosyncratic risk in the world portfolio. Hence,

$$\frac{d\eta_t}{\eta_t} = x_t^2 (1 - \eta_t) \tilde{\sigma}^2 dt = \left(\frac{q_t}{p_t + q_t} \right)^2 \frac{1 - \eta_t}{\eta_t^2} \tilde{\sigma}^2 dt. \tag{48}$$

Step 4: Derive ODEs for the postulated price processes q and p as a function of the state variable η. We omit this step as it is similar to the previous section.

4.3 Observations and Limit Case

The increase in experts' wealth share η_t, or equivalently the decline of households' wealth share, $1 - \eta_t$, results in part from the fact that experts earn a risk premium from taking on idiosyncratic risk. The higher the idiosyncratic risk $\tilde{\sigma}^2$, the faster experts' wealth share rises toward 100%. Interestingly, it is the fact that experts are unable to share idiosyncratic risk which makes them richer over time compared to households.

Money allows for some sharing of idiosyncratic risk, since the experts' exposure to idiosyncratic risk of $x_t \tilde{\sigma}$ is less than what it would have been without money, ie, $\tilde{\sigma} / \eta_t$, as long as $x_t < 1/\eta_t$ or $p_t > 0$.

4.3.1 Comparison with Real Model

It is instructive to contrast the settings of this section with that of Section 2, where households hold the real risk-free asset instead of money. The evolution η follows now (48) instead of (11). Note that in both settings the experts' wealth share drifts towards 100%. However, there are crucial differences. In the setting with nominal money, aggregate risk is shared fully between experts and households. Hence, both groups receive a risk premium and therefore aggregate risk does not impact the wealth share in the model with money. In contrast, in the real model experts hold all the aggregate risk and hence only they earn a risk premium, leading to a positive drift in η. More importantly, aggregate risk sharing with money makes the evolution of experts' wealth share deterministic. In contrast, in the real model that experts' wealth share is necessarily stochastic, as revealed by (11).

4.3.2 The Only Experts Case

Finally, we are able to derive a closed form solution for the absorbing state $\eta = 1$ to which the system drifts. When the state $\eta = 1$ is reached $\mu^q(1) = \mu^p(1) = 0$ and thus experts' asset pricing Eq. (46) and capital market clearing (47) can be combined and simplified as follows

$$\frac{1}{\tilde{\sigma}^2}\frac{a-\iota}{q} = \frac{E[dr^k - dr^M]/dt}{\tilde{\sigma}^2} = x_t = \frac{q}{p+q} \tag{49}$$

Combining Eq. (49) with the goods market clearing condition

$$\rho(p+q)K_t = (a-\iota)K_t \tag{50}$$

and the optimal investment rate

$$\iota = \frac{q-1}{\kappa}, \tag{51}$$

for the functional form $\Phi(\iota) = \frac{1}{\kappa}\log(\kappa\iota + 1)$ one obtains the "money equilibrium," in which money is a bubble with

$$q = \frac{1+\kappa a}{1+\kappa\sqrt{\rho}\,\tilde{\sigma}} \quad \text{and} \quad p = \frac{\tilde{\sigma}-\sqrt{\rho}}{\sqrt{\rho}}q.$$

The "money equilibrium" exists as long as $\tilde{\sigma} > \sqrt{\rho}$.

In addition, there exists a "moneyless equilibrium," obtained by setting $p = 0$ and solving (50) with (51) to obtain

$$q^0 = \frac{1+\kappa a}{1+\kappa\rho} \quad \text{and} \quad p^0 = 0.$$

Eq. (49) is no longer relevant because money is no longer an asset in which agents can put their wealth.

Note that the price of capital for the "moneyless" equilibrium is the same as in the real economy of Section 2. The growth rate of the economy in both equilibria is given by $g = \frac{1}{\kappa}\log q - \delta$. In the money equilibrium, q is lower and so is overall economic growth, but experts have to bear less risk.

4.3.3 Financial Deepening

Financial deepening or innovation that lower the amount of idiosyncratic risk households have to bear also lowers the value of money, p. However, it increases the price of capital q and with it, the investment rate, ι, and the overall economic growth rate g. Surprisingly, $q + p$ declines. That is, financial deepening lowers total wealth in the economy.

4.3.4 The Paradox of Prudence

The paradox of prudence arises when experts try to lower their risk by tilting their portfolio away from real investment and towards safe asset, money. Scaling back risky asset holding can be micro-prudent, but macro-*im*prudent. As experts try to lower their (idiosyncratic) risk exposure, the price of capital falls in Brunnermeier and Sannikov (2015a). This behavior lowers overall economic growth and with it the real return on money holdings. Since each individual expert takes prices and rates of return as given, they do not internalize this pecuniary externality. As shown in Brunnermeier and Sannikov (2015a), money holdings in this model are inefficiently high if $\tilde{\sigma}(1-\kappa\rho) > 2\sqrt{\rho}$. Our paradox of prudence is analogous to Keynes' Paradox of Thrift, but the former is about changes in portfolio choice and risk, while the latter refers to the consumption–savings decision.[r]

5. CRITICAL ASSESSMENT AND OUTLOOK

The economy with two types of agents gives rise to a number of general ideas—we describe these broader ideas in this section. We would like to make the point that continuous time has the capacity to build upon many ideas present in the literature, with fuller and less stylized models, and to drive a deeper understanding of financial frictions in the macroeconomy in new ways. We comment on how the methodology we presented above can be extended, and used fruitfully, in higher-dimensional state spaces. We also comment on the issues of uniqueness of equilibria and the characterization of the full set of equilibrium possibilities when multiple equilibria exist.

One key idea is that the wealth distribution in the economy matters. In the models we solved in Sections 2 and 3, the wealth distribution is characterized by a single state variable, the wealth share of experts η_t. When η_t is low, experts become undercapitalized. More generally, other sectors can become undercapitalized. Mian and Sufi (2009) argue that a big drag on the economy in the recent financial crisis has been the fact that many households are undercapitalized. Caballero et al. (2008) discuss how during Japan's lost

[r] Keynes' Paradox of Thrift states that an increase in the savings propensity can paradoxically lower aggregate savings. An increase in savings propensity lowers consumption demand. If the increased savings are "parked in (bubbly) money" instead of additional real investments, aggregate demand becomes depressed. This lowers aggregate income. Saving a fraction of now lower income can lower overall dollar savings.

decade it was the corporate sector that became undercapitalized. The general message here is that the wealth distribution across sectors matters for the level of economic activity—asset allocation—as well as the rates of earnings and risk exposures of various sectors. These earnings and risk exposures in turn drive the stochastic evolution of the wealth distribution.

The idea that the wealth distribution drives economic cycles is not new in the literature. Kiyotaki and Moore (1997) and Bernanke et al. (1999) consider the fluctuations of the wealth of a class of agents near the steady state. Of course, continuous-time methods facilitate a full solution of this type of a model. He and Krishnamurthy (2013) consider a model similar to the ones we presented here, but without asset misallocation and with a somewhat different assumption of the earnings of the households' holdings of expert equity.[s]

More broadly, several papers introduce the idea of intergenerational wealth distribution. This idea exists already in Bernanke and Gertler (1989), where the wealth of old entrepreneurs affects wages in the labor market, which in turn impact the accumulation of wealth by young entrepreneurs. Myerson (2012) builds a model with T generations of bankers, in which the wealth distribution evolves in cycles, causing cycles in real activities. When the wealth of old bankers is high, risk premia are low, and hence earnings of young bankers are low. Wealth distribution across sectors also matters. Brunnermeier and Sannikov (2015b) develop a rather symmetric model, in which there are two sectors that produce two essential goods, and either one of the sectors can become undercapitalized. Brunnermeier and Sannikov (2012) discuss the idea that multiple sectors can be undercapitalized, and that monetary policy can affect "bottlenecks" through its redistributive consequences. They envision an economy in which multiple assets are traded, and agents within various sectors hold specific portfolios, backed by a specific capital structure. Brunnermeier and Sannikov (2015a) provide formal backing of these ideas using a three-sector model, in which traded assets include capital, money and long-term bonds, and monetary policy can affect the prices of these assets (and hence affect the sectors that hold theses assets) in various ways.

This leads us to the obvious question about the capacity of continuous-time models to develop these complex ideas. Can continuous-time methods successfully handle models with multiple state variables, which describe, eg, the distribution of wealth across sectors together with the composition of productive capital? We believe that yes—we are highly optimistic about the potential of continuous-time models. Certainly, the curse of dimensionality still exists. However, models with as many as four state variables should be solvable through a system of partial differential equations in a matter of minutes, if not faster,

[s] In that model, households earn more than their required return, and therefore there is rationing of experts' shares. Effectively, the alternative assumption gives households some market power, which intermediaries do not have. This leads to a lower intermediary earnings rate and a slower recovery from crisis.

through the use of efficient computational methods. The authors of this chapter have some experience with computation, and on a personal level many possibilities seem feasible now which appeared out of reach 5 years ago. To gauge computational speed, DeMarzo and Sannikov (2016) solve a model with three state variables, using a system of two partial differential equations. In addition the procedure involves an integration step somewhat reminiscent of the "static step" of the procedure in Section 3.5. With 201 × 51 × 51 grid points, the procedure using the explicit(!) method takes only a minute to compute the optimal contract. The implicit method of solving partial differential equations, which we use to compute the examples in Section 3.6 is significantly faster. For example, when solving a partial differential equation of the parabolic type in two dimensions (all equations for computing the value function using the iterative method are parabolic), with N grid points in space, one needs $O(N^2)$ grid points in time to ensure that the computational procedure is stable, when using the explicit method. In contrast, when using the implicit method, stability does not depend on the length of the time step, ie, the time step can be kept constant when greater resolution is required along the space dimension. Hence, we believe that by making a claim that models with four state variables are feasible to solve, we are in fact quite conservative.

We think that the iterative method, based on value function iteration for each type of agent, should prove quite fruitful. This method is based on backward induction starting from a terminal condition on the state space. At each new time interval, we start with value functions computed for the end of the interval. These value functions determine the agents' incentives through their continuation values from various portfolio choices. As a result, we can determine at each time point the allocations of assets and risk consistent of equilibrium—this is the "static step"—and hence we can compute the value function one period earlier. We see this method as fairly general and suitable for multiple dimensions.

In contrast, the shooting method aims at solving for the fixed point—equilibrium value functions and allocations in an infinite-horizon economy—up front. The straightforward extension of this method to multidimensional state spaces may be difficult to implement, as one would have to guess functions that match boundary conditions on the entire periphery of the state space, instead of just two endpoints. Nevertheless, procedures that use variations of policy iteration may lead to an efficient way of solving for a fixed point.

What makes continuous-time models particularly tractable is that transitions are local (when shocks are Brownian)—hence it is possible to determine the agents' optimal decisions and solve for their value functions by evaluating only first and second derivatives. In discrete time, with discrete transitions, the agents' decisions at any point may depend on entire value functions.

What about environments with so many dimensions that the straightforward discretization of the state space makes computation infeasible, due to the curse of dimensionality? Here, we are curious about the idea of describing state variables through certain essential moments—following the suggestions of Krusell and Smith (1998). We have not processed this possibility sufficiently to comment on it in the chapter, but generally we are very eager to know about ways to choose moments that describe the state space in a meaningful way for a given model. We should say, however, that continuous time can be helpful here as well, for describing continuation values and prices as functions of moments.

We finish this section by discussing the question of equilibrium uniqueness in the model we presented and in more complex models we envision. First, consider a finite-horizon economy that we are solving for via an iterative procedure. The procedure has two steps—the time step of value function iteration and the static step that determines prices and allocation. The time step cannot be a source of nonuniqueness—given continuation values, transition probabilities and payoff flows, the value function one period earlier is fully determined. The static step may or may not lead to nonuniqueness. In the model of Section 3 there are multiple nonstationary equilibria. For example, at any time point, the price of capital q_t can jump. If q_t jumps up by 10% then the risk-free asset must have an instantaneous return of 10% as well to ensure that the markets for capital and the risk-free asset clear. Of course, by the market-clearing condition for output (43), the price of capital q_t must correspond to the allocation of capital $\psi_t \in [0, 1]$. The allocation itself must be justified by the local volatility of capital, so that all agents have incentives to hold their portfolios. However, the possibility of jumps opens up room to many possibilities.

We compute the Markov equilibrium, in which prices and allocations are functions of η. If so, then the price of capital $q(\eta_t)$ must satisfy the differential equation that follows from (41) and (42). Notice that there are two values of $\sigma + \sigma_t^q$ consistent with the quadratic Eq. (41), positive and negative. We select the positive value, since otherwise amplification is negative, in the sense that a positive fundamental shock would result in a drop in the value of capital. Hence, the equilibrium we compute is the unique Markov equilibrium, in which the return on capital is always positively correlated with fundamental shocks to capital.

In more general models, we envision that some of the same forces are present. We also anticipate that, when there are multiple equilibria, it may be of interest to characterize the whole set of equilibria via an appropriate recursive structure. To answer this question, one may need to construct/compute a correspondence from the state space to the vector of equilibrium payoffs of all agent types. We envision that this correspondence can be found recursively by solving for the boundaries of attainable equilibrium payoff sets backwards in time, but the details of this procedure are certainly work in progress.

ACKNOWLEDGMENTS

We are grateful to comments by Zhiguo He, Lunyang Huang, Ji Huang, Falk Mazelis, Sebastian Merkel, Greg Phelan, and Christian Wolf, as well as the editors John Taylor and Harald Uhlig.

REFERENCES

Achdou, Y., Han, J., Lasry, J.M., Lions, P.L., Moll, B., 2015. Heterogeneous agent models in continuous time. Working Paper, Princeton University.

Adrian, T., Boyarchenko, N., 2012. Intermediary leverage cycles and financial stability. Working Paper, Federal Reserve Bank of New York.

Adrian, T., Boyarchenko, N., 2013. Intermediary balance sheets. FRB of New York Staff Report, Number 651.

Aiyagari, S.R., 1994. Uninsured idiosyncratic risk and aggregate saving. Q. J. Econ. 00335533. 109 (3), 659–684. http://www.jstor.org/stable/2118417.

Angeletos, G.M., 2007. Uninsured idiosyncratic investment risk and aggregate saving. Rev. Econ. Dyn. 1094-2025. 10 (1), 1–30. http://dx.doi.org/10.1016/j.red.2006.11.001. http://www.sciencedirect.com/science/article/B6WWT-4MR7DG4-1/2/321ab3b301256e6f17bf2b4003c7218d.

Basak, S., Cuoco, D., 1998. An equilibrium model with restricted stock market participation. Rev. Financ. Stud. 08939454. 11 (2), 309–341.http://www.jstor.org/stable/2646048.

Bernanke, B., Gertler, M., 1989. Agency costs, net worth, and business fluctuations. Am. Econ. Rev. 79 (1), 14–31.

Bernanke, B., Gertler, M., Gilchrist, S., 1999. The financial accelerator in a quantitative business cycle framework. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics. 1, Chapter 21. Elsevier, Amsterdam, The Netherlands, pp. 1341–1393.

Bewley, T.F., 1977. The permanent income hypothesis: a theoretical formulation. J. Econ. Theory 0022-0531. 16 (2), 252–292. http://dx.doi.org/10.1016/0022-0531(77)90009-6. http://www.sciencedirect.com/science/article/B6WJ3-4CYGG80-1SC/2/301c685f23755550247618450b40f612.

Bewley, T.F., 1980. The optimum quantity of money. In: Kareken, J.H., Wallace, N. (Eds.), Models of Monetary Economies. Federal Reserve Bank of Minneapolis, pp. 169–210. http://minneapolisfed.org/publications_papers/books/models/pcc169.pdf.

Bianchi, J., 2011. Overborrowing and systemic externalities in the business cycle. Am. Econ. Rev. 101 (7), 3400–3426. http://dx.doi.org/10.1257/aer.101.7.3400.

Black, F., Scholes, M., 1973. The picing of options and corporate liabilities. J. Polit. Econ. 00223808. 81 (3), 637–654. http://www.jstor.org/stable/1831029.

Brock, W., Mirman, L., 1972. Optimal economic growth and uncertainty: the discounted case. J. Econ. Theory 4 (3), 479–513.

Brunnermeier, M.K., Sannikov, Y., 2012. Redistributive monetary policy. In: Jackson Hole Symposium.1, pp. 331–384.

Brunnermeier, M.K., Sannikov, Y., 2014. A macroeconomic model with a financial sector. Am. Econ. Rev. 104 (2), 379–421.

Brunnermeier, M.K., Sannikov, Y., 2015a. The I theory of money. Working Paper, Princeton University.

Brunnermeier, M.K., Sannikov, Y., 2015b. International credit flows and pecuniary externalities. Am. Econ. J. Macroecon. 7 (1), 297–338. http://dx.doi.org/10.1257/mac.20140054.

Brunnermeier, M.K., Sannikov, Y., 2016. On the optimal inflation rate. Am. Econ. Rev. 106 (5), 484–489.

Caballero, R.J., Hoshi, T., Kashyap, A.K., 2008. Zombie lending and depressed restructuring in Japan. Am. Econ. Rev. 98 (5), 1943–1977. http://dx.doi.org/10.1257/aer.98.5.1943.

Candler, G.V., 1999. Finite-difference methods for continuous-time dynamic programming problems. In: Marimon, R., Scott, A. (Eds.), Computational Methods for the Study of Dynamic Economies. Cambridge University Press, Cambridge, England, pp. 172–194.

Chandler, L.V., 1948. The Economics of Money and Banking. Harper Brothers Publishers, New York, US.

Cox, J.C., Ingersoll, J.E., Ross, S.A., 1985. A theory of the term structure of interest rates. Econometrica 53 (2), 385–408.

DeMarzo, P., Sannikov, Y., 2016. Learning, termination and payout policy in dynamic incentive contracts. Working Paper, Princeton University.

Diamond, P.A., 1965. National debt in a neoclassical growth model. Am. Econ. Rev. 00028282. 55 (5), 1126–1150. http://www.jstor.org/stable/1809231.

DiTella, S., 2013. Uncertainty shocks and balance sheet recessions. Working paper, Stanford University.

Drechsler, I., Savov, A., Schnabl, P., 2016. A model of monetary policy and risk premia. J. Finance (forthcoming).

Fernandez-Villaverde, J., Rubio-Ramirez, J.F., 2010. Macroeconomics and volatility: data, models, and estimation. Working Paper, University of Pennsylvania.

Fisher, I., 1933. The debt-deflation theory of great depressions. Econometrica 00129682. 1 (4), 337–357. http://www.jstor.org/stable/1907327.

Gurley, J.G., Shaw, E.S., 1955. Financial aspects of economic development. Am. Econ. Rev. 00028282. 45 (4), 515–538. http://www.jstor.org/stable/1811632.

He, Z., Krishnamurthy, A., 2012. A model of capital and crises. Rev. Econ. Stud. 79 (2), 735–777.

He, Z., Krishnamurthy, A., 2013. Intermediary asset pricing. Am. Econ. Rev. 103 (2), 732–770.

He, Z., Krishnamurthy, A., 2014. A macroeconomic framework for quantifying systemic risk. Working Paper, National Bureau of Economic Research.

Hicks, J., 1937. Mr. Keynes and the 'classics': a suggested interpretation. Econometrica 3 (2), 147–159.

Huang, J., 2014. Banking and shadow banking. Working Paper, Princeton University.

Isohätälä, J., Milne, A., Roberston, D., 2014. The net worth trap: investment and output dynamics in the presence of financing constraints. Working Paper.

Isohätälä, J., Klimenko, N., Milne, A., 2016. Post-crisis macrofinancial modelling: continuous time approaches. In: Emmanuel, H., Philip, M., John, O.S.W., Sergei, F., Meryem, D. (Eds.), Handbook of Post-Crisis Financial Modelling, Chapter 10. Palgrave Macmillan, London, UK, pp. 235–282.

Judd, K.L., 1998. Numerical Methods in Economics. MIT Press, Cambridge, MA.

Keynes, J.M., 1936. The General Theory of Employment, Interest and Money. Macmillan, London, UK.

Kindleberger, C.P., 1978. Manias, Panics, and Crashes: A History of Financial Crises. Basic Books, New York, US.

Kiyotaki, N., Moore, J., 1997. Credit cycles. J. Polit. Econ. 00223808. 105 (2), 211–248. http://www.jstor.org/stable/2138839.

Klimenko, N., Pfeil, S., Rochet, J.C., 2015. Bank capital and aggregate credit. Working Paper, University of Zürich.

Krusell, P., Smith Jr., A.A., 1998. Income and wealth heterogeneity in the macroeconomy. J. Polit. Econ. 00223808. 106 (5), 867–896. http://www.jstor.org/stable/2991488.

Kydland, F.E., Prescott, E.C., 1982. Time to build and aggregate fluctuations. Econometrica 00129682. 50 (6), 1345–1370. http://www.jstor.org/stable/1913386.

Leland, H., 1994. Corporate debt value, bond covenants, and optimal capital structure. J. Finance 49 (4), 1213–1252.

Maggiori, M., 2013. Financial intermediation, international risk sharing, and reserve currencies. Working Paper, NYU.

Mankiw, G., Romer, D., 1991. New Keynesian Economics, Vol. 1: Imperfect Competition and Sticky Prices. MIT Press, Cambridge, MA.

Markowitz, H., 1952. Portfolio selection. J. Financ. 00221082. 7 (1), 77–91. http://www.jstor.org/stable/2975974.

Mendoza, E.G., 2010. Sudden stops, financial crisis, and leverage. Am. Econ. Rev. 100, 1941–1966.

Mian, A., Sufi, A., 2009. The consequences of mortgage credit expansion: evidence from the US mortgage default crisis. Q. J. Econ. 124 (4), 1449–1496.

Minsky, H.P., 1957. Central banking and money market changes. Q. J. Econ. 00335533. 71 (2), 171–187. http://www.jstor.org/stable/1883812.

Mittnik, S., Semmler, W., 2013. The real consequences of financial stress. J. Econ. Dyn. Control. 37 (8), 1479–1499.

Moll, B., 2014. Productivity losses from financial frictions: can self-financing undo capital misallocation? Am. Econ. Rev. 104 (10), 3186–3221.

Moreira, A., Savov, A., 2016. The macroeconomics of shadow banking. J. Finance (forthcoming).

Myerson, R.B., 2012. A model of moral hazard credit cycles. J. Polit. Econ. 120 (5), 847–878.

Oberman, A.M., 2006. Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton-Jacobi equations and free boundary problem. SIAM J. Numer. Anal. 44 (2), 879–889.

Patinkin, D., 1956. Money, Interest, and Prices: An Integration of Monetary and Value Theory. Row, Peterson, Evanston, IL.

Phelan, G., 2014. Financial intermediation, leverage, and macroeconomic instability. Working Paper, Williams College.

Rappoport, D., Walsh, K., 2012. A discrete-time macroeconomic model with a financial sector. Mimeo, Yale University.

Samuelson, P.A., 1958. An exact consumption-loan model of interest with or without the social contrivance of money. J. Polit. Econ. 00223808. 66 (6), 467–482. http://www.jstor.org/stable/1826989.

Sannikov, Y., 2008. A continuous-time version of the principal-agent problem. Rev. Econ. Stud. 0034652775 (3), 957–984. http://www.jstor.org/stable/20185061.

Silva, D.H., 2016. The Risk Channel of Unconventional Monetary Policy. Working Paper, MIT.

Sims, C.A., 1980. Macroeconomics and reality. Econometrica 48 (1), 1–48.

Solow, R.M., 1956. A contribution to the theory of economic growth. Q. J. Econ. 00335533. 70 (1), 65–94. http://www.jstor.org/stable/1884513.

Stokey, N., Lucas, R., 1989. Recursive Methods in Economics Dynamics. Harvard University Press, Cambridge, MA.

Taylor, J.B., 1979. Estimation and control of a macroeconomic model with rational expectations. Econometrica 00129682, 14680262. 47 (5), 1267–1286. http://www.jstor.org/stable/1911962.

Tobin, J., 1969. A general equilibrium approach to monetary theory. J. Money Credit Bank. 1 (1), 15–29.

Werning, I., 2012. Managing a liquidity trap: monetary and fiscal policy. Working Paper, MIT.

Woodford, M., 2003. Interest and Prices: Foundations of a Theory of Monetary Policy. Princeton University Press, Princeton, NJ.

CHAPTER 19

Housing and Macroeconomics

M. Piazzesi[*,†], **M. Schneider**[*,†]
[*]Stanford University, Stanford, CA, United States
[†]NBER, Cambridge, MA, United States

Contents

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.06.003

Abstract

This chapter surveys the literature on housing in macroeconomics. We first collect facts on house prices and quantities in both the time series and the cross section of households and housing markets. We then present a theoretical model of frictional housing markets with heterogeneous agents that nests or provides background for many studies. Finally, we describe quantitative results obtained during the last 15 years on household behavior, business cycle dynamics, and asset pricing, as well as boom bust episodes.

Keywords

Booms and busts in housing markets, land prices, market segmentation, mortgages, residential investment, expectations, collateral constraints, illiquidity, transaction costs

JEL Classification Codes

R2, R3, E2, E3, E4, G1

1. INTRODUCTION

The first volume of the *Handbook of Macroeconomics*, published in 1999, contains essentially no references to housing. This statistic accurately summarizes the state of the field at the time. Of course, housing was not entirely absent from macroeconomic studies, which typically account for all production, consumption and wealth in an economy. The lack of references instead reflected the treatment of housing as simply one component of capital, consumption or household wealth that does not deserve special attention.

At the turn of the millennium, housing was implicitly present in three loosely connected literatures. One is work on aggregate fluctuations that studies the sources of business cycles and the response of the economy to fiscal and monetary policy. In the typical 20th century model, residential structures were part of capital, or sometimes "home capital" (together with consumer durables). Housing services were part of nondurables (or home good) consumption. Models of financial frictions and the role of capital as collateral focused on borrowing by firms. Volatility of house prices played no role—in fact, any volatility of asset prices was largely a sideshow.

Second, housing was implicitly present in the large body of work on asset pricing concerned with differences in average returns and price volatility across assets. Studies in this area used to largely stay away from properties of house prices and returns. At the same time, a common modeling exercise identified a claim to all consumption with equity and tried to explain the volatility of its price with a consumption-based stochastic discount factor. Housing thus played an implicit role as part of payoffs and risk adjustment. Finally, there is work on heterogenous households that seeks to understand the role of frictions and policy for inequality as well as distributional effects of shocks. Here, housing was included as a large implicit component of household wealth as well as a share of consumption.

The first half of the 2000s saw not only the largest housing boom in postwar US history, but also new research that introduced an explicit role for housing in macroeconomics. The new research studies the interaction of house prices and collateralized household borrowing with business cycles and monetary policy. It also explores how the role of housing as a consumption good as well as a collateralizable asset affects savings, portfolio choice, and asset pricing. By the time the US housing boom turned into a spectacular bust in 2007, housing was already a prominent topic in macroeconomics. The Great Recession added important new data points and further underscored the importance and unique properties of housing. As a result, housing now routinely receives special attention in macroeconomic discussions.

While the new literature grew out of the three lines of research described earlier, the focus on housing brought out several distinctive features. First, it naturally pushed researchers toward integration of themes and tools from all three lines of research. It is difficult to describe household behavior while ignoring uncertainty about house prices, or to think about mortgage debt without heterogeneous agents. Many papers surveyed below thus employ tools from financial economics to study exposure to uncertainty, and many quantitative models are analyzed with computational techniques that allow rich heterogeneity within the household sector.

The second feature is familiar from urban economics: "the housing market" is really a collection of many markets that differ by geography as well as other attributes. Disaggregating not only the household sector but also the housing stock provides valuable insights into the transmission of shocks and alters policy conclusions. For example, shocks to financial intermediaries or policies that change the cost of mortgage credit might have stronger effects on prices in markets where the typical buyer is also a borrower. Moreover, those shocks might have larger aggregate effects if their impact cannot be shared across subpopulation of agents. Availability of new large scale micro data sets has made it possible to explicitly study the interactions of many agents in many markets, and derive the aggregate effects of those interactions.

A third, related, feature is that the literature on housing has brought to bear a lot of evidence from the cross section of markets in a single episode to complement time series evidence that is common in macroeconomics. To illustrate, one can learn a lot about the role of technology shocks for residential investment from recurrent time series patterns in postwar history. In contrast, to assess the role of recent financial innovation for house prices, such patterns are less informative. Fortunately, though, we can learn from cross-sectional patterns in financing and prices across submarkets and types of households.

The literature shows how both time series and cross-sectional patterns on housing markets lend themselves to the same style of analysis that is common elsewhere in macroeconomics. Reduced form statistical tools are used to document facts and sometimes to isolate certain properties of equilibrium relationships. Insights on the quantitative importance of different mechanisms as well as policy counterfactuals are derived from multivariate structural models. In many ways, modeling the cross-sectional comovement in a single period of, say, mortgage borrowing and wealth across households and house prices across market segments, is conceptually similar to modeling the time series comovement of, say, residential and business investment, GDP and house prices in postwar history. Both exercises require tracing out the effect of exogenous variation in some features of the environment jointly on many endogenous variables.

This chapter describes work on housing in macroeconomics in three parts. Section 2 collects the new facts that emerge once disaggregation makes housing explicit. We first document business cycle properties of housing consumption, residential investment and mortgage debt. We then look at the dynamics of house prices at the national, regional and within-city level, and compare price volatility and trading volume for housing and securities. Finally, we document the dual role of housing as a consumption good as well as an asset in household portfolios.

Section 3 describes a theoretical framework that nests or provides background for many studies in the literature. It allows for four special features of housing that are motivated by facts from Section 2: *indivisibility*, *nontradability of dividends*, *illiquidity*, and *collateralizability*. Indeed, many homeowners own only their residence, directly consume its dividend in form of housing services and bear its idiosyncratic risk. Moreover houses are relatively costly to trade and easy to pledge as collateral even for individual households. In contrast, securities such as equity and bonds are typically held in diversified portfolios, have tradable payoffs, are traded often at low cost, and are harder to use as collateral, at least for individual investors.

Section 4 summarizes quantitative results derived from versions of the general framework over the last two decades or so. While no study contains all the ingredients introduced in Section 3, each one quantifies one or more of the tradeoffs discussed there. We start by reviewing work on consumption, savings and portfolio choice. We also consider mortgage choice and the role of financial innovation for household decisions. We then move on to general equilibrium analysis of the business cycle, monetary policy and asset prices. Finally, we consider boom–bust episodes, with an emphasis on the 1970s and 2000s US housing cycles.

We interpret results from different types of quantitative exercises in light of the general framework. One approach studies structural relationships with an explicit shock structure. For example, large bodies of work assess the ability of life cycle models of consumption, savings and portfolio choice to explain cross-sectional patterns as well as the ability of DSGE models to match time series patterns. An alternative approach investigates families of Euler equations for different agents and/or markets to reconcile allocations and asset prices. A third approach tries to isolate properties of the decision rules or the equilibrium law of motion with reduced form approaches.

What have we learned so far? We highlight here two key takeaways from the new literature that underlie the quantitative successes reported in detail below. First, *frictions matter*: Quantitative modeling of household behavior now routinely relies on collateral constraints, incomplete markets and transaction costs as key ingredients. Incompleteness of markets means in particular that homeowners bear property-level price risk. A large body of reduced form evidence provides additional support for this approach. Second, *heterogeneity of households matters*: Models with heterogeneous households and frictions introduce powerful new amplification and propagation mechanisms. In particular, they provide more scope for effects of shocks to the financial sector which have become important in accounts of postwar US history, especially the recent boom–bust cycle.

We also conclude that making housing explicit improves our understanding of classic macroeconomic questions, previously studied only with models that provide an implicit role for housing. For thinking about business cycles, the comovement and relative volatility of residential and business investment provide discipline on model structure. For thinking about asset pricing, the role of housing as a consumption good as well as a collateralizable asset generate the type of slow moving state variables for model dynamics that are needed in order to understand observed low frequency changes in the risk return tradeoffs for many assets, including housing itself. Finally, financial frictions in the household sector change the transmission of both aggregate and distributional shocks and policy interventions, especially to consumption.

At the same time, many open questions remain and there is ample opportunity for future research. One issue is the tradeoff between tractability and detail faced by any macroeconomic study. There are three areas in particular where more work is needed to converge on the right level of abstraction—with possibly different outcomes depending on the question. One is aggregation across housing markets: do we gain, for example, from building more models that treat the United States as a collection of small countries identified with, say, states or metropolitan areas? Another area is choosing dimensions of household heterogeneity: since observable demographic characteristics such as age, income, and wealth explain only a small share of cross-sectional variation, how should unobservable heterogeneity be accommodated? Finally, the majority of studies reviewed below capture financial frictions by assuming short term debt and financial shocks as changes to maximum loan-to-value ratios. Given the rich and evolving contractual detail we see in the data, what are the essential elements that should enter macroeconomic models?

A major outstanding puzzle is the volatility of house prices—including but not only over the recent boom–bust episode. Rational expectations models to date cannot account for house price volatility—they inevitably run into "volatility puzzles" for housing much like for other assets. Postulating latent "housing preference shocks" helps understand how models work when prices move a lot, but is ultimately not a satisfactory foundation for policy analysis. Moreover, from model calculations as well as survey evidence, we now know that details of the expectation formation by households—and possibly lenders and developers—play a key role. A promising agenda for research is to develop models of expectation formation that can be matched to data on both market outcomes and survey expectations. A final point is that most progress we report is in making sense of household behavior. The supply side of housing as well as credit to fund housing has received relatively less attention, another interesting direction for future work.

To keep the length of chapter manageable, we have narrowed focus along some dimensions where other recent survey papers already exist. In particular, the *Handbook* *of Urban and Regional Economics* contains chapters on search models of housing (Han and Strange, 2015) as well as US housing policy (Olsen and Zabel, 2015).[a] Since we focus on work that is already published, we have also left out much of the important emerging literature on the housing bust and Great Recession, as well as policy at the zero lower bound for nominal interest rates. Finally, our chapter deals almost exclusively with facts and quantitative studies about the United States. This reflects the focus of the literature, which in turn has been driven in part by availability of data. Another exciting task for future research is to use the tools discussed in this chapter to study the large variation in housing market structure and housing finance across countries, surveyed for example by Badarinza et al. (2016).

2. FACTS

2.1 Quantities

Fig. 1 plots the aggregate expenditure share on housing from the National Income and Product Account (NIPA) tables. The numbers in NIPA table 2.3.5 are based on survey data. The questionnaires in these surveys (for example, the Residential Finance Survey conducted by the Census Bureau) ask renters about their actual monthly rent payments. These payments are imputed to comparable owner-occupied units (Mayerhauser and Reinsdorf, 2007). The sample consists of quarterly data from 1959:Q1 to 2013:Q4.

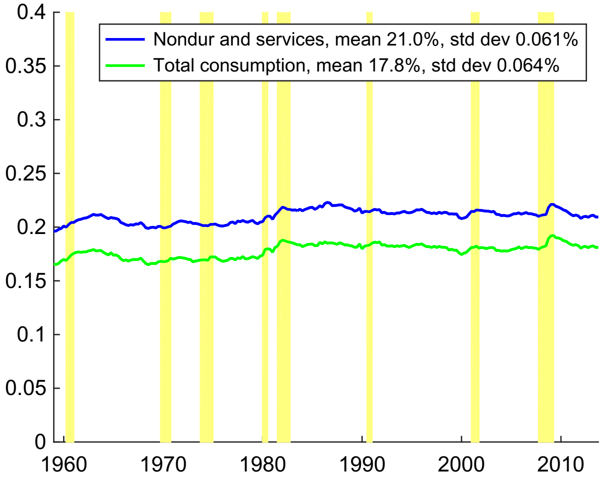


Fig. 1 Aggregate expenditure share on housing, 1959:Q1–2014:Q4.

[a] The same handbook contains a chapter on housing, finance and the macroeconomy (Davis and Van Nieuwerburgh, 2015) that also discusses some of the material covered in the present chapter.

We compute the expenditure share in two ways. The blue line shows housing expenditures as a fraction of expenditures on nondurables and services. This series has a mean of 21% and a standard deviation of 0.061%. The green line shows housing services as a fraction of total consumption (including durables). This series has a slightly lower mean of 17.8% and a bit higher standard deviation of 0.064%. The yellow bars indicate NBER recessions.

The overall impression from Fig. 1 is that the aggregate expenditure share is pretty flat over time. The expenditure share on housing is also similar across households in micro data, as shown by Piazzesi et al. (2007). Their table A.1 shows evidence from the Consumer Expenditure Survey, where the definition of housing expenditures depends on tenure choice. The CEX asks renters about their rent payments, while owner occupiers are asked about their interest payments on mortgages and other lines of credit, property taxes, insurance, ground rents, and expenses for maintenance or repairs. Davis and Ortalo-Magné (2011) use micro data on the expenditure share of renter households alone. The paper shows that individual expenditure shares based on the 1980, 1990, and 2000 Decennial Housing Surveys do not vary much within or across the top 50 US metropolitan statistical areas.

Fig. 2 plots three series: residential investment, nonresidential investment and output. The series are from NIPA table 1.1.3; they are all logged and detrended using the Hodrick–Prescott filter. The figure illustrates that both investment series are more volatile than output. Also, residential investment is twice as volatile as nonresidential investment. The volatility of residential investment is 9.7%, while nonresidential investment

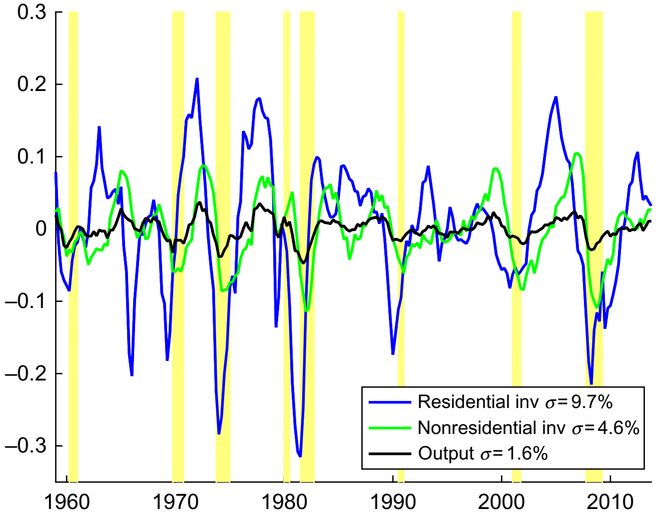


Fig. 2 Aggregate residential investment, nonresidential investment, and output; logged and detrended with Hodrick–Prescott filter.

has a volatility of 4.6% and the volatility of output is 1.6%. The figure also shows that the series for residential investment tends to increase before nonresidential investment and output, and it tends to decrease before the other two series. In other words, residential investment leads the cycle.

Once investment has created housing capital, it stays around for a long time. As reported by Fraumeni (1997), structures depreciate at rates of 1.5–3% per year. The depreciation rates for nonresidential capital are higher, between 10% and 30%. Moreover, housing combines housing capital with land, which is a fixed factor.

2.1.1 Constraints on the Supply of Housing

The degree to which new developments can increase the supply of housing varies across geographic areas. For example, developers in Indianapolis and Omaha find it easier to buy land and construct new homes than developers in San Francisco and Boston. There are two popular indices that carefully measure such housing supply constraints.

The first index is by Saiz (2010) and captures physical constraints. These geographical constraints capture two main features of land topology that make new developments difficult or impossible. The first feature is the presence of water. Saiz measures the area within 50 km from cities that is covered by oceans, lakes, rivers, and other water bodies such as wetlands. The second feature of land topology is steep slopes. Saiz computes the share of the area with a slope above 15% within a 50-km radius around an MSA.

The second measure of supply constraints captures regulatory restrictions. These are measured by the Wharton Residential Urban Land Regulation Index created by Gyourko et al. (2008). This index captures the stringency of residential growth controls in terms of zoning restrictions or project approval practices.

2.2 Prices

Fig. 3 shows the price–dividend ratio for stocks as a green line which measured on the left axis. The figure also shows the price–rent ratio for housing as a blue line with units indicated on the right axis. The figure illustrates the large volatility of the two series. The price–dividend ratio for stocks uses data from the Flow of Funds and represents the overall valuation of companies in the United States. The dividend series includes net repurchases. The price–dividend ratio fluctuates between 20 and 65, as measured on the left axis.

The numerator of the price–rent series for housing is the value of residential housing owned by partnerships, sole proprietors, and nonfinancial corporations, which are landlords for many rental units, as measured by the Flow of Funds. The denominator of the price–rent series is rents from the NIPA table 2.3.5, which includes actual rent payments as well as imputed rents for owner–occupiers (as discussed in the context of Fig. 1). The price–rent ratio fluctuates between 11 and 19, as measured on the right axis.

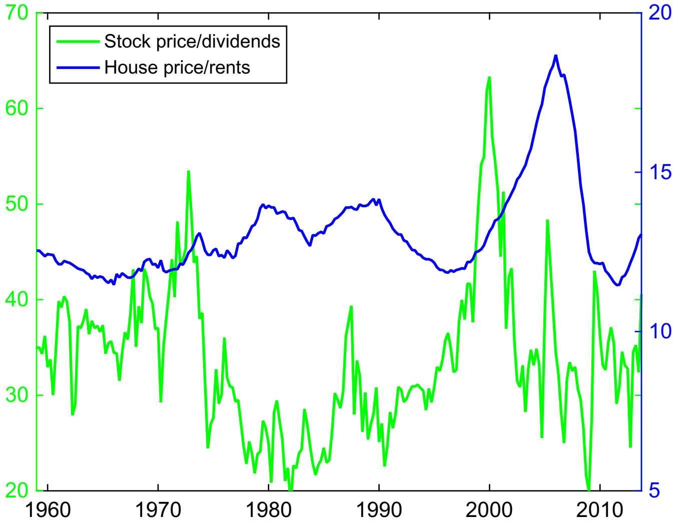


Fig. 3 Aggregate price/dividend ratio for stocks and price/rent ratio for housing.

The two valuation ratios often move inversely. For example, stocks tanked during the housing booms in the 1970s and 2000s. By contrast, stocks appreciated during the 1990s while housing did poorly. The recent boom–bust episode in housing stands out in the postwar experience.

2.2.1 Excess Volatility of Individual House Prices

House prices, like the prices of other assets, are highly volatile. The prices of individual houses are especially volatile. The volatility of various house price indices is smaller, but still a challenge for economic models—this is the excess volatility puzzle.

Most house price indices are constructed from repeat sales—average price changes in houses that sell more than once in the sample. CoreLogic constructs such city-wide indices for many metropolitan areas, various tiers of these markets, as well as the US national index. These indices are published as the S&P/Case–Shiller Home Price Indices by Standard & Poor's. The Federal Housing Finance Agency also constructs such indices from repeat sales or refinancings on the same properties (formerly called the OFHEO index). Zillow also publishes such indices for cities, states or the nation.

Case and Shiller (1989) estimate the standard deviation of annual percentage changes in individual house prices to be close to 15%. The paper concludes that individual house prices are similar to individual stock prices that are also very volatile. City-wide indices are less volatile than individual house prices. Flavin and Yamashita (2002) estimate a 14% volatility for individual house prices in their table 1A. Their table 1B reports a 4% volatility for Atlanta, 6% for Chicago, 5% for Dallas, and 7% for San Francisco. Landvoigt et al. (2015) estimate the volatility of individual house

Table 1 House price volatility

	Individual house	City	State	Aggregate
Volatility	14%	7%	5%	2–3%

Note: This table is from tables B1 and B2 in Piazzesi et al. (2007).

prices in different years. Their table 1 shows estimates that range between 8–11% during the 2000s boom and 14% during the bust.

Compared to stocks, which commove strongly with the aggregate stock market, a larger share of the volatility in individual house prices is idiosyncratic, as documented in Case and Shiller (1989). Their evidence stems from regressions of individual house price change on city-wide price changes. The regressions have low R^2s: 7% for Atlanta, 16% for Chicago, 12% for Dallas, and 27% for San Francisco.

Table 1 summarizes information from tables B1 and B2 from Piazzesi et al. (2007). The table illustrates the rule of thumb that 1/2 of the volatility in individual house prices is city-level variation, while 1/4 of the individual volatility is aggregate house price variation. This volatility decomposition illustrates the importance to understand the variation within narrow locations or individual houses. The high volatility of individual house prices together with high transaction costs lead to low Sharpe ratios (defined as average excess return on an asset, divided by its volatility) on housing. In other words, individual houses are not as attractive as an investment.

Idiosyncratic shocks to house prices are difficult to diversify. The problem with houses is that they are *indivisible*—they are sold in their entirety, not in small pieces. As a consequence, households own 100% of a specific house rather than small portions of many different houses. Moreover, the market for housing indices is not very liquid. In any given month, only a couple of futures contracts on city-wide house price indices trade on the Chicago Mercentile Exchange, if they trade at all.[b]

The ease of diversification distinguishes houses from other assets such as stocks. For example, households can save a small amount of money and invest it in a stock market index (such as the S&P 500) that tracks the value of a large stock portfolio. Alternatively, households can buy a few shares from several companies. The conventional wisdom in finance is that a small number of different stocks—such as five or six companies—are sufficient to achieve a high degree of diversification in a portfolio.

2.2.2 Momentum and Reversal

House prices have more momentum than other assets and also exhibit long-run reversal. The changes in log real prices of houses are more highly serially correlated compared to other assets. Case and Shiller (1989) provide the first evidence of such high serial

[b] The data on volume in these markets is here http://www.cmegroup.com/market-data/volume-open-interest/real-estate-volume.html

correlation. They document that a change in the log real price index in a given year and a given city tends to be followed by a change in the same direction the following year between 25% and 50% as large. Englund and Ioannides (1997) provide cross-country evidence where changes are followed by changes between 23% and 74% the next year. Glaeser et al. (2014) find changes the next year between 60% and 80%.

Cutler et al. (1991) compare the serial correlation in housing markets to that in other asset markets across many countries. For example, stocks, bonds and foreign exchange exhibit weak momentum for horizons less than a year. The monthly autocorrelation in excess stock returns is 10%, for US bonds it is 3%, 24% for foreign bonds, and 7% for foreign exchange. The excess returns on all these assets are essentially uncorrelated from year to year. In contrast, the excess returns on housing in their table 4 has an autocorrelation of 21% from year to year.

Over longer periods, house prices experience reversal. Englund and Ioannides (1997) document that changes in log real prices are followed by changes in the opposite direction after 5 years. Glaeser et al. (2014) also provide evidence of such reversal in their table 4. They estimate the autocorrelation of real house price changes over 5 years to be −0.80.

2.2.3 Predictable Excess Returns on Housing

The excess returns on many assets, including housing, are predictable. Case and Shiller (1989) show that excess returns on the city indices are predictable with excess returns in the previous year in their table 3. Case and Shiller (1990) provide further evidence of predictability for excess returns. Their table 8 runs regressions of city excess returns on rent–price ratios and construction costs divided by price. The coefficient on the rent–price ratio is positive: a high rent–price ratio predicts high excess returns over the next year.

Cochrane (2011) compares the predictability regressions for houses and stocks. Table 2 replicates his table 3. "Houses" in Table 2 refers to the aggregate stock of housing in the United States. "Stocks" refers to a value-weighted index of US stocks. The estimated slope coefficients indicate that high rents relative to prices signal high subsequent returns, not lower subsequent rents. The results for housing in the left panel look remarkably similar to those in the right panel for stocks. The returns are predictable for both, but dividend growth and rent growth are not predictable. The ratio of rents or dividends to prices is highly persistent, but stationary.

Campbell et al. (2009) decompose house price movements with the Campbell–Shiller linearization of the one-period return

$$r_{t+1} \approx \text{const.} + \rho(p_{t+1} - d_{t+1}) - (p_t - d_t) + \Delta d_{t+1},$$

where $r_{t+1} = \log R_{t+1}$ is the log housing return, $p_t = \log(P_t)$ is the log house price, $d_t = \log(D_t)$ is the log rent, $\Delta d_{t+1} = d_{t+1} - d_t$ is rent growth, and $\rho = 0.98$ is a constant in the approximation. This return identity simply says that high returns either come from higher prices (future $p - d$), lower initial prices, or higher dividends.

Table 2 House price and stock price regressions

	Houses			Stocks		
	b	***t***	***R*²**	***b***	***t***	***R*²**
r_{t+1}	0.12	(2.52)	0.15	0.13	(2.61)	0.10
Δd_{t+1}	0.03	(2.22)	0.07	0.04	(0.92)	0.02
dp_{t+1}	0.90	(16.2)	0.90	0.94	(23.8)	0.91

Note: This table is table 3 from Cochrane (2011). It reports results from regressions of the form

$$x_{t+1} = a + b \times dp_t + \varepsilon_{t+1}$$

where dp_t is either the log rent–price ratio in the left panel or the log dividend–price ratio in the right panel. In the left panel, x_{t+1} is either log annual housing returns r_{t+1}, log rent growth Δd_{t+1}, or the log rent–price ratio dp_{t+1} measured with annual data for the aggregate stock of housing in the United States, 1960–2010, from http://www.lincolninst.edu/subcenters/land-values/rent-price-ratio.asp In the right panel, x_{t+1} is either log stock returns r_{t+1}, dividend growth Δd_{t+1} , or the log dividend–price ratio dp_{t+1} measured with annual CRSP value-weighted return data, 1947–2010.

By iterating the return identity forward, we get the present value identity

$$dp_t \approx \text{const.} + \sum_{j=1}^{k} \rho^{j-1} r_{t+j} - \sum_{j=1}^{k} \rho^{j-1} \Delta d_{t+j} + \rho^k dp_{t+k}, \tag{1}$$

where $dp_t = d_t - p_t$ is the log rent–price ratio. The present value identity holds state-by-state as well as in expectation. Any movement in the rent–price ratio on houses therefore has to be associated with a movement in either the conditional expected value of future returns r_{t+j}, expected future rent growth Δd_{t+j} or a bubbly anticipation of future high prices dp_{t+k}.

Campbell et al. estimate a vector-autoregression that includes real interest rates, rent growth and excess returns on housing. The housing data are from various metropolitan regions and US aggregate data. Based on the estimated VAR, the paper evaluates the expected infinite sums of future returns and future rent growth on the right-hand side of Eq. (1) for $k \to \infty$ by imposing the no-bubble condition[c] $\lim_{k\to\infty} \rho^k dp_{t+k} = 0$. It finds that movements in price–rent ratios can be attributed to a large degree to time variation in risk premia and less so to expectations of future rent growth. The time variation in real interest rates does not explain price–rent ratio movements. Their fig. 2 also shows that the 2000s boom is hard to explain through the lense of their estimated VAR which predicts low price–rent ratios throughout the 2000s.

[c] Giglio et al. (2016) provide direct evidence on the no-bubble condition in housing markets by comparing the value of freeholds (infinite maturity ownerships of houses) with the value of leaseholds with maturities over 700 years in the United Kingdom and Singapore.

2.2.4 Value of Land vs Structures

Fig. 4 plots the value of the residential housing stock together with its two components, the value of the residential structures and the value of land. All series are from the Flow of Funds and are reported as multiples of GDP. The figure illustrates that movements in the value of the residential housing stock are mostly due to movements in the value of land. The value of structures fluctuates much less. The figure again highlights the importance of the recent boom–bust episode in the postwar housing experience.

Knoll et al. (2014) collect data on house values in many industrialized countries going back to 1870. The paper documents that real house values in most countries were largely constant from the 19th to the mid 20th century. Over the postwar period, real house prices approximately tripled. The majority of this increase in real house prices is associated with rising land prices, while real construction costs have been roughly constant.

There is also large cross-sectional variation in the share of land in the overall house value. A key component of this variation is what realtors call "location, location, location": each location is unique. There may be attractive locations with unique characteristics in fixed supply such as lake and oceanfronts, locations with strict zoning rules, outstanding amenities such as good schools or opera houses, low crime, etc. For example, table 4 in Davis and Heathcote (2007) reports that houses in San Francisco have a land share of 80.4% while houses in Oklahoma City have a land share of 12.6%. The table shows that areas with higher land shares tend to have higher house prices, higher average house price growth and more volatile house prices.

Another source of cross-sectional variation is differences in the durability and/or attractiveness of the existing structures. For example, the building material for structures in earthquake prone areas like California tends to be wood, which is cheaper and

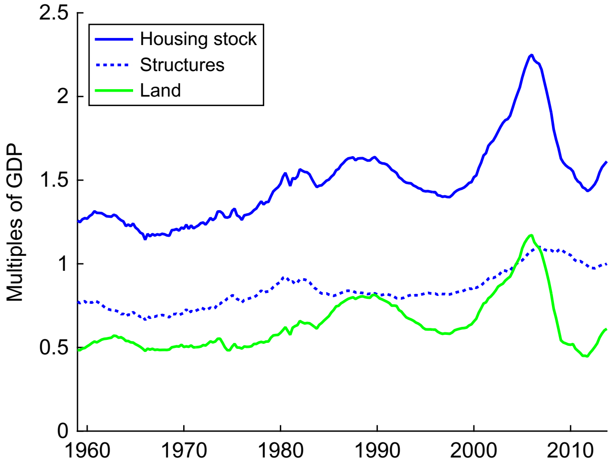


Fig. 4 The value of the residential housing stock together with its individual components, the value of residential structures, and the value of land.

deteriorates faster than brick which is used for most constructions in Pennsylvania. Architectural styles may also matter. For example, Victorian homes are valued at a premium, while 1950s postwar structures come at a discount.

2.2.5 Cross Section of House Prices

There are important cross-sectional patterns in house prices that help understand the variation across and within narrow areas. For example, during the 2000s, cheaper houses experienced a stronger boom–bust than more expensive houses. This pattern is different from previous boom–busts, where cheaper houses have experienced weaker boom–busts (such as in the 1970s.) Gentrification matters for poorer neighborhoods within a city that are in close proximity to more expensive neighborhoods. These low-price neighborhoods experience stronger booms–bust episodes than other low-price neighborhoods as well as high-price neighborhoods. Finally, the recent experience of the sand states challenges the notion that house prices in areas with an elastic housing supply should be less volatile.

Fig. 5 plots median house prices by city and tiers starting in the mid 1990s. The series are defined and constructed by Zillow Research. The top left panel shows that median house prices in the top tier of Los Angeles, California, gained 22% per year during the recent housing boom (1996–2006). The bottom tier gained *additional* 6 percentage points per year. During the housing bust (2006–11), the top tier made 4% capital losses per year, while the bottom tier dropped 5 percentage points more than the top tier.

The main stylized fact—houses in the bottom tier experienced a stronger boom–bust episode during the 2000s than houses in the upper tiers—can also be observed in other cities. In Las Vegas, houses in the bottom tier appreciated by 16% per year, while houses in the top tier only appreciated by 13%. During the bust, bottom-tier house prices fell by 14% while top-tier house prices fell only by 10% per year. In Chicago, capital gains across these tiers were the same on the way up, but there were larger losses in the lower tiers on the way down. In Omaha, the boom was not as pronounced, but still the bottom tier appreciated by 2 percentage points more than the top tier and was the only tier to experience a capital loss during the bust.

Landvoigt et al. (2015) estimate these patterns for the metro area of San Diego based on individual transaction data. The paper documents a roughly 20% difference between capital gains on the cheapest houses and most expensive houses between the years 2000 and 2005. The Zillow tiers group the cross section of houses and thereby reduce these cross-sectional differences. Kuminoff and Pope (2013) show a similar pattern for the land component of house values: cheap land appreciated more than expensive land during the 2000s boom.

Guerrieri et al. (2013) document that gentrification matters for poorer neighborhoods that are geographically close to high-price neighborhoods within a city. Their table 3

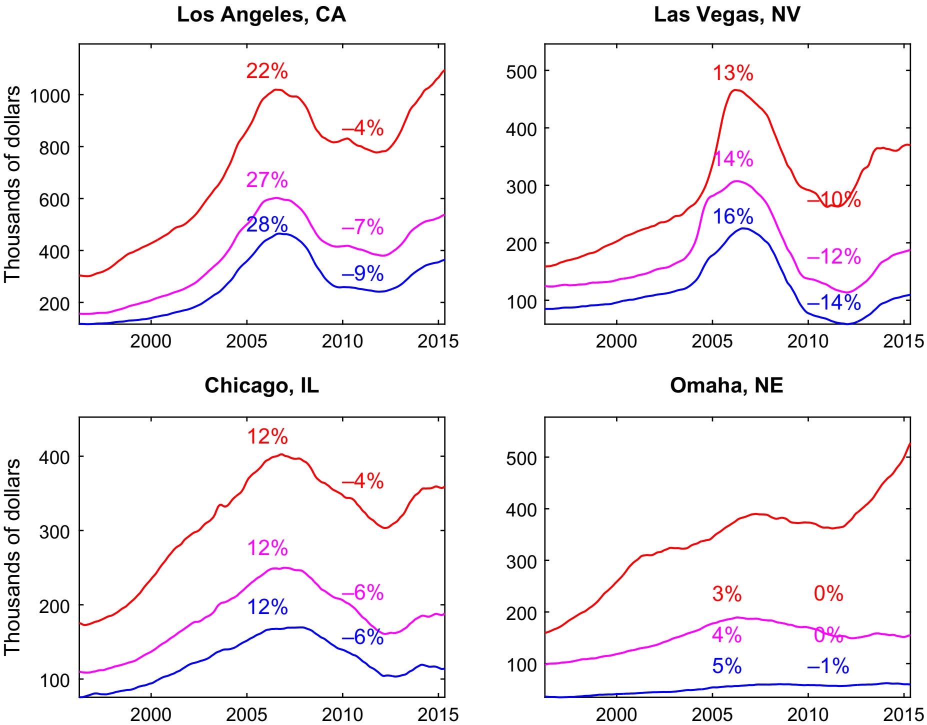


Fig. 5 Median house prices (in thousands of dollars) by city and tier: top tier (red (dark gray in the print version) line), medium tier (magenta (gray in the print version) line), and low tier (blue (black in the print version) line). The colored numbers indicate the tiered capital gains in percent per year during the housing boom (1996–2006) and during the bust (2006–11). *The data are from Zillow Research.*

shows that neighborhoods with an initially low price which were in close proximity to high-price neighborhoods appreciated more than otherwise similar initially low-price neighborhoods. For example, low-priced neighborhoods that were roughly 1 mile away from high-price neighborhoods appreciated by 12.4 percentage points more than low-priced neighborhoods that were roughly 4 miles away.

The recent experience in the "sand states"—Arizona, Florida, Nevada, and inland California—has challenged the notion that supply constraints amplify house price cycles. Fig. 1 in Davidoff (2013) shows that the magnitude of the house price cycle in the early 2000s in the sand states was larger than the cycle in coastal markets. His fig. 2 documents that the increase in the number of housing units was also larger in the sand states. Nathanson and Zwick (2015) argue that some cities, such as Las Vegas, do not have an abundance of land. Instead, these cities face long-run supply constraints in the form of tight virtual urban growth boundaries, formed by encircling federal and state lands.

2.3 Financing

Fig. 6 shows aggregate household debt from the Flow of Funds as multiple of GDP in the United States over the postwar period. The increase in the series happened in three discrete steps: right after World War II, the 1980s, and the 2000s. After the collapse of the housing market in 2006, households have been deleveraging. The red line in Fig. 6 is mortgage debt/GDP, which is roughly 3/5 of overall household debt. Most of household debt is thus collateralized. The plot shows that mortgage debt is chiefly responsible for the three discrete steps in which debt drastically increased. Household debt, especially mortgage debt, has also increased in other countries over the postwar period, as documented by Cardarelli et al. (2008). Jordà et al. (2016a) document this increase for many industrialized countries in a sample that goes back to 1870.

Jordà et al. (2016b) document that asset price boom–bust episodes that are combined with prior run-ups in leverage are associated with larger output costs during their bust. The data sample covers many industrialized countries going back to 1870. Moreover, boom–busts in housing have more severe output costs than those in equity markets.

2.3.1 Mortgage Growth During the 2000s

Mian and Sufi (2009) investigate who borrowed more during the 2000s. Did these borrowers expect higher future income growth? To address this question, Mian and Sufi use IRS data on income and mortgage debt data from the "Home Mortgage Disclosure Act" (HMDA). Their fig. 1 shows that income growth and mortgage growth are positively correlated across metro areas between 2002 and 2005 (in their top right panel). The evidence within metro areas, however, shows a *negative* correlation between income growth

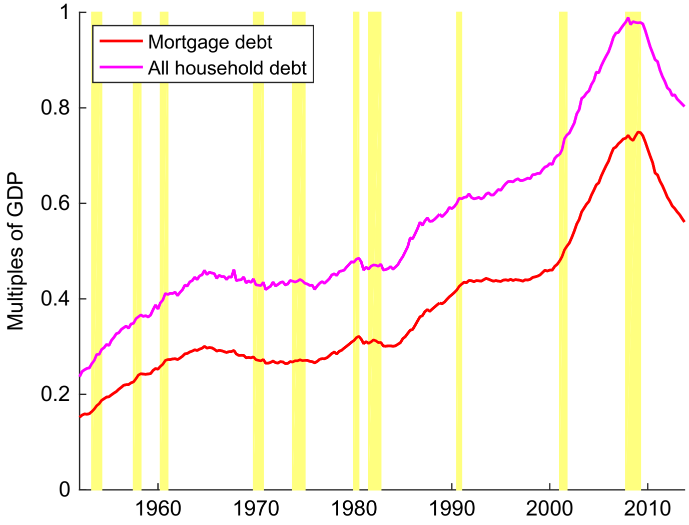


Fig. 6 Aggregate household debt and mortgage debt as fraction of GDP.

and mortgage growth across zip codes (in the lower right panel.) Moreover, they show that this negative correlation at the zip code level is unique to the 2002–05 period. These findings suggest that the 2000s were a unique episode in which mortgage debt increased in zip codes that experienced lower income growth.

Adelino et al. (2015) decompose mortgage growth into the extensive margin—the growth rate in the number of mortgages in a zip code—and the intensive margin—the growth rate in the size of individual mortgages. Their table 2 shows that the extensive margin is responsible for the negative correlation between IRS income growth and mortgage growth across zip codes. In fact, the intensive margin is *positively* correlated with IRS income at the zip code level. Moreover, Adelino et al. show that the growth rate of individual HMDA income—borrowers' income as indicated on their mortgage applications—is *positively* correlated to individual mortgage size across households. The paper argues that the negative correlation between income and mortgage growth documented by Mian and Sufi (2009) may be explained by a change in buyer composition (ie, richer buyers in poorer zip codes).

Mian and Sufi (2015) present evidence that the growth rate of HMDA income is higher than IRS reported income growth at the zip code level. They argue that the difference between the two growth rates represents mortgage fraud. Of course, the comparison of HMDA income and IRS income is tricky, because mover households who purchase a home have different characteristics than stayer households, especially during the 2000s boom. Table 2 in Landvoigt et al. (2015) compares the characteristics of home buyers and homeowners in 2000 Census data and 2005 data from the American Community Survey. They find that the median buyer in 2005 has more income and is richer than in 2000 in real terms.

Another important component of the increase in mortgage debt is existing homeowners who borrowed against the increased value of their house. Mian and Sufi (2011) document that especially homeowners in areas with stronger house price appreciation extracted equity from their houses with home equity lines of credit. Chen et al. (2013) report that a large fraction of refinancing during the 2000s were cash-outs, defined as more than 5% increases in loan amounts.

2.3.2 *Mortgage Contracts*

In the United States, the predominant mortgage contract is a fixed-rate mortgage with long maturity, usually 30 years. The main alternative is an adjustable-rate mortgage. In a basic adjustable-rate mortgage, the initial rate is set as a markup (or margin) on top of a benchmark, such as the 1-year Treasury rate. Adjustable rates are periodically reset to the current benchmark. During the recent housing boom, hybrid adjustable-rate mortgages became more popular. These hybrid contracts have a fixed rate for an initial period up to 10 years and adjusted periodically thereafter.

Campbell and Cocco (2003) report that fixed-rate mortgages accounted for 70% of newly issued mortgages on average during the period 1985–2001, while adjustable-rate mortgages accounted for the remaining 30%. The share of fixed-rate mortgages in new originations fluctuates over time. Fig. 2 in Campbell and Cocco (2003) shows the evolution of the share of fixed-rate mortgages, which is strongly negatively correlated with long-term interest rates.

Cardarelli et al. (2008), Andrews et al. (2011), and Badarinza et al. (2016) provide cross-country evidence on mortgage contracts. Table 4 in Andrews et al. shows that the typical mortgage maturity varies across countries between 10 years in Slovenia and Turkey to 30 years in Denmark and the United States. Table 3 in Badarinza et al. shows wide differences in the use of adjustable-rate mortgages and prepayment penalties. For example, the majority of mortgages in Australia, Finland, Portugal and Spain have an adjustable rate, while Belgium, Denmark, Germany, and the United States have mostly fixed-rate mortgages. Belgium and Germany have prepayment penalties, which make these fixed-rate mortgages highly risky. Table 3.1 in Cardarelli et al. (2008) shows that the countries with the largest fractions of securitized mortgages are the United States, Australia, Ireland, Greece, United Kingdom, and Spain.

2.3.3 Recent Financial Innovation and Lender Incentives

Leading up to the recent housing boom, the banking sector underwent a profound transformation. The traditional role of banks was to originate mortgages and hold them on their books until they are repaid. More and more, modern banks "originate-to-distribute"; banks originate mortgages, pool and tranche them, as resell them via the securitization process. In other words, mortgages are not kept on the balance sheet of the originating bank but are sold to investors. This transformation of the banking sector has changed the incentives of banks to screen mortgages. The resulting decline in lending standards has lead to a large expansion in credit.

Financial innovation also helped create new types of mortgages. Many mortgage contracts were designed to defer amortization, for example, with teaser rates or no interest rate payments during an initial period (such as "2–28 mortgages"). The share of alternative mortgages increased from below 2% until 2003 to above 30% during the peak years of the US housing boom (as documented, for example, in fig. 1 of Amromin et al., 2013). Another aspect of the deterioration of lending aspects were "no doc" loans, which did not require any documentation of income, or NINJA ("no income, no job or assets") loans.

Keys et al. (2010) provide evidence that securitization was associated with laxer screening of mortgages. The idea of the paper is to compare the performance of mortgages that are securitized with those that are not securitized. Since the 1990s, credit scoring has become the key tool to screen borrowers. The guidelines established by the government-sponsored enterprises, Fannie Mae and Freddie Mac, cautioned against lending to risky borrowers with a FICO score below 620. The 620 cutoff is also important for securitization as mortgages above the cutoff are easier to securitize. The paper studies the performance of a million mortgages over the years 2001–06. It finds that mortgages with a FICO score right above 620 performed *worse* than mortgages slightly below the 620 cutoff.

2.4 Market Structure

Housing has *broad ownership*. Roughly two thirds of US households own a house. Over the postwar period, the home ownership rate varied between 62% and 69%. It peaked at 69.2 at the end of 2004, toward the peak of the recent boom. The current ownership rate is down to 63.7%.

More households own a house than stocks. The ownership rate for stocks crucially depends on whether indirect holdings (through mutual funds and pension funds) are included or not. But even if we include indirect holdings, the ownership rate for stocks is below 50%.

Housing markets are *illiquid* relative to other asset markets. Turnover (per year) in housing markets is low relative to the stock market. The average turnover rate in the stock market is 110%, which means that every stock changes hands at least once in any given year. By contrast, the average turnover rate in the housing market is only 7%. This illiquidity is manifested in the fact that time on market—the number of days or months between listing and selling a house—is a key statistic in housing markets, while time on market plays no role in stock markets.

An important aspect of housing is that it is more *difficult to short* than other assets such as stocks. Because houses are unique and indivisible, an investor may not be able to take a short position in a particular house. The low liquidity in house price indices and their derivatives makes it either impossible or costly to take large short positions in the overall market. It is possible to short REITs, which are indexed to the value of commercial real estate. However, REITs are not perfectly correlated with the value of residential real estate. During recent housing booms, investors have used creative strategies to short housing. For example, during the recent housing boom, investors were short in mortgage-backed securities. In the ongoing Chinese boom, investors short the stock of large developers. Many of these investment strategies are costly and require sophistication, and are not perfect shorts for residential real estate.

Bachmann and Cooper (2014) document a secular decline in the turnover rate (the sum of their owner-to-owner and renter-to-owner moves) from the mid 1980s to 2000s in data from the Panel Study of Income Dynamics (PSID). Moreover, the paper documents that the turnover rate (in particular, the rate of owner-to-owner moves) is procyclical. Kathari et al. (2013) document a secular decline in moving rates of both renters and owners since the mid 1980s based on the Current Population Survey.

2.5 Household Portfolios

A sizeable literature uses various household level data sets to document cross-sectional patterns in housing consumption and the role of housing and mortgages in household portfolios. We summarize here key cross-sectional patterns that have been fairly stable over time. In particular, housing choices depend significantly on age and net worth.

It is well known that expenditure on nondurable consumption is hump-shaped over the life cycle (eg, Deaton, 1992). Fernandez-Villaverde and Krueger (2007) document a similar hump-shaped life cycle pattern for expenditure on durables. Their definition of durables includes purchases of consumer durables as well as housing expenditure by renters and owners in the CEX. Their fig. 6 shows that the hump peaks roughly at the age of 50 years, similar to the pattern for nondurables. After the peak, durables expenditure declines substantially with age. For example, durables expenditure at age 50 is twice as large as expenditure at 75.

Yang (2009) distinguishes expenditure on housing from that on other durables. For renters, housing expenditure is from CEX data. For owners, housing expenditure is from the SCF, assuming that expenditure is proportional to house value. Her fig. 4 shows that housing expenditure for owners also increases with age similar to durable expenditures. However, it peaks later in life—at age 65—rather than at age 50. Moreover, housing expenditure flattens out after age 65; unlike durable expenditure, it does not decline with age.

The homeownership rate is also hump-shaped over the life cycle. For example, table 6 in Chambers et al. (2009b) shows the homeownership rate first increases from roughly 40% for young households (aged 20–34 years) to twice that share for older households (aged 65–74 years). The homeownership rate then declines slightly for very old households.

The homeownership rate also increases with income. For example, Gyourko and Linneman (1997) study decennial census data from 1960 until 1990 to show that homeownership rates increase with income even after conditioning on age. There is also evidence that low income and minority households are less able to sustain homeownership than high income and white households. For example, Turner and Smith (2009) examine data from the PSID spanning the years 1970–2005 and document that homeowners in these groups have consistently higher exit rates from ownership.

The portfolio share on housing depends on both age and wealth. It declines monotonically with age. Young households are *house poor*: they choose highly leveraged positions in housing. As they age and accumulate wealth, they lower their portfolio weight on housing and pay down their mortgages. For example, table 2 in Flavin and Yamashita (2002) shows that young homeowners (aged 18–30) have an average portfolio weight of 3.51 on housing and −2.83 on mortgages in the PSID. Middle-aged households (aged 41–40 years) have an average weight of 1.58 on housing and −0.88 on mortgages. Older households (aged 71+) have an average weight of 0.65 on housing and −0.04 on mortgages.

The portfolio share on housing is hump-shaped in wealth. For example, table 1 of Campbell and Cocco (2003) shows that households in the bottom third of the wealth distribution are renters—they do not own a home, so their portfolio share on housing is zero. Wealthier households have a large fraction of their wealth, between 60% and 70%, invested in housing. For rich households (in the top 20% of the wealth distribution), the portfolio share on housing rapidly declines with wealth. These households shift more and more of their portfolio into stocks.

Wealth is also hump-shaped over the life cycle. Fig. 7 in Piazzesi and Schneider (2009a) uses the Survey of Consumer Finances to document the hump in wealth for middle-aged households (aged 53 years). The figure plots wealth of "rich households"—defined as the top 10% of net worth in their cohort—separately from cohort totals. These rich households own more than half of the cohort wealth—indicating a high concentration of wealth.

The hump in wealth over the life cycle multiplied by portfolio shares on housing that decline with age results in a hump-shaped pattern in housing wealth over the life cycle (third left panel in fig. 7 of Piazzesi and Schneider, 2009a). This housing wealth is somewhat concentrated—rich households own roughly a third of the housing wealth in their cohort. However, most of the overall wealth concentration can be attributed to the extremely high concentration of wealth invested in stocks: rich households own almost all of the stock wealth in their cohort.

3. THEORY

This section describes a theoretical framework that nests or provides background for many studies in the literature. At its heart is the intertemporal household decision problem with housing as both an asset and a consumption good. The papers discussed below all share a version of this problem. They differ in what other aspects of housing are included—in particular, the option to rent, collateral constraints or transaction costs—in whether equilibrium is imposed and, if yes, in how the supply side is modeled.

We thus begin with a "plain vanilla" household problem. It assumes that houses of every quality as well as other assets and consumption of the nonhousing good are all traded in competitive markets. The only friction is that consumption of housing services requires ownership of a house. Housing thus differs from other assets because of *indivisibility* and *nontradability of dividends*. Indeed, households hold either zero or one units of the housing asset and the "dividend"—that is, the value of housing services less maintenance cost—cannot be sold in a market to other households.

After introducing the plain vanilla problem, we discuss household optimization, derive asset pricing equations and define an equilibrium with a fixed aggregate supply of housing services. Here, we highlight the distinction between an exogenous distribution of house qualities and a fixed stock of housing that developers can costlessly convert into one of many distributions with the same mean. We also discuss the role of expectation formation. In later sections, we then add further key ingredients one by one: production and land, a rental market, collateral constraints and transaction costs.

3.1 Basic Setup

We work in discrete time. Studies differ in how long the economy lasts and what households' planning horizons are. To explain the basic tradeoffs, these details are not important, so we do not take a stand on them now. Instead, we focus on the period t decisions of a household who expects to also live in period $t+1$. Studies also typically assume a large number of different households who may differ in characteristics such as age, income or beliefs. We do not make such heterogeneity explicit, but instead describe a generic household problem with minimal notation.

To represent uncertainty, we fix a probability space $(\Omega, \mathcal{F}, P^0)$. The set Ω contains states of the world. Events in the σ-field $\mathcal{F}$ correspond to all exogenous events that can occur. For example, each state of the world could imply a different sequence of shocks to a household's income over his lifetime. The probability measure P^0 says how likely it is that each event $F \in \mathcal{F}$ occurs. In other words, it tells us with what probability nature draws a state of the world $\omega \in F$. In general, the "physical" probability P^0 need not coincide with the belief of a household.[d]

3.1.1 Preferences

The evolution of the households' information is summarized by a filtration $\mathcal{F}_t$ on Ω: $F \in \mathcal{F}_t$ means that the household knows in period t whether event F has occurred or not. The household's belief about states of the world is described by a probability P. In what follows, we keep these objects in the background and instead work directly with random variables and conditional moments. Our convention is that random variables dated t are contained in the household's period t information set. For example, c_t is (random) consumption of nonhousing goods and we write $E_t c_{t+1}$ for the household's expected period $t+1$ consumption given period t information.

Households derive utility from housing services s and other consumption c. Utility is state and time separable; in particular, period t utility from the two goods is given by

$$U(g(s_t, c_t)),$$

where $g: \mathbb{R}^2 \rightarrow \mathbb{R}$ is an "aggregator function" that is homogeneous of degree one and $U: \mathbb{R} \rightarrow \mathbb{R}$ is strictly increasing and concave. Decomposing utility in this way helps

[d] The physical probability is what one would use to compute or simulate the distribution of outcomes of the economy. It thus coincides with the belief of an outside observer, for example, an econometrician, who observes a large sample of data generated from the model.

distinguish substitution across goods within a period from substitution of consumption bundles $g(s_t, c_t)$ across periods.

The aggregator g describes households' willingness to substitute housing services for other consumption within a period. A common example is the CES functional form

$$g(s_t, c_t) = \left(c_t^{(\varepsilon-1)/\varepsilon} + \omega s_t^{(\varepsilon-1)/\varepsilon}\right)^{\varepsilon/(\varepsilon-1)}, \tag{2}$$

where ε is the *intra*temporal elasticity of substitution and ω is a constant. Agents are more willing to substitute within the period the higher is ε. As $\varepsilon \to \infty$, the two goods become perfect substitutes and as $\varepsilon \to 0$, they become perfect complements. The limit $\varepsilon \to 1$ represents the Cobb–Douglas case with constant expenditure shares.

The function U captures agent's willingness to substitute consumption *bundles* g over time (as well as states of nature). A common example is the power function $U(g) = g^{1-1/\sigma}/(1-1/\sigma)$ where σ is the *inter*temporal elasticity of substitution among bundles at different points in time. For $\sigma \to 0$, households want to maintain a stable bundle over time whereas for $\sigma \to \infty$ utility becomes linear in bundles. The limit $\sigma \to 1$ corresponds to logarithmic utility. With a CES aggregator, the special case $\sigma = 1/\varepsilon$ results in utility that is separable across the two goods.

While our assumptions on utility are convenient for exposition, several straightforward extensions are also common in the literature. First, some papers replace time separable utility by recursive utility, for example, the tractable functional form introduced by Epstein and Zin (1989). To deal with multiple goods, the usual recursive utility formulation is applied directly to bundles aggregated by g.[e] Second, some papers add preference shocks; in particular, a "housing preference shock" is often introduced via a random weight ω in (2). Finally, labor is often added as a third good in utility.

3.1.2 Technology

Households obtain housing services by living in exactly one house. Houses come in different qualities $h \in \mathcal{H} \subset \mathbb{R}$ where the set $\mathcal{H}$ can be either discrete or continuous. Our convention is that $\mathcal{H}$ may contain zero to accommodate households who do not live in a house. A household who lives in a house of quality h_t from t to $t + 1$ obtains a flow of housing services $s_t = h_t$ that enters period t utility. In quantitative applications, the flows s_t and c_t are typically identified with the household's consumption over a time range

[e] Formally, let $W : \mathbb{R}^2 \to \mathbb{R}$ denote a function that captures substitution over time and let $\nu : \mathbb{R} \to \mathbb{R}$ denote a function that captures aversion to risk about utility gambles. Utility from a consumption process (c_t, g_t) is defined recursively by

$$U_t = W\left(g(c_t, s_t), \nu^{-1}(E_t[\nu(U_{t+1})])\right).$$

Our time separable case obtains if $\nu = U$ and $W(x, y) = U(x) + \beta U(y)$. Epstein and Zin propose a CES aggregator for W and a power function for ν.

that includes date t, and the quality of his residence h_t is an average over that time range. Our timing convention implies that the house quality h_t relevant for period t consumption is chosen based on the period t information set.[f]

The one-dimensional quality index h orders houses from low to high qualities. In general, it captures many characteristics of a house—its location, the size of the land, square footage of lot and structure, its view, amenities, etc. The underlying assumption is that households agree on the ranking of all houses within the housing market that is being studied. At the same time, households may differ in their taste for house quality relative to other consumption and hence be willing to pay different amounts for any given house.

A household who lives in a house of quality h_t from t to $t+1$ must undertake maintenance worth $I(h_t)$ units of the other (nonhousing) good. The quality of the house then evolves over time according to

$$h_{t+1} = H_{t+1}(h_t), \tag{3}$$

where the subscript $t+1$ indicates that the evolution may be random. We highlight two popular special cases. The first assumes that all depreciation is "essential maintenance" without which the house is uninhabitable. As long as essential maintenance is performed, house quality is constant, that is, $I(h_t) = \delta_h h_t$ and $H(h_t) = h_t$. A second special case is that households do not pay for maintenance but average quality deteriorates geometrically, that is, $I(h_t) = 0$ and $H(h_t) = (1-\delta_h)h_t$. In both cases, $\delta_h h_t$ is depreciation of housing. The first approach is convenient when the set of qualities $\mathcal{H}$ is finite.

3.1.3 Housing Markets

Houses are traded in competitive markets. The only friction is that consumption of housing services requires ownership of a house. Housing thus differs from other assets because of *indivisibility* and *nontradability of dividends*. Indeed, it is held in indivisible units and its "dividend"—that is, the value of housing services less maintenance cost—cannot be sold in a market. This assumption is relaxed in Section 3.6 where we introduce a market for rental housing. In line with our timing convention, utility from a house bought at date t is enjoyed at date t itself—date t house prices are thus "cum dividend."

A house of quality h_t trades in period t at the price $p_t(h_t)$, denominated in units of the nonhousing good which serves as numeraire. The price function is increasing in quality. If the set $\mathcal{H}$ consists of a finite number of house types, then house prices can be summarized by a vector. With a continuum of qualities, it often makes sense to assume that the price function is smooth—a small change in quality leads to a small change in price. For example, in some applications the price function is linear, that is, there is a number $\bar{p}_t$ such that $p_t(h) = \bar{p}_t h$ for all quality levels h.

[f] Alternative timing conventions are possible and sometimes used in the literature. For example, we might assume that quality chosen at date t yields housing services only at date $t+1$.

3.1.4 What Is a House?

The setup emphasizes indivisibility and quality differences: housing services are provided by a distribution of housing capital stocks of different qualities, one for each household. In general, pricing is nonlinear: each quality level represents a different good and relative prices depend on relative demand and supply. This approach goes back to Rosen (1974) who studied competitive equilibrium with consumers who choose one "design" of a product that is identified by a vector of characteristics.

Braid (1981, 1984) and Kaneko (1982) studied housing with a one-dimensional quality index in static models with a continuum and a finite set of qualities, respectively. Caplin and Leahy (2014) characterize comparative statics of competitive equilibria in a static setting with a finite number of agents and goods. The dynamic setup here follows the finite quality models in Ortalo-Magné and Rady (1999, 2006) and Ríos-Rull and Sánchez-Marcos (2010) as well as the continuum approach in Landvoigt et al. (2015).

At first sight, allowing for nonlinear pricing may appear unnecessary: why not assume that there is a homogenous housing capital good—akin to physical capital in many macroeconomic models—with households choosing different quantities of that good at a common per-unit price? The latter approach is a special case of the setup that obtains when some market participants can convert houses of different quality with a marginal rate of transformation of one. For example, in Section 3.4, we derive it from the presence of a developer sector who undertakes this activity.

Work on housing has more often gone beyond setups with homogeneous capital and linear pricing than work on, say, business capital. One likely reason is measurement. The difficulties with measuring house prices in national accounts have been discussed frequently. At the same time, new micro data provide evidence on price dynamics at fine levels of disaggregation by geography and type of house. The evidence in Section 2.2 suggests that linear pricing is perhaps too restrictive, since both volatility conditional on quality is high and conditional means vary systematically by quality. We return to this issue below.

While our setup nests essentially all specifications in the macro literature, it is restrictive in at least two ways. First, it may not be possible or desirable to represent the cross section of houses by a one-dimensional index. A more general approach could follow Rosen (1974) and directly model preference over many characteristics. In particular, households may rank houses differently because they disagree about the weighting of characteristics. Second, a more general approach to household capital accumulation might start from an evolution equation

$$h_{t+1} = z_{t+1} H(h_t, I_t),$$

where z_{t+1} is a depreciation shock. In this equation, initial quality h_t and improvements I_t are imperfect substitutes, so that upkeep of the house is an explicit margin for the household. This approach could generate a distribution of houses in different states of disrepair.

3.2 Household Choice

We now consider the household's decision problem when houses as well as other assets and consumption of the nonhousing good are all traded in competitive markets. The household receives an exogenous labor income stream y_t. Securities, such as equity or bonds, trade at date t at prices collected in a $J \times 1$ vector q_t and provide payoffs at date $t+1$ summarized by a $J \times 1$ vector π_{t+1}. For long-lived securities such as equity, the payoff may contain the date $t+1$ price. We make no further assumption on market structure. Markets may be incomplete in the sense that it is not possible for households to assemble a portfolio of securities in period t with payoff equal to any given consumption plan that depends on date $t+1$ information. With incomplete markets, households may not be able to insure against future labor income risk.

3.2.1 Recursive Household Problem

Without trading frictions, past portfolio choice—including housing choice—affects the household at date t only through its effect on wealth. We can thus formulate the problem recursively with a single endogenous state variable *cash on hand w* that comprises housing wealth, other wealth plus income from labor and securities. To start off the recursion, we define a terminal value function $V_T(w_T)$. In a finite horizon life cycle problem, this function captures utility at the end of life, perhaps including bequests. In an infinite horizon setup ($T=\infty$), existence of a value function can be derived from trading restrictions that prevent Ponzi schemes.

For a household who expects to live for an additional period, the Bellman equation is

$$V_t(w_t) = \max_{c_t,\,,\theta_t;h_t \in \mathcal{H}} U(g(c_t,h_t)) + \beta E_t[V_{t+1}(w_{t+1})] \tag{4}$$

$$c_t + p_t(h_t) + I(h_t) + \theta_t^\top q_t = w_t,$$

$$w_{t+1} = \theta_t^\top \pi_{t+1} + p_{t+1}(H_{t+1}(h_t)) + y_{t+1}.$$

The first condition is the current budget constraint that says how cash on hand is split into consumption, asset purchases and maintenance. The second constraint describes the evolution of cash on hand which depends on future security payoffs, house value and labor income.

The same Bellman equation works for problems with random horizon. Indeed, a common approach assumes that households survive with a probability that can depend on age. Those survival probabilities are then used in computing the conditional expectation in the Bellman equation. In the terminal period of life, households learn that this is their last period, sell all assets and either consume the proceeds or transfer wealth to children. Given our timing convention on housing services and the need for ownership, we also assume that households do not consume housing services in the terminal period of life.

3.2.2 Two Stage Solution Approach

We consider household choice in two stages. The household first decides on house quality and thus how much of cash on hand to spend on housing. In a second stage, he allocates the remaining funds to numeraire consumption and securities. The split is helpful because indivisibility and nontradability make the housing choice special. On the one hand, indivisibility means that house quality may be discrete and the pricing of house quality may be nonlinear. On the other hand, nontradability means that housing and securities are imperfect substitutes even if there is no risk—a case when all other securities become perfect substitutes.

We write the second stage problem with returns and portfolio weights, rather than asset prices and quantities. The gross return on the jth security is $R_{t+1,j} = \pi_{t+1,j}/q_{t,j}$. We assume that the Jth security is a risk-free bond and denote the gross risk-free rate by R_t^f. Moreover, the returns on securities $j = 1, \ldots, J-1$ are risky and collected in a vector R_{t+1}. The household selects a portfolio weight $\alpha_{t,j}$ for each of the risky assets j. These weights are collected in a $J-1$ vector α_t, so that $1-\alpha_t^\top \iota$ is invested in risk-free bonds, where ι is a $J-1$ vector of ones. There are no restrictions on the sign of the portfolio position: the households can short a risky asset by choosing $\alpha_{t,j} < 0$ or borrow at the riskless rate by choosing $\alpha_t^\top \iota > 1$. The return on the portfolio is $\widetilde{R}_{t+1}(\alpha_t) = \alpha_t^\top R_{t+1} + (1-\alpha_t^\top \iota)R_t^f$.

The second stage problem is

$$\widetilde{V}_t(\widetilde{w}_t, h_t) = \max_{c_t, \alpha_t} U(g(c_t, h_t)) + \beta E_t[V_{t+1}(w_{t+1})]$$

$$w_{t+1} = (\widetilde{w}_t - c_t)\widetilde{R}_{t+1}(\alpha_t) + p_{t+1}(H_{t+1}(h_t)) + y_{t+1}. \tag{5}$$

The household starts with cash $\widetilde{w}_t = w_t - p_t(h_t) - I(h_t)$ left over after housing expenditure. He then chooses numeraire consumption c_t and invests the remaining funds $\widetilde{w}_t - c_t$ in securities. Cash next period consists of savings in securities multiplied by their average return plus the payoffs from housing and human capital, both of which are nontradable assets in the second stage problem.

Optimal choice depends on risk in house values, labor income and securities returns. To illustrate, we perform a second-order Taylor expansion of the future value function in (5) around expected future wealth to obtain

$$\widetilde{V}_t(\widetilde{w}_t, h_t) \approx U(g(c_t, h_t)) + \beta V_{t+1}(E_t w_{t+1}) + \frac{1}{2}\beta E_t V''_{t+1}(E_t w_{t+1}) var_t(w_{t+1}).$$

Without risk, the last term vanishes and the problem has a solution only if all returns are the same. Securities are then perfect substitutes and portfolio choice is indeterminate. More generally, for a risk-averse household with $V''_{t+1} < 0$, welfare declines with the volatility of future wealth. As a result, securities are imperfect substitutes. Moreover, utility declines with the volatility of future house values as well with the covariance of house values and labor income.

3.2.3 Housing Choice

The first stage problem takes as given the maximized objective $\widetilde{V}_t$ from the second stage. We assume that $\widetilde{V}_t$ is increasing in both its arguments and smooth as a function of $\widetilde{w}_t$; properties usually inherited from g, U and V_T. The first stage problem is then to choose optimal house quality to solve

$$V_t(w_t) = \max_{h_t \in \mathcal{H}} \widetilde{V}_t\left(w_t - p_t(h_t) - I(h_t), h_t\right). \tag{6}$$

The household thus trades off expenditure on a house against its indirect utility value. From (5), the latter comes from two sources: housing not only earns capital gains, but also enters utility as a consumption good—it delivers a nontradable dividend. Nontradability thus implies that housing and other assets can be imperfect substitutes even when there is no risk.

In the typical application, optimal house quality is increasing in wealth, other things equal. Indeed, the objective on the right-hand side of (6) is typically supermodular in (w, h), that is, the benefit from additional cash is increasing in house quality and vice versa. Intuitively, one key force is diminishing marginal utility of numeraire consumption and future wealth: if more is spent on housing then extra cash becomes more valuable. However, we also need that the utility value of house quality does not overturn this effect. This might happen, for example, if housing services are not a normal good in the aggregator g or if the distribution of capital gains R^h becomes much more attractive at higher qualities.

With a discrete set of house qualities, an increasing policy function is a step function in wealth: there are cutoff wealth levels at which households are indifferent between two adjacent quality levels. Households with wealth in between two cutoffs all choose the same quality level which they strictly prefer. Moreover, our setup allows for zero holdings of housing—in general, marginal utility need not increase without bound as consumption of housing services tends to zero. As a result, there can be a wealth cutoff at which households are indifferent between the lowest available house quality and not buying any house. With continuous house quality, we work with a smooth price function and also assume further that the objective $\widetilde{V}_t$ is smooth in h_t. At the optimal quality, a household is then indifferent between his optimal quality and a slightly better or worse house. Optimal choice is characterized by the first order condition

$$p_t'(h_t) + I'(h_t) = \frac{\widetilde{V}_{t,2}(w_t - p_t(h_t) - I(h_t), h_t)}{\widetilde{V}_{t,1}(w_t - p_t(h_t) - I(h_t), h_t)}, \tag{7}$$

where $\widetilde{V}_{t,i}$ are the partial derivatives of $\widetilde{V}_t$.

The marginal rate of substitution of housing for other expenditure is equated to the *slope* of the house price function at quality h. The slope appears because of indivisibility: the quantity of housing is one for all households, and indifference is across nearby quality levels. In contrast to a competitive model with divisible goods, the marginal rates of substitution of different households are not necessarily equated in equilibrium. The only exception is the case of a linear function for prices as well as linear improvements (for example, either one of the two special cases for technology highlighted above $I(h_t) = \delta_h h_t$ and $I(h_t) = 0$.). Indeed, if the slopes on the left-hand side are the same everywhere, then h_t can equivalently be interpreted as the quantity of a divisible housing capital.

3.2.4 Consumption and Savings

Consider now the second-stage problem for given house quality. The first-order conditions with respect to nonhousing consumption c_t as well as portfolio weights α_t on the $J-1$ risky securities can be arranged as

$$\begin{aligned} U'(g(c_t,h_t))g_1(c_t,h_t) &= \beta E_t\left[V'_{t+1}(w_{t+1})\right]R_t^f \\ 0 &= \beta E_t\left[V'_{t+1}(w_{t+1})\left(R_{t+1} - \iota R_t^f\right)\right]. \end{aligned} \tag{8}$$

The first equation says that households are indifferent at the margin between consumption and borrowing or lending at the risk-free rate. The second equation shows the portfolio choice margin: households are indifferent between risk-free investment and investment in any of the risky securities.

The first equation helps understand which households hold leveraged positions in housing. Indeed, suppose there are no risky securities. The first equation then determines optimal consumption, the only variable affecting future cash on hand w_{t+1} in (5) that is not predetermined given h_t, $\widetilde{w}_t$ and y_{t+1}. In particular, if the household has more labor income next period, he consumes more so that his bond position $\widetilde{w}_t - c_t$ declines and may become negative. We would thus expect homeowners with an upward sloping labor income profile and little initial financial wealth to leverage up. This intuition is quite general and continues to hold when labor income or security returns are risky. It allows life cycle models to successfully replicate the age profile of household portfolios in the data.

We emphasize that borrowing (that is, negative $\widetilde{w}_t - c_t$) does not imply negative savings, because savings also include the positive housing position. This feature is important for matching the data where savings are rarely negative. In the model, savings can be positive because the purpose of borrowing is not necessarily to move income from the future to the present—in fact, a borrower household with positive savings moves income from the present to the future. Instead, the purpose of borrowing for such a household is to buy a large enough house to enjoy his desired flow of housing services.

3.2.5 Securities Portfolios

To get intuition on the role of housing in portfolio choice, suppose that the continuation value function V_{t+1} is known as of date t.[g] From the first order conditions for risky

[g] This is literally true only under restrictive conditions, for example, when asset returns are iid and income is deterministic. More generally, V_{t+1} is random conditional on date t because continuation values depend on state variables that forecast future asset returns and income. The optimal portfolio weights then contain an additional term that reflects "intertemporal hedging demand"—agents prefer assets that insure them against bad realizations of the state variables. We simplify here to focus on the new effects introduced by housing.

securities and using the definition of cash on hand, we can then approximate the optimal portfolio weights on risky securities by

$$\alpha_t \approx \frac{E_t w_{t+1}}{\tilde{w}_t - c_t} var_t(R_{t+1})^{-1} \left(\rho_{t+1}^{-1} \left(E_t R_{t+1} - R_t^f \right) - cov_t \left(R_{t+1}, \frac{y_{t+1} + p_{t+1}(H_{t+1}(h_t))}{E_t w_{t+1}} \right) \right), \tag{9}$$

where $\rho_{t+1} = -E_t w_{t+1} V''_{t+1}(E_t w_{t+1}) / V'_{t+1}(E_t w_{t+1})$ reflects curvature in the value function and can be interpreted as a measure of relative risk aversion.

The optimal portfolio equation resembles textbook formulas, but makes important corrections for the presence of nontradable assets, here human capital and housing. To interpret it, consider first the scale factor $E_t w_{t+1} / \left(\tilde{w}_t - c_t \right)$. If there are no nontradable assets, then this factor equals the expected return on the entire securities portfolio and typically has only a small effect on the optimal weights. More generally, it says that the weights should be scaled up if there a lot of nontradable assets. This is because total wealth is not only $\tilde{w}_t - c_t$ but includes the present value of those nontradable assets.

Consider now the big bracket in (9). The first term reflects the desire to exploit premia on securities—expected returns that differ from the risk-free rate. To illustrate, suppose there is a security with payoffs that are orthogonal to any other shock including house prices and labor income. Up to the scale factor, the optimal weight on that security is simply its expected excess return divided by its variance as well as risk aversion. The household thus exploits a nonnegative premium on the security, and more so if there is less risk and he is less risk averse. The sign of the premium determines the direction of trade: the household holds the security if the premium is positive and shorts it otherwise.

The second term reflects hedging of labor income and housing risk. Consider first the role of labor income. If markets are complete, then there exists a portfolio of securities, θ_t^y say, that exactly replicates labor income, that is, $y_{t+1} = \pi_{t+1}^\top \theta_t^y$. Optimal portfolio choice for any risk-averse investor then involves a term that shorts the portfolio θ^y. Intuitively, the household wants to avoid risks that he is already exposed to via his nontradable human capital position. With incomplete markets, it may not be possible to short labor income. Instead, the household trades against labor income "as much as he can" with the existing set of assets. The precise meaning of "as much as he can" is given by the projection of labor income on returns $var_t(R_{t+1})^{-1} cov_t(R_{t+1}, y_{t+1})$.

Housing enters the optimal portfolio formula (9) in much the same way as labor income: it affects the demand for securities through the second "hedging demand" term. The presence of housing thus generally changes the optimal mix of securities. For example, households who work at local companies with payoffs that are correlated with their house price would optimally short the stocks of those companies. This type of interaction effect is present whether or not housing is traded in every period.

An interesting special case arises when labor income is uncorrelated with all risky securities. In this case, labor income enters (9) only because its mean increases the scale factor. The portfolio weights on all risky assets are thus scaled up along with mean labor income, regardless of labor income risk. At the same time the riskless asset position $1-\alpha_t^{\top}\iota$ is decreased. Households again trade away labor income, except that the portfolio best suited to do so now consists entirely of the risk-free security.

3.3 Asset Pricing

The previous section characterized households' optimal decision rules given prices. In particular, we have used household first-order conditions to interpret model implications for savings and portfolio choice that can be evaluated with data on household asset positions. As usual, the same first-order conditions imply restrictions on asset prices given consumption and payoffs. In fact, a large literature in asset pricing uses household Euler equations to test assumptions on preferences and market structure. Since Euler equations describe an equilibrium relationship between observables, they can be tested without taking a stand on other features of the economy such as asset supply.

This section considers household Euler equations for housing and contrasts them with those for securities. We thus move from the decisions of a generic household to restrictions on asset prices due to optimization by an entire population of possibly heterogeneous households. In order not to clutter notation, we mostly continue our practice of not explicitly labeling individual characteristics and choices such as income and consumption. At the same time, the discussion emphasizes that there is a large number of households whose first-order conditions hold simultaneously and whose choices and characteristics are observable.

3.3.1 Families of Euler Equations

So far, we have taken as given a household's subjective probability P and written subjective conditional expectations as E_t. When discussing asset prices, it is useful to distinguish between investor beliefs that relate prices and choices, and the "physical" probability that governs the data-generating process and is therefore relevant for describing measures of conditional moments constructed from the data. For example, an econometrician may measure expected excess returns $E_t^0 R_{t+1} - R_t^f$ by regressing excess returns on public information. From now on we assume that household beliefs and the physical probability agree on probability zero events next period and use the random variable ξ_{t+1} to indicate a change of measure: for any random variable Y, $E_t[Y] = E_t^0[\xi_{t+1} Y]$. Under rational expectations, we have $\xi_{t+1} = 1$.

Pricing Securities

We denote a generic household's intertemporal marginal rate of substitution (MRS) adjusted by the change of measure by

$$M_{t+1} = \beta \frac{U'(g(c_{t+1}, h_{t+1}))g_1(c_{t+1}, h_{t+1})}{U'(g(c_t, h_t))g_1(c_t, h_t)} \xi_{t+1}. \tag{10}$$

From (8), any household MRS serves as a stochastic discount factor: returns satisfy $E_t^0 M_{t+1} R_{t+1} = 1$ and securities prices can be written as $q_t = E_t^0 M_{t+1} \pi_{t+1}$. If markets are complete, all MRSs are equated in equilibrium and there is a unique M_{t+1} that represents the prices of contingent claims normalized by one-step-ahead conditional probabilities under P_0.

The standard pricing equation is often used (together with the definition of covariance) to decompose asset prices into expected discounted payoffs plus risk premia:

$$q_t = E_t^0 \pi_{t+1} / R_t^f + cov_t^0 (M_{t+1}, \pi_{t+1}). \tag{11}$$

The risk premium required by investors is larger (and the price therefore lower) if a security pays off little when the MRS is high. A positive risk premium is equivalent to a positive expected excess return $E_t^0 R_{t+1} - R_t^f$. Measures of conditional moments $E_t^0 \pi_{t+1}$ or $E_t^0 R_{t+1}$ constructed from the data—for example, by regression on public information—imply that expected payoffs are much more stable than prices, and that expected excess returns are predictable. Similar results obtain for housing, as reviewed in Section 2.2. If investors have rational expectations and have no or mild risk aversion, this finding cannot be reconciled with (11)—the excess volatility puzzle.[h]

We say that an agent is a *marginal investor* for an asset if any small change in its price or return distribution changes his optimal position in that asset. This concept is key for understanding asset pricing in heterogeneous agent models: it tells us whose behavior changes along with asset prices. For example, shocks that mostly affect *infra*marginal agents (that is, agents who are not marginal) are unlikely to move prices. Conversely, if a shock moves the price, it must also affect the positions of marginal agents. In our setup, the first-order conditions (8) imply that all households are marginal for all assets. Asset prices thus change if and only if all households adjust their positions. This is true whether or not markets are complete.

House Prices with a Finite Number of Qualities

Indivisibility means that only few households may be marginal for houses of any given quality. Indeed, with a finite set of qualities $\mathcal{H} = \{h^1, \ldots, h^n\}$, a household who strictly prefers his optimal quality in the first stage problem (6) will not respond to a change in price. At the same time, for every quality h^k except the highest, there are marginal

[h] Eq. (11) indicates two reasons why asset prices could exhibit premia that are on average high but also volatile. First, if investors have rational expectations then the covariance of the MRS with payoffs must be negative and variable. Alternatively, investors may be more pessimistic than the econometrician (that is, ξ is high when π is low) and their relative pessimism moves over time.

investors who are indifferent at date t between h^k and the next highest quality. The indifference conditions

$$\widetilde{V}_t\left(w_t - p_t\left(h^k\right) - I\left(h^k\right), h^k\right) = \widetilde{V}_t\left(w_t - p_t\left(h^{k+1}\right) - I\left(h^{k+1}\right), h^{k+1}\right) \qquad (12)$$

relate price steps between quality levels to the characteristics of the marginal investors. The marginal investors are thus particularly important for pricing houses.

Restrictions on house values $p_t\left(h^k\right)$ are obtained by adding up price steps implied by (12). In applications, there is typically an additional household optimality condition that serves as a boundary condition. In particular, we assume in what follows that there is always a household who is indifferent between the worst quality house or no house at all. For such a household, the indifference condition (12) holds at $h^1 = 0$ and $p_t\left(h^1\right) = 0$. Alternatively, the price of the worst quality house may be given by its value in some alternative use that leaves the house vacant.

Example 1 There are two periods, two states of nature and three house qualities 0, h^1 and h^2. The only security is risk free with a zero interest rate. There is no maintenance and future house values are h^i in state 1 and zero in state 2. There is a continuum of households with linear utility in both goods as well as future wealth. Households share the same discount rate of zero and the same wealth, but differ in their subjective probability of the high price state, say ρ. The household characteristic ρ is distributed uniformly on $[0,1]$. We consider an allocation with $1 - \rho_2$ houses of quality h^2 and $\rho_2 - \rho_1$ houses of quality h^1.[i]

The following prices and individual choices are consistent with individual optimization. There are cutoff households with subjective probabilities ρ_1 and ρ_2 who are indifferent between zero and h^1, as well as h^1 and h^2, respectively. Households with beliefs ρ_1 determine the value of a house of quality h^1 as "dividend" (housing services) plus expected resale value, $p\left(h^1\right) = h^1 + \rho_1 h^1$. Households with ρ_2 value house quality h^2 as $p\left(h^2\right) = p\left(h^1\right) + (1+\rho_2)\left(h^2 - h^1\right)$. Both expressions satisfy (12). Households with $\rho \in [\rho_2, 1]$ choose quality h^2, households with $\rho \in [\rho_1, \rho_2]$ choose h^1 and households with $\rho < h^1$ choose zero. In this example, the second stage is trivial: agents are indifferent between current and future consumption. In the first stage problem, higher-probability households buy high quality houses. Lower probability households perceive those houses as too expensive.

House Prices with Continuous Quality

With continuous house quality, every household is marginal for houses of his own optimal quality, but not necessarily for any other quality. To see this, start from the first-order

[i] One way to think of this setup is as an equilibrium model with fixed supply. More generally, it simply describes a family of households who buy a set of houses, for example, all movers in a given period.

condition (7) and substitute for the derivatives of $\widetilde{V}_t$ using (8) and the envelope theorem to obtain

$$p_t'(h_t) = \frac{g_2(c_t, h_t)}{g_1(c_t, h_t)} - I'(h_t) + E_t^0\left[M_{t+1} p_{t+1}'(H_{t+1}(h_t)) H_{t+1}'(h_t)\right]. \tag{13}$$

A household who chooses h_t is indifferent between h_t and a slightly better or worse house: the slope of the equilibrium price function must equal the change in the "dividend" $g_2/g_1 - I'$ plus the change in the risk adjusted future value of the house. If now some range of houses becomes more expensive while prices around quality h_t remain unchanged, this does not affect the optimal choice of h_t. No household needs to be marginal for any quality other than his own.

The Euler equation (13) restricts the slope of the price function, much like (12) restricts price steps along a discrete quality ladder. Restrictions on house values are again derived using a boundary condition. To illustrate, we select for each quality one household who buys that quality, and denote his numeraire consumption and MRS by $\left(c_t^*(h), M_{t+1}^*(h)\right)$. We then integrate (13) starting from $p_t(0) = 0$ to write the house price at quality h as

$$p_t(h) = \int_0^h \frac{g_2(c_t^*(\widetilde{h}), \widetilde{h})}{g_1(c_t^*(\widetilde{h}), \widetilde{h})} d\widetilde{h} - I(h) + E_t^0\left[\int_0^h M_{t+1}^*(\widetilde{h}) p_{t+1}'(H_{t+1}(\widetilde{h})) H_{t+1}'(\widetilde{h}) d\widetilde{h}\right]. \tag{14}$$

With indivisibility, the "dividend" of a house of quality h reflects an average of intratemporal MRSs of households who purchase qualities less of equal to h. Similarly, risk adjustment reflects an average of intertemporal MRSs of those households.

A popular special case restricts price functions to be linear. Linear pricing can be derived from the assumption that developers can freely convert houses of different quality into each other, as discussed further in Section 3.4. With the same slope $p'(h_t) = \bar{p}_t$ at every quality level, the Euler equation (13) applies to the price per unit of quality $\bar{p}_t$. The value of a house of quality h is

$$\bar{p}_t h = \frac{g_2\left(c_t^*(h), h\right)}{g_1\left(c_t^*(h), h\right)} h - I'(h) h + E_t^0\left[M_{t+1}^*(h) \bar{p}_{t+1} h H_{t+1}'(h)\right]. \tag{15}$$

With linear pricing, markets for different quality housing are tied more tightly together. As a result, every household is marginal for every house, as is the case for securities. The dividend and risk adjusted payoff at quality h can then be related to the MRSs of households who buy quality h.

Example 2 To illustrate nonlinear pricing with a continuum of qualities, we adapt the simple example from above. The set of households is unchanged, but the set of qualities is now $\mathcal{H} = [0, 1 - \rho_1]$. We consider an allocation with θ_1 houses of quality zero and $1 - \theta_1$

houses of positive quality uniformly distributed along the interval. Without maintenance, the first order condition (13) simplifies to $p_t'(h) = 1 + \rho$. We again construct prices and choices that satisfy all optimality conditions. There is a cutoff household who has subjective probability ρ_1 and is indifferent between no house and an infinitesimal house. Households with $\rho < \rho_1$ buy no house, while households with $\rho > \rho_1$ buy a house of quality $h = \rho - \rho_1$. House values are $p_t(h) = h(1 + \rho_1) + \frac{1}{2}h^2$.

Again higher probability households buy better houses, which lower probability households perceive to be overpriced. Let ρ_0 denote the true probability of state 1. The stochastic discount factors for securities $M_{t+1}^*(h)$ are equal to $(h + \rho_1)/\rho_0$ in state 1 and $(1 - (h + \rho_1))/(1 - \rho_0)$ in state 2. Since the discount rate is zero and utility is linear, they differ only by the change of measure from the subjective probability of households who buy quality h to the true probability ρ_0.[j] Every stochastic discount factor correctly prices the riskless bond, the only available security.

So far the example emphasizes heterogeneity in risk assessment through beliefs. The essential feature, however, is only that agents disagree about the future value of houses. We can thus alternatively assume that there is no risk ($\rho_0 = 1$) but ρ represents households' discount factors. For the above choices to remain optimal, we also assume that there is no risk-free security so houses are the only traded assets. Prices are then the same as above: the interpretation is that more patient households buy larger houses since they want to save more. The absence of a risk-free security is important to ensure a solution to households' problems without borrowing constraints.

3.3.2 Limits to Arbitrage

In general, there need not exist a stochastic discount factor that prices all houses. This is a key difference between houses and divisible securities with tradable payoffs. The existence of a stochastic discount factor says that all investors who choose to buy assets discount risk-free payoffs at the same rate and pay the same risk premia per unit of payoff. In a frictionless market, these properties are guaranteed by the absence of arbitrage opportunities, which in turn is necessary for the existence to a solution to the investor's optimization problem.[k]

A stochastic discount factor need not exist because indivisibility and nontradability introduce limits to arbitrage. In fact, each friction separately is sufficient to preclude discount rates or risk premia to be equated across houses. If either the quantity of assets is restricted to zero or one, or if all dividends have to be consumed, then fewer arbitrage

[j] The example relies on differences in beliefs for tractability. For the issues discussed below, it does not matter whether differences in risk attitude stem from beliefs or other household characteristics such as risk aversion or nontradable income risk.

[k] If an investor perceives two assets with same exposure but different risk premia, he expects unlimited profits from shorting the expensive portfolio and buying the cheaper one.

trades are feasible or desirable and the absence of arbitrage places weaker restrictions on prices. We consider the mechanisms in turn and then draw conclusions for matching observed prices.

Indivisibility and the Valuation of Quality Steps

Example 2 above illustrates the role of indivisibility. Suppose there was a stochastic discount factor M_{t+1} pricing all houses. With two states of nature, M_{t+1} consists of two numbers. Since houses pay off zero in state 2, the risk-adjusted future payoff from a house of quality h would have to equal h multiplied by the value of M_{t+1} in state 1. However, in Example 2 the risk adjusted payoff is $\rho_1 h + \frac{1}{2}h^2$, a contradiction. The result does not depend on a continuum of house qualities—a similar contradiction can be shown in Example 1. Moreover, it does not depend on nontradability; in fact, in the examples the housing dividend per unit of quality $g_2/g_1 - I'$ is independent of h, as it would be if dividends could be sold at a per-unit price in a rental market.

Why do optimizing households not arbitrage away differences in the valuation of house payoffs? Consider the pricing equation (13): it resembles a standard pricing equation $q_t = E_t^0 M_{t+1}\pi_{t+1}$, except that it is applied only to the quality step from h to $h + dh$.[l] The pricing of that quality step reflects the valuation of buyers of h. Buyers of lower quality houses may not share the same valuation—in fact, in the examples they perceive a lower probability of a positive payoff and would like to short the quality step at h. However, quality steps are not by themselves traded in markets: households can only trade houses, that is, portfolios of quality steps. Moreover, households cannot sell houses short. As a result, they cannot in general generate a synthetic claim that replicates the change in payoff at a quality step.[m]

If other forces equate risk-adjustment factors, a stochastic discount factor exists even with indivisibility and short sale constraints. For example, suppose that housing risk is spanned by the securities, that is, for every house there exists a portfolio of securities with the same payoff profile. Every $M_{t+1}^*(h)$ is then a valid stochastic discount factor for all houses.[n] If optimizing investors can replicate houses by trading securities, they equate

[l] Another difference is that by our timing convention house prices are always "cum dividend"—they include the current flow payoff from housing—whereas securities prices are ex dividend. This convention is not central to the discussion that follows.

[m] The effect of indivisibility is different from that of short sale constraints with divisible securities. Indeed, in models with only short sale constraints and no other constraints a stochastic discount factor does exist: for any given risk, it reflects the MRS of the investors who are most optimistic about that risk and end up as the only investors exposed to that risk in equilibrium. As a result, investors do not differ in the risk premia they pay for risks they are actually exposed to.

[n] If housing risk is spanned, the marginal rates of substitution $M_{t+1}^*(h)$ are equated on all events over which house payoffs are constant and can be pulled out of the integral in (14). Integrating over $\tilde{h}$, we obtain a standard risk-adjusted payoff.

their assessment of housing risk. For example, in the special case when markets are complete, all $M_{t+1}^*(h)$ are equal.

Nontradability and Individual Specific Returns

We refer to $g_2/g_1 - I'$ as the dividend from housing because it records the flow benefit to homeowners, as does the dividend on a security such as equity. However, nontradability implies that the housing dividend may differ across households who consume h because those households have different consumption bundles and preferences. As a result, the returns earned on the same housing position may differ across households, in contrast to the return on securities.[o] Returns on owner-occupied housing are thus more difficult to observe than those on other assets, including rental housing where the dividend to the landlord can be observed in the rental market.

Nontradability implies that a stochastic discount factor need not exist even when the pricing of houses is linear. Indeed, (15) says that with linear pricing the MRS of buyers of quality h determines the price of assets and houses of quality h. However, it is not necessarily true that the same MRS determines house prices for any other quality $h' \neq h$. Arbitrage is limited because households who disagree about the required risk-free return on assets or on the risk premium on houses may also obtain different marginal benefits from housing services, or "marginal dividends." More patient or more optimistic households thus buy larger houses, while more impatient or more pessimistic households buy smaller houses.

3.3.3 Pricing Houses vs Pricing Equity

What are the testable restrictions on the prices of houses and other assets that are implied by optimizing behavior in our framework? The large literature on the pricing of equity employs two working hypotheses. The first is that equity, or firm capital, is a divisible asset that is priced linearly, so that it suffices to focus on the properties of a single per-unit price. Second, there exists a stochastic discount factor that can be inferred from optimality conditions of some investor, for example, certain households or institutional investors. Success of a model is then measured by whether the family of stochastic discount factors implied by the model can explain how the price of equity moves relative to dividends. Moreover, one can learn about desirable features of a model up front from a reduced form approach that postulates a specific functional form for the stochastic discount factor and infers its properties from securities prices.

The previous discussion shows that models of owner-occupied housing satisfy these two working hypotheses only under restrictive assumptions. On the one hand, indivisibility implies that pricing may be nonlinear for any given observable concept of

[o] In fact, when we select households with $\left(c_t^*(h), M_t^*(h)\right)$ such that (13) holds, it is not necessarily the case that all households who buy h share those characteristics.

quality—houses of different qualities are different assets. The challenge for a model is then not to reconcile movements in one price with many household MRSs, but rather to generate the right cross-sectional links between different prices and MRSs. On the other hand, when markets are sufficiently incomplete, limits to arbitrage preclude the existence of a stochastic discount factor altogether. In this case, reduced form frictionless pricing exercises do not help infer how pricing works—a more explicit analysis of frictions is called for.

Whether or not pricing is linear or a stochastic discount factor exists, models of optimizing households imply strong testable restrictions on the joint distribution of house prices, house quality choices and household characteristics. Suppose, for example, that according to the model, wealth is the only dimension of heterogeneity among households. Optimal choice of housing implies an assignment of house qualities to wealth levels. Given that assignment, (14) predicts a cross section of prices by quality. Success of a study then depends on how well it can match the cross-sectional comovement of wealth, quality and prices when compared to micro data. The restrictions are derived from household optimization alone, much like standard Euler equation tests.

Nonlinear Pricing and the Cross Section of House Prices

With indivisibility and nontradability, the cross section of house prices is especially informative about the merits of different models. In particular, nonlinear pricing can account for richer patterns in the cross section of capital gains than linear pricing. We have seen in Section 2.2 how capital gains systematically differ across the quality spectrum over the recent US boom–bust cycle. With linear pricing, capital gains are

$$\frac{p_{t+1}(H_{t+1}(h_t))}{p_t(h_t)}=\frac{\bar{p}_{t+1}}{\bar{p}_t}\frac{H_{t+1}(h_t)}{h_t}.$$

The conditional distribution of capital gains depends on current quality h_t only via actual changes in quality between t and $t+1$. In contrast, the effects of valuation are the same for all qualities. This feature implies that models with linear pricing have trouble generating the large differences in average capital gains across quality tiers. If pricing is instead nonlinear, then changes in the characteristics of marginal investors along the quality spectrum can also affect capital gains.

Nonlinear pricing of houses can reflect various dimensions of heterogeneity. Example 2 highlights how differences in risk assessment or discount factors affect *inter*temporal MRSs. However, even if all intertemporal MRSs agree, so that a stochastic discount factor exist, the *intra*temporal MRSs (7) are not necessarily equated because of nontradability. Nonlinear payoffs and hence prices can thus obtain even in a static setting or if all intertemporal MRSs agree so that a stochastic discount factor exists.

With nonlinear pricing, individual characteristics of marginal investors at a given quality matter for the relative price of that quality. The same property arises in markets that are segmented by quality. The difference between nonlinear pricing and segmentation is that nonlinear pricing creates spillovers in pricing across qualities. For example, changes in the preferences of households who buy low quality houses affect also the values of higher quality houses.

Volatility of House Values in Heterogeneous Agent Models

Indivisibility—and to some extent also nontradability—provide extra scope for heterogeneity of agents to affect the volatility of house prices. This is promising because standard heterogeneous agent models face a challenge when it comes to generating volatility. The challenge arises because optimizing households respond to shocks by reallocating assets until all Euler equations hold jointly. If a stochastic discount factor exists, discount rates $E_t^0 M_{t+1}$ and risk premia $cov_t^0(M_{t+1}, \pi_{t+1})$ are equated across agents. Any shocks to the distribution of agent characteristics or shocks that affect a subset of agents have only a muted impact on prices because portfolio adjustments keep MRSs similar. As the simplest example, if markets are complete, pure changes in the distribution of individual income risks are offset by portfolio adjustment and prices remain unchanged.

With indivisible housing and markets sufficiently incomplete so that housing risk is not spanned, intertemporal MRSs $M_{t+1}^*(h)$ are not equated. Suppose there is a shock that affects the income or beliefs of low quality home buyers. The shock can change the MRS of low quality buyers and hence the slope of the price function at low qualities, but have no effect on the MRS of high quality buyers. Reallocation of housing risk is limited since no household buys more than one house. As a result, the shock will likely have a stronger impact on house prices in the low quality segment than the high segment, and the aggregate market will move together with the price of low quality houses.

To illustrate the implication of cross-sectional shocks on risk premia in the standard pricing equation (14), we let $D_t(h)$ denote the housing dividend and rewrite the pricing equation as

$$p_t(h) = D_t(h) + E_t^0[p_{t+1}(H_{t+1}(h))]/R_t^f + h\, cov_t^{0,U}(M_{t+1}^*(h)), p'_{t+1}(H_{t+1}(h))H'_{t+1}(h)),$$

where we have exchanged expectation and integration, and used the fact that all $M_{t+1}^*(h)$ agree on the risk-free rate. The second term on the right-hand side is the expected present value of the house discounted at the risk-free rate. The notation $cov_t^{0,U}$ indicates that the random variables vary not only across states of nature, but also across qualities, where quality is uniformly distributed on $[0, h]$ by construction—this is because we have selected one household for quality level.

For securities, MRSs are all equal and risk premia depend on variation common to all MRSs and payoffs across states of nature. Any excess volatility of prices is due to changes in this common variation. With indivisible housing, excess volatility can also be due to changes of the cross-sectional distribution of agent characteristics. In particular, changes in the environment that affect only a subset of agents that buy low quality houses can shift the distribution and affect many prices and hence the aggregate market.

When pricing is linear, nontradability implies that MRSs are still not equated across investors. However, the same per-unit price $\bar{p}_t$ appears in all Euler equations (15). Hence, the per-unit price will only change if the Euler equations of buyers at all quality levels are affected. A shock that affects only a subset of households can thus only matter for prices if it also changes the Euler equation of high quality buyers. This requires changes in either the intertemporal or the intratemporal MRS of high quality buyers. Models with linear pricing thus imply that the distribution of house choices respond more strongly, which dampens the effect on prices. Overall, the scope for price volatility is reduced.

3.4 Equilibrium

In this section we take a first look at equilibrium. We close the frictionless model presented so far by introducing an exogenous supply of securities as well as an exogenous endowment of numeraire consumption. We also assume a fixed *aggregate* supply of housing services. To emphasize the role of indivisibility, we compare two stark special cases for technology that are common in applications: a fixed distribution of house qualities, and free conversion of house qualities into each other.

We take a general approach to expectation formation that can accommodate various concepts in the literature. We first define a *temporary equilibrium* for date t as a collection of prices and allocations such that markets clear given beliefs and agents' preferences and endowments. Following Grandmont (1977), temporary equilibrium imposes market clearing and individual optimization, but does not require that each agent's belief coincide with the physical probability P^0. We then discuss further restrictions on expectations and their role in quantitative work. In particular, we compare rational expectations equilibrium and self-confirming equilibrium—a common shortcut that simplifies computations in heterogeneous agent models—as well as temporary equilibria with directly measured expectations.

3.4.1 Housing Market Clearing

We denote the mass of households that makes decisions at date t by $\mathcal{I}_t$. For each individual $i \in \mathcal{I}_t$, the solution of the individual household problem delivers decision rules for consumption, savings and portfolio choice that depend on calendar time, the endogenous state variable cash on hand and current prices. Moreover, household decisions depend on preferences and in particular beliefs about future income, prices and asset payoffs. Let P_t^i be the belief of household i at date t and $h_t^i\left(p_t, q_t; w_t^i, P_t^i\right)$ be his housing demand at date t.[p]

[p] Here, P_t^i represents a probability on infinite sequences. Beliefs at different information sets therefore do not have to be derived as conditionals from a single probability. This generality is useful to accommodate, for example, beliefs that are derived from a forecasting model estimated with data up to date t.

We assume that there are always at least as many households as houses of positive quality.[q] We thus fix the mass of houses at $\mathcal{I}_t$ and let G_t denote the date t cumulative density function of available house qualities, defined on $[0, \infty)$. If the households' choice set $\mathcal{H}$ is finite, then G_t is a step function. If $G_t(0) > 0$, then not every household will be able to buy a house of positive quality in equilibrium. The housing market clears if at every quality $h > 0$, the number of households who choose a house of quality h or better is the same as the number of houses with these qualities:

$$\Pr\left(h_t^i\left(p_t, q_t;\ w_t^i, P_t^i\right) \geq h\right) = 1 - G_t(h). \tag{16}$$

3.4.2 Fixed Supply vs Linear Conversion

If the distribution of house qualities G_t is exogenous, prices adjust so that the household sector absorbs the given distribution. The endogenous objects in this case include equilibrium house prices—one for every quality level—as well as the assignment of individual houses to individual households. In general, house prices are nonlinear in quality and reflect the distributions of qualities and household characteristics. The simple examples of Section 3.3.2 show how this can happen: if we take the set of houses used there to describe an exogenous supply, then the prices and individual choices characterize an equilibrium given that supply.

The polar opposite of a fixed quality distribution is *linear conversion*. Suppose that the total housing stock (measured as the aggregate supply of housing services) is fixed at some number H_t, but that it can be divided up every period into individual houses without cost. Since the marginal transformation across quality types is now fixed at one, pricing must be linear. There is only one price that reflects the value of a unit of housing in terms of numeraire. The distribution of qualities G_t becomes an endogenous object that is determined in equilibrium subject to a constraint on the mean

$$\int_0^1 h dG_t(h) = H_t. \tag{17}$$

A fixed distribution is interesting in applications that consider the short-term response to shocks. It is also useful for longer-term analysis if the market can be viewed as a collection of segments fixed by geography or regulation such as zoning. In contrast, linear conversion is an interesting assumption in applications that consider long-run outcomes or when studying new developments where developers design the distribution of houses from scratch. Beyond these polar opposites, it could be interesting to explore intermediate cases of costly conversion by developers. The macroeconomics literature has yet to consider this explicitly.

[q] This assumption covers most applications we discuss below. Alternatively, we would have to develop further the use of a vacant house.

To decentralize an economy with linear conversion, we assume that there is a competitive developer sector that buys existing houses and sells new houses. The endogenous distribution of houses will then satisfy our earlier assumption that the number of houses is always less or equal than the number of households. Since households have no use for more than one house, developers never create more than $\mathcal{I}_t$ houses at date t. Moreover, competition among developers and linear conversion force linear pricing: the relative price of any two qualities must equal the unitary marginal rate of transformation.

With either technology, the housing component of equilibrium includes a price function $p_t(.)$ as well as an allocation of house qualities such that the market clearing condition (16) holds. In an *equilibrium with fixed quality distribution*, (16) holds for the exogenous cdf G_t. In contrast, an *equilibrium with linear conversion*includes an equilibrium distribution of house qualities G_t that satisfies (17) and moreover features a linear price function $p_t(h) = \bar{p}_t h$.

3.4.3 Temporary Equilibrium

We assume that household $i \in \mathcal{I}_t$ enters period t endowed with a house of quality $\bar{h}_t^i$, securities $\bar{\theta}_t^i$ as well as y_t^i units of numeraire. We allow for households in their last period of life who mechanically sell any housing and securities and consume all the proceeds. To accommodate long-lived securities, we write payoffs as price plus dividend, that is, $\pi_t = \hat{\pi}(q_t) + D_t$. For example, the Jth security is a risk-free one-period bond, so $\pi_{t,J} = 1$. If the jth security is equity then $\pi_{t,j} = q_{t,j} + D_{t,j}$ where $D_{t,j}$ is the dividend.[r]

In addition to a price function, a house allocation and—with linear conversion—a distribution of house qualities, a *date ttemporary equilibrium* consists of securities and consumption allocations as well as security prices such that housing, numeraire and securities markets clear at the optimal demand, with initial wealth evaluated at the equilibrium prices. The conditions for wealth, numeraire and securities are

$$
\begin{aligned}
w_t^i = y_t^i + \bar{\theta}_t^{i\top} \hat{\pi}_t(q_t) + p_t\left(\bar{h}_t^i\right); \quad i \in \mathcal{I}_t \\
\int_{\mathcal{I}_t} c_t^i\left(p_t, q_t; w_t^i, P_t^i\right) + I\left(h_t^i\left(p_t, q_t; w_t^i, P_t^i\right)\right) di = \int_{\mathcal{I}_t} \left(y_t^i + \theta_t^{i\top} D_t\right) di \\
\int_{\mathcal{I}_t} \theta_t^i\left(p_t, q_t; w_t^i, P_t^i\right) di = \int_{\mathcal{I}_t} \bar{\theta}_t^i di.
\end{aligned}
\tag{18}
$$

A *sequence of temporary equilibria* is a collection of date t temporary equilibria that are connected via the updating of endowments. In particular, for any household $i \in \mathcal{I}_t$ who was already alive at date $t-1$, we impose $\bar{h}_t^i = h_{t-1}^i\left(p_{t-1}, q_{t-1}; w_{t-1}^i, P_{t-1}^i\right)$ and similar for the securities holdings. Agents who enter the economy at date t are endowed only with labor

[r] The function $\hat{\pi}$ helps accommodate debt with longer but finite maturity. For example, if the kth security is a risk-free two-period zero-coupon bond, then $\pi_{t,k} = \hat{\pi}_{t,k} = q_{t,J}$ since the two-period bond turns into a one-period bond after one period.

income y_t^i. While a sequence of temporary equilibria tracks the distribution of asset holdings over time, it still does not restrict expectations.

3.4.4 Rational Expectations Equilibrium vs Self-confirming Equilibrium

A *rational expectations equilibrium* is a sequence of temporary equilibria such that $P_t^i = P^0$ for every period t and agent i. Beliefs thus coincide with the physical probability for all events: all agents agree with the econometrician on the distribution of all exogenous and endogenous variables. Rational expectations equilibrium is common in macroeconomic studies, especially when the model has few agents and assets or when there is no aggregate risk. In such cases, it is straightforward to move from the recursive formulation of decision problems to the definition of a recursive equilibrium that expresses prices as a function of a small set of state variables.

For the simplest example, suppose there is a representative agent. Since we have assumed a fixed supply of assets, there are no endogenous state variables. Prices only depend on current variables such as consumption as well as current variables required to forecast future exogenous variables such as income and asset payoffs. With rich heterogeneity, rational expectations equilibria become more difficult to characterize. With incomplete markets as well as other frictions described below, defining a recursive equilibrium may require a large dimensional state space that contains the distribution of not only wealth but also individual asset holdings—for example, housing and long term mortgages—as well as their dependence on age.

To avoid explicitly dealing with a large state space and the resulting complicated distribution of endogenous variables, studies with heterogenous agents and aggregate risk often look for a *self-confirming equilibrium* in which agent beliefs coincide with the physical probability P^0 only on a subset of events.[s] A common approach follows Krusell and Smith (1998) and parametrizes agent beliefs about future prices with "forecast functions" that map future prices to a simple set of current predictor variables (such as the current cross-sectional mean of asset holdings) and shocks. A self-confirming equilibrium requires that the forecast functions match prices also under the physical probability.

Self-confirming equilibrium imposes different restrictions on allocations and prices than rational expectations equilibrium since the forecast functions only involve a limited set of moments of the state variables. In general, there can be other self-confirming equilibria with other forecast functions, and there is no guarantee that any particular self-confirming equilibrium is a rational expectations equilibrium.[t] Applying self-confirming

[s] The labeling here follows Sargent (1999) who in turns builds on the game theoretic concept in Fudenberg and Levine (1998). Krusell and Smith (1998) refer to "approximate equilibria."

[t] At the same time, if there exists a recursive rational expectations equilibrium, then it is also a self-confirming equilibrium for *some* forecast function (not necessarily simple). In sufficiently tractable models, one can try out different forecast functions systematically so as to establish that a self-confirming equilibrium is indeed close to a rational expectations equilibrium. This route is taken by Krusell and Smith (1998), but not in the typical application on housing reviewed below.

equilibrium with a given forecast function thus calls for justifications of assumptions on beliefs, perhaps by appealing to bounded rationality.

3.4.5 Temporary Equilibrium with Measured Expectations

An alternative approach implements temporary equilibria by directly measuring expectations about future variables that are relevant for agent decisions. The temporary equilibrium then provides a map from technology and the distribution of household characteristics *as well as expectations* into prices and allocations. To specify beliefs, one relevant source is survey data which can be informative in particular about the cross-sectional relationship between expectations and other characteristics (for example, Piazzesi and Schneider, 2009a). Alternatively, expectations about prices can be specified using a forecasting model (Landvoigt et al., 2015).

Temporary equilibrium with measured expectations also simplifies computation. It is helpful to think of the computation of equilibrium in two steps—first individual optimization given prices and then finding market clearing prices. To find temporary equilibrium prices for a given trading period means finding a solution to the nonlinear equation system (18) in as many unknowns as there are prices. This is in contrast to rational expectations equilibrium where one looks for an entire price function. Since the price finding step for temporary equilibrium is easier, the optimization step can be made more difficult: the concept lends itself well to models with a rich asset structure, for example, with many house types or many risky assets.

A conceptual difference between temporary equilibrium with measured expectations and rational expectations equilibrium is that the modeler does not a priori impose a connection between expectations at any given date and model outcomes at future dates. Of course, if the model is well specified, then this does not matter for the fit of the model: any rational expectations equilibrium gives rise to a sequence of temporary equilibria given the set of beliefs that agents hold in the rational expectation equilibrium. With a well specified model, that same set of beliefs should be apparent in expectation surveys or in a good forecasting model.

The conceptual difference is thus in how we assess the fit of a misspecified model and how we achieve identification of parameters. Rational expectations equilibrium and self-confirming equilibrium view both prices and the cross section of endowments as a function of state variables. To identify parameters that affect the coefficients in prices and decision rules requires controlled variation of the state variables. The concepts are thus most easily and most commonly applied when variables display recurrent patterns: the empirical moments of prices and other variables can then be compared to the stationary equilibrium implied by the model. In contrast, temporary equilibrium can be implemented even with data on only a single trading period. Prices are then a single set of numbers and endowments are measured directly. Identification of parameters that affect prices comes from cross-sectional variation in prices and allocations.

There is also a difference in how we deal with misspecification and counterfactuals. Rational expectations insist that expectations are "consistent with the model," so beliefs are as misspecified as the model itself. Moreover, counterfactuals—such as changes in a policy parameter—vary expectations in a way that is consistent with the model. Temporary equilibrium with measured expectations instead emphasizes that expectations are "consistent with the data" at the initial equilibrium. As a result, there is no prediction on how expectations change with parameters; any counterfactual requires a reassessment of the assumptions on expectations.[u]

There are two reasons why the use of temporary equilibrium with measured expectations is particularly attractive in models of housing. First, as we have discussed, there are payoffs from including a rich set of assets, in particular houses of many different qualities. Second, the postwar data on housing is shaped by the two boom periods—the 1970s and the 2000s—that saw several unusual shocks, as discussed in Section 4.5. Given this data situation, identification of a stationary equilibrium price function from regular patterns is less powerful. In contrast, there is much to learn from the cross section and from data on expectations for both boom episodes.

3.5 Production and Land

In this section, we describe models of housing supply that are common in the applications below. We start from a general setup that allows for land and structures as separate factors of production. We then explain when housing can nevertheless be represented by the single state variable "quality," as we have done throughout this chapter. Finally, we review additional restrictions on house prices derived from firm optimization.

Consider a general production function at the property level. When a new structure of size k^0 is paired with a lot of size l, initial house quality is $h = F^0\left(k^0, l\right)$. Once a house has been built, its lot size remains the same, whereas the structure may depreciate or improve. With a stream of investments i_t, the quality of a house of age τ is given by

$$h_{t+\tau} = z_{t+\tau} F^{\tau}\left(k_t^0, i_{t+1}, \ldots, i_{t+\tau}, l_t\right), \tag{19}$$

where $z_{t+\tau}$ is a productivity shock. The production function F^{τ} may depend on the vintage τ.

Both new structures and improvements to existing houses are produced by a construction sector from capital K_t^c and labor N_t^c. As before, the mass of houses is $\mathcal{I}_t$ and we index individual houses by $j \in [0, \mathcal{I}_t]$. We further assume that it is costless to scrap an existing house. Construction output—or residential investment—is then

[u] Predictions on expectations can be obtained by imposing more structure on expectation formation, for example, via learning rules. For a survey on learning in macroeconomics, see Evans and Honkapohja (2009).

$$\int_0^{\mathcal{I}_t} \left(k_t^0(j) + i_t(j)\right) dj = I_t^c = Z_t^c F^c\left(K_t^c, N_t^c\right), \tag{20}$$

where F^c and Z_t^c are the production function and the productivity shock for the construction sector, respectively.

We distinguish the construction sector labeled c from the rest of the business sector–labeled y—that makes numeraire from capital and labor. Capital in both sectors is made from numeraire one for one without adjustment costs and depreciates at constant rates δ^c and δ^y. The resource constraints for numeraire and the capital accumulation equations are

$$\begin{aligned} C_t + I_t^y + I_t^c &= Z_t^y F^y\left(K_t^y, N_t^y\right), \\ K_{t+1}^s &= (1-\delta^s)K_t^s + I_t^s, s = y, c. \end{aligned} \tag{21}$$

It remains to describe how costly it is to change the distribution of existing individual housing units. We distinguish different scenarios below.

In each case, the definition of equilibrium is amended by adding (*i*) construction output as a separate intermediate good that trades in a competitive market at the relative price p_t^c, (*ii*) both types of capital as securities in the households' problems that trade at a price of one and yield a net return equal to the marginal product of capital less depreciation, (*iii*) as market clearing conditions for construction output and numeraire (20) and the first equation in (21), respectively, (*iv*) labor income as labor times the competitive wage in the household budget constraint.

3.5.1 From Land and Structures to House Quality

In principle, the above technology could give rise to rich dynamics for the distribution of house types. For example, if different vintages of houses have different capital–land ratios, they may yield the same housing services, but depreciate at different rates. The macroeconomics literature has by and large sidestepped this issue with assumptions that allow housing to be summarized by one number, quality. We now discuss several special assumptions that accomplish the same outcome even when land is present. The simplest approach is to leave out land altogether, as in the literature on home production. Housing is then identified with structures only.

3.5.2 The Tree Model

Another simple approach is a "tree model" of housing that can motivate setups with a fixed or slow-moving quality distribution. Suppose that structures depreciate at rate δ, but that a house remains inhabitable (that is, yields positive housing services) only as long as structures and land are always paired in exact proportions.[v] All owners who hold

[v] In terms of the above notation, assume first that $F^0(k, l) = l$ if $k = \kappa l$ and $F^0(k, l) = 0$ otherwise, so every inhabitable house built must have a structure–land ratio of κ. Assume further that future quality F^τ is equal to l_t if $i_s = \delta k_s^0$ for all $s = t, \ldots, t + \tau$ and zero otherwise.

a house from one period to the next then make the improvement $i_t = \delta k_t^0$ every period. In other words, a house works like a tree that yields fruit equal to housing services less improvements.

The tree model implies that the state of a house can be summarized by a single variable, quality. From the perspective of households, quality is constant as long as maintenance is performed, the case of "essential maintenance" discussed in Section 3. When the distribution of lots is fixed, one can apply the definition of equilibrium with a fixed quality distribution from Section 3.4. Alternatively, we could add a technology by which lots are converted. For example, if it was possible for developers to freely redivide lots, then we would obtain an equilibrium with linear conversion.

3.5.3 A Frictionless Model

Suppose that the production of housing from land and structures has constant returns and that structures depreciate at a constant rate. Suppose further that houses are produced by a competitive developer sector who can linearly convert both land and structures. We thus have a frictionless model with two factors of production.[w] All houses built at the same point in time will share the same ratio of structures to land. From the perspective of households, the change in house quality depends on the land share together with the depreciation rate of structures. The household problem thus looks like one with geometric depreciation of quality, determined endogenously from the equilibrium land share.

The frictionless model imposes a supply-side restriction on house prices that must hold together with Euler equations from the household side discussed earlier. Indeed, from the first-order condition of a developer, we have

$$\bar{p}_t F_1^0(K_t, L) = p_t^c,$$

where K_t is aggregate structures, L is aggregate land, assumed constant, and p_t^c is the relative price of construction output. If there are many structures, then the scarcity of the fixed factor land drives up the per-unit price $\bar{p}_t$ of housing. Since aggregate structures move slowly over time, this type of model typically has trouble generating a lot of volatility in house prices relative to the price of construction output. The problem is similar to that encountered by models of the firm without adjustment costs to capital.

[w] In terms of the above notation, let

$$F^\tau(k_t^0, i_{t+1}, \ldots, i_{t+\tau}, l_t) = F^0(k_{t+\tau}, l_t),$$

where $k_{t+\tau} = (1-\delta)k_{t+\tau-1}$ is recursively defined.

3.5.4 Land as a Flow Constraint

An alternative frictionless model uses land as a constraint on the flow of new housing, as opposed to as a factor of production for all housing as above. Since the model assumes linear conversion, we write technology directly in terms of aggregate quality:

$$H_t = (1-\delta)H_{t-1} + \widetilde{F}^h\left(Z_t^c F^c\left(K_t^c, N_t^c\right), \bar{L}\right). \tag{22}$$

Here, $\widetilde{F}^h$ is a constant returns production function that transforms construction output (that is, housing investment) and a constant flow of new land into new housing. The technology is decentralized via competitive firms.

The flow constraint approach also reduces the state variables to only house quality. It does so by applying the depreciation rate directly to the bundle of land and structures. Even though different vintages of new houses will generally have different land shares, they are nevertheless assumed to depreciate at the same rate. The flow constraint also differs from the frictionless model above in the restriction on prices. Firm first-order conditions deliver

$$\bar{p}_t \widetilde{F}_1^h\left(Z_t^c F^c\left(K_t^c, N_t^c\right), \bar{L}\right) = p_t^c.$$

The ratio of house prices to the price of construction output now relates to residential investment, which is much more volatile than the level of capital.

3.6 Rental Housing

So far we have focused on owner-occupied housing, that is, we have forced households to own a house if they want to consume housing services. We now modify the model to allow for rental housing. We discuss implications for portfolio choice and discuss how additional restrictions on house prices can be derived from household as well as from landlord decisions to invest in tenant-occupied housing.

We continue to assume that households have exactly one residence that is now either owned or rented. We denote the quality of a rented residence by s_t and the rental rate at that quality by $p_t^s(s_t)$. We then modify the second-stage problem to

$$\begin{aligned} \widetilde{V}_t\left(\widetilde{w}_t, h_t\right) &= \max_{c_t, \alpha_t} U(g(c_t, h_t + s_t I_{h_t=0})) + \beta E_t[V_{t+1}(w_{t+1})] \\ w_{t+1} &= \left(\widetilde{w}_t - c_t - p_t^s(s_t)\right) R_{t+1}(\alpha_t) + p_{t+1}(H_{t+1}(h_t)) + y_{t+1}. \end{aligned} \tag{23}$$

In the budget constraint, expenditure now includes rent. The indicator function in the objective ensures that only households who have not chosen to own (that is, $h_t = 0$) obtain utility from a rented residence.

To handle the landlord side of renting, we assume that tenant-occupied houses of a given quality are held in real estate investment trusts (REITs) and households can purchase shares in those trusts subject to short-sale constraints. REIT shares then enter the second stage problem much like standard securities. The dividend earned by the REIT from a house of quality h_t is given by the rent net of maintenance cost $p_t^s(h_t) - I_r(h_t)$. We allow maintenance cost to be higher when the house is tenant occupied than when it is owner occupied.

The formulation here thus introduces one advantage of ownership—lower maintenance cost—that is traded off against the disadvantage of bearing housing price risk. This approach to studying rental markets and tenure choice in an otherwise frictionless equilibrium model goes back to Henderson and Ioannides (1983). Their paper also provides microfoundations for the difference in maintenance cost using a moral hazard problem between landlord and tenant. A closely related approach assumes that homeowners receive more housing services from owned houses. In addition to the tradeoff studied here, differences in tax treatment as well as the interaction of tenure choice with collateral constraints and transaction costs are also important; they are discussed further below.

3.6.1 Optimality Conditions and Tenure Choice

Renters' first-order condition is one of intratemporal choice between the two goods, housing services and numeraire. We focus on the case of a continuum of qualities. With a smooth rent function, a household who rents a house of quality h must be indifferent between renting that house or renting a slightly better or worse house:

$$p_t^{s\prime}(h_t) = \frac{g_2(c_t, h_t)}{g_1(c_t, h_t)}. \tag{24}$$

Much like for owner occupiers, a renter of quality h_t is marginal for houses of quality h_t, but not necessarily for house of any other quality. As a result, the rent function can in general be nonlinear—a linear rent function obtains under special assumptions such as when rental houses of different qualities can be converted one for one.

The first-order conditions for REIT shares at different quality levels work like those for stocks of different companies. The intertemporal MRS of a landlord household serves as a stochastic discount factor for tenant-occupied houses. Without frictions, the typical landlord household will build a diversified portfolio that contains houses of all qualities. For tenant-occupied houses, discount rates and risk-adjustment factors are thus also equated across quality levels. This does not mean, however, that prices become linear in quality: rent and hence the dividend to the landlord is generally nonlinear due to indivisibility in the rental market.

The presence of a rental market separates the roles of housing as a consumption good and asset. While owners must commit more savings toward the housing asset and bear housing risk, renters simply pay the flow expenditure of housing services. At the same time, the difference in maintenance cost implies that the rent for a house of given quality may be higher than the dividend that a household would earn if he instead were to own the house. In the current setup with indivisibility as the only friction, we would thus expect households who perceive a higher risk-adjusted payoff from housing to become owners.

3.6.2 The User Cost of Housing

Consider a household who is indifferent between owning and renting a house of quality h. Suppose further that housing risk is spanned so that the stochastic discount factor is the intertemporal MRS of an indifferent household.[x] The indifference condition now equates the rent $p_t^s(h)$ to the "user cost of housing," that is, price less discounted payoff. Equivalently, we can write the current price as

$$p_t(h_t) = p_t^s(h_t) - I(h_t) + E_t[M_{t+1} p_{t+1}(H_{t+1}(h_t))]. \tag{25}$$

Here, the payoff from ownership includes the maintenance cost $I(h_t)$ of an owner-occupied house. We thus obtain a conventional asset pricing equation for houses at quality h.

An alternative derivation starts from the first-order condition of landlords and assumes that there is free conversion between tenant and owner-occupied houses. We can then use the landlord's MRS as a stochastic discount factor:

$$p_t(h_t) = p_t^s(h_t) - I_r(h_t) + E_t[M_{t+1} p_{t+1}(H_{t+1}(h_t))]. \tag{26}$$

For both equations to hold at the same time, we must either have no difference in maintenance cost, or the intertemporal MRSs of landlords and owners must be different. This might be, for example, because landlords are more optimistic than owners and are thus willing to incur more housing risk.

If we solve the user cost (25) forward, and impose a transversality condition on the expected weighted house price in the distant future, the price of a house of quality h can be written as the present value of future rents

$$p_t(h_t) = E_t\left[\sum_{\tau=0}^{\infty} \prod_{j=1}^{\tau} M_{t+j}\left(p_{t+\tau}^s(h_{t+\tau}) - I(h_t)\right)\right]. \tag{27}$$

Since we have assumed that housing risk is spanned, we can further aggregate across quality levels and obtain pricing equations for the entire housing market.

Applied studies often take (27) as a starting point and construct a reduced-form pricing kernel. The test is analogous to testing whether a particular candidate stochastic discount factor prices equity given observable prices and dividends. As we have seen, user-cost equations hold only under special assumptions. We also emphasize that even when those assumptions are met, they represent additional restrictions on prices that hold on top of the equations already discussed above that characterize optimal quality choice conditional on owning or renting.

[x] This is true in particular if the household is a landlord and there is free conversion between tenant and owner occupancy—the household can then assemble REITs portfolios with the same payoffs as any individual house. Of course, the indifferent household may not be a landlord—in the presence of short-sale constraints not all households need to participate in the market for tenant-occupied housing.

3.7 Collateral Constraints

Much of the literature captures the role of housing as collateral by a linear constraint on the amount of short term risk-free debt, our Jth security:

$$-q_{t,J}\theta_{t,J} \le \phi_t p_t(h_t). \tag{28}$$

Households who take out a mortgage must make a large enough downpayment so that the loan-to-value ratio remains below ϕ_t. The maximum loan to value ratio can be random—exogenous variation in ϕ_t is a popular example of a "financial shock" that either loosens or tightens household borrowing capacity. We also shut down borrowing opportunities through risky securities by imposing short sale constraints $\theta_{t,j} \ge 0$ for $j = 1, \ldots, J-1$.

The downpayment constraint (28) goes back to theoretical work on optimal savings by Artle and Varaiya (1978). Slemrod (1982) used it in an early quantitative life cycle model. In 1990s several papers explored equilibrium effects. In static setups, Shleifer and Vishny (1992) stressed the potential for asset fire sales, while Stein (1995) considered its role in generating comovement of house prices and housing volume. Kiyotaki and Moore (1997) emphasized the amplification effects from collateral constraints in a dynamic model. Detemple and Serrat (2003) and Chien and Lustig (2009) study economies with contingent claims subject to collateral constraints. Geanakoplos (2011) endogenizes the downpayment constraint in a model that allows for default.

While the simple constraint (28) provides a tractable way to capture the benefit of housing as collateral, it leaves out several features of observed mortgages. First, while it is in principle possible for the price to drop below the face value of the mortgage over the next period, the chance of this happening is negligible in most quantitative studies. In contrast, in the data many households with long-term mortgages are "under water." Moreover, a key decision for households is whether to prepay and/or refinance mortgages in response to changes in house prices or interest rates. The simple constraint effectively assumes that refinancing is costless, so that an increase in house prices translates directly into higher borrowing capacity. While it may capture the basic tradeoffs well when the period length is relatively long, or when adjustment of mortgage terms is cheap, several applications discussed below show that details of mortgage contracts can matter significantly for quantitative results.

3.7.1 *Household Optimization*

The collateral constraint restricts the choice of the risk-free security in the second stage problem (5): we thus modify that problem by adding the constraints $-\left(\widetilde{w}_t - c_t\right)(1-\alpha'\iota) \le \phi_t p_t(h_t)$ and $\alpha \ge 0$. Denoting the multipliers on these constraint by ν_t and μ_t, respectively, the first order conditions (8) become

$$\begin{aligned} U'(g(c_t,h_t))g_1(c_t,h_t) &= \beta E_t[V'(w_{t+1})]R_t^f + \nu_t \\ \nu_t \iota &= \beta E_t\left[V'(w_{t+1})\left(R_{t+1} - \iota R_t^f\right)\right] + \mu_t. \end{aligned} \tag{29}$$

As long as the constraints do not bind, the conditions are unchanged. If a household runs up against his borrowing constraint, however, the marginal cost of borrowing includes not only the expected repayment, but also the shadow cost of the constraint. This affects indifference conditions at both the borrowing/lending and portfolio choice margins. In particular, if the borrowing constraint is tight (high ν_t) and the expected excess return on a risky security is low, then it may be optimal to not hold that security at all ($\mu_{t,j} > 0$).

If housing serves as collateral, its marginal benefit in (6) reflects its marginal collateral benefit, in addition to the utility benefit from housing services and the expected capital gain. To compare the three components, we focus on the case of continuous housing quality. The counterpart of (13) is

$$p_t'(h_t)\left(1-\phi_t\left(1/R_t^f - E_t^0 M_{t+1}\right)\right) = \frac{g_2(c_t,h_t)}{g_1(c_t,h_t)} - I'(h_t) + E_t^0\left[M_{t+1}p_{t+1}'(H(h_t))H_{t+1}'(h_t)\right]. \tag{30}$$

On the left hand side, the collateral benefit is expressed as a percentage discount to the pricing step p'. From (29), the discount is zero if the household is unconstrained (that is, $\nu_t = 0$ and $E_t M_{t+1} R_t^f = 1$). It is higher if the lower is the intertemporal MRS: collateral is more useful if the household has a greater need for borrowing.

3.7.2 Savings and Portfolio Choice

The constraints imply that household net worth $p_t(h_t) + q_t'\theta_t$ is nonnegative. This feature is useful for matching household portfolios in the data since negative net worth is not common. It also implies that borrowing does not move future income to the present, in contrast to a simple permanent income model. Instead borrowing is a portfolio choice decision, undertaken in order to build a large enough housing position. The forces discussed in Section 3.2 remain at work: households with a lot of future income should choose leveraged housing positions, especially if their labor income is uncorrelated with housing payoffs.

In the cross section, optimal savings depend on the relative abundance of current wealth relative to future income as well as the remaining life span. When wealth is low relative to income, households do not save at all. Young households with low wealth–income ratios save to be able to make a downpayment. As soon as they have saved enough, they build leveraged portfolios in housing and also some attractive other assets, such as stocks. Older households have higher wealth–income ratios and are thus long in all assets.

As wealth rises relative to income, households start saving until their savings rate approaches an unconstrained optimal savings rate that depends on the distribution of returns—it is constant when returns are iid. Younger households have a longer planning horizon and therefore spread their savings over more years. This effect tends to increase the savings by the young. However, middle aged households have more income, so that they can save more. The higher savings rates of young households dominate when labor income is not important, which means at high wealth–income ratios. For empirically relevant ranges of the wealth–income ratio, the higher savings of the middle aged dominate and create a hump-shaped wealth pattern, which we also see in the data.

Another implication is that constrained households are more reluctant to buy risky securities. Indeed, consider the first-order conditions (29) for households who hold securities: constrained households are indifferent between risky securities and risk-free investment only if the marginal utility weighted expected excess return is strictly positive. In contrast to housing, securities do not come with collateral benefits, and thus require higher premia in order to be held. This feature helps in applications to explain why young households with low cash relative to income do not hold equity even though the equity premium is high.

3.7.3 The Pricing of Securities

The first order conditions (29) suggest that the presence of a collateral constraint might help generate more volatile expected excess returns on risky securities, and hence help resolve the volatility puzzle. Indeed, changes in the tightness of the constraint do affect conditional risk premia. However, a problem with this effect is that it also tends to generate volatility in the risk-free interest rate. Combining the first-order conditions, we obtain

$$U'(g(c_t, h_t))g_1(c_t, h_t) = \beta E_t[V'(w_{t+1})R_{t+1}].$$

The marginal condition for the level returns of risky securities is thus the same as without a collateral constraint. In applications that generate volatility in expected excess returns, that volatility is typically due to volatility in the risk-free rate moves, as opposed to volatility in conditional risky returns as in the data.

3.7.4 House Prices

The presence of a collateral constraint also alters the pricing of houses. The most immediate effect is that if constrained households buy houses, then the collateral benefit increases house prices, holding fixed payoffs and the households' intertemporal MRS. This is a liquidity effect that occurs even with linear pricing and when dividends are tradable. Dividing (29) by the big bracket on the left hand side, we have that housing payoffs are discounted at a lower rate to price in the collateral benefit.

Whether the liquidity effect is important for price movements in a heterogeneous agent model depends on market structure and the presence of other constraints. Collateral constraints provide an important example why households can be affected differently by shocks—for example, the financial shock ϕ_t affects (30) if and only if the household is constrained. At the same time, as discussed in Section 3.3, shocks alter a family of Euler equations via both price and quantity adjustment. The effect on prices will be higher if market structure requires price adjustment because quantities do not move.

To illustrate, suppose ϕ_t increases to relax the downpayment constraint. In the typical population, some households are constrained while others are unconstrained. Suppose now the model assumes linear pricing because houses of different quality can be converted freely. In order for a housing boom to occur, the price per unit of housing will move only if the shock is strong enough to alter the valuation of payoff by unconstrained households; otherwise quantity adjustment will provide more housing for constrained households accompanied by a smaller price reaction. In contrast, with nonlinear pricing and indivisibility, the shock can strongly affect the prices of houses bought by constrained households, without a big impact on the Euler equation of the unconstrained. With limited quantity adjustment, the overall effect on prices can thus be bigger.

3.8 Transaction Costs

We now introduce a proportional transaction cost κ whenever a household sells a house. This tractable and popular specification is often motivated by the rule of thumb that about 6% of the house price are typically paid to the seller's agent in a transaction. It was first studied by Flemming (1969) in a deterministic context and by Grossman and Laroque (1990) in a stochastic model. Beyond these direct costs, it is plausible that most households face other moving costs, either pecuniary—such as changing local services—or possibly nonpecuniary, for example, disutility from leaving a familiar environment. Such costs sometimes motivate a fixed component to moving costs. In what follows we work only with proportional costs since those are sufficient to understand the key effects.

Once transaction costs are taken into account, the existing house becomes *illiquid* and its quality at the beginning of the period $\tilde{h}_t = H_t(h_{t-1})$ becomes a separate state variable in the household problem. We thus write the value function as $V_t(w_t, \tilde{h}_t)$ where w_t is total wealth at the beginning of period t as before. We introduce separate notation for $\tilde{h}_t$ since it depends not only on quality chosen in the previous period, but may also depend on random events such as depreciation. The presence of transaction costs does not affect choices in the second stage problem from Section 3.2. To keep track of the new state variable, we only need to modify the expected continuation utility in the objective to $E_t[V_{t+1}(w_{t+1}, H_{t+1}(h_t))]$.

Let $m_t \in \{0,1\}$ denote the moving choice. The first stage problem is now

$$V_t(w_t, \tilde{h}_t) = \max_{m_t, h_{t+1} \in \mathcal{H}} m_t \tilde{V}_t\left(w_t - \kappa p_t(\tilde{h}_t) - p_t(h_t), h_t\right) + (1 - m_t)\tilde{V}_t(w_t - p_t(\tilde{h}_t), \tilde{h}_t) \quad (31)$$

The first term is the utility of a mover who sells the old house, incurs the transaction cost and buys a new house. The second term is the utility of a stayer: house quality remains unchanged, and the disposable funds for consumption and securities in the second stage problem consist of *liquid wealth*, that is, wealth net of the illiquid house.

To illustrate the benefits of illiquid housing, consider the model with continuous quality. The marginal benefit of house quality at the beginning of the period consists of the effect of housing on wealth as well as the direct benefit. From the envelope theorem, the total marginal benefit is

$$V_{t,1}p'(\widetilde{h}_t) + V_{t,2} = m_t \widetilde{V}_{t,1}^{\text{move}}(1-\kappa)p_t'(\widetilde{h}_t) + (1-m_t)\widetilde{V}_{t,2}^{\text{stay}}, \tag{32}$$

where the subscripts in $\widetilde{V}_t^{\text{move}}$ and $\widetilde{V}_t^{\text{stay}}$ indicate whether $\widetilde{V}_t$ is evaluated at the first or second arguments in (31). A mover household, enjoys the marginal benefit of liquid funds conveyed by an extra unit of house quality. In contrast, a stayer household experiences no increase in liquid fund and only enjoys the continuation utility benefit of house quality.

The household problem illustrates three key new features of pricing introduced by transaction costs. First, only movers can be marginal investors in housing in any given period. Since housing has low turnover, the characteristics of only a few people matter directly for determining prices. Second, the value of housing depends less on future prices if moving is less likely. Indeed, (32) shows that the price matters more the higher is m_t. In the extreme case where households know they will never move in the future, their benefit from housing is independent of future prices. Finally, transaction costs lower marginal benefit, and this effect is capitalized into house prices. Other things equal, we thus expect lower prices in markets with higher turnover.

Transaction costs also alter portfolio choice tradeoffs described earlier. First, they make ownership more expensive than renting, and more so for households who expect to move again quickly. In a market with heterogenous agents, the price will compensate the average investor for future transaction costs. It is more likely then that frequent movers prefer renting. Second, households with rising income profiles may leverage even more so they can lock in a large housing position early and do not have to move later. Finally, collateral constraints are more likely to bind even for rich households: whether constraints bind depends on the amount of liquid resources $\widetilde{w}_t$ in the second stage problem. With transaction costs, household may let $\widetilde{w}_t$ decline even though total wealth w_t is large.[y]

4. THEORY VS DATA

We are now ready to discuss work that quantifies the framework in Section 3 and studies its implications in various applications.

4.1 Magnitudes

At the core of any quantitative work based on the framework in Section 3 is the individual household problem. For a problem in which households may choose to buy an

[y] The issue is compounded in a model with long term mortgages that are costly to adjust so that the mortgage position also becomes illiquid. The household then faces a liquidity constraint unless he either sells the house or adjusts the mortgage. As long as he does neither, a change in house prices does not alter funds available for spending.

individual house, it is important to correctly specify the risk-return trade-offs involved. As discussed in Section 2.2, the prices of individual houses are highly volatile. Moreover, a large component of this volatility is idiosyncratic. House prices may also covary with income and other asset prices. These return moments can be taken from empirical studies that estimate their means and covariances with micro data, such as individual property level data and the PSID. Below we discuss whether and how the magnitudes matter in applications.

4.1.1 Preference Parameters

Since housing expenditure shares in the data are similar over time as well as across households (as discussed in Section 2.1), a common specification of the aggregator (2) over housing consumption and nonhousing consumption is Cobb–Douglas. The preference parameter is set equal to the expenditure share on housing, which is roughly 20%.

The choice of the risk aversion parameter depends on whether the portfolio choice problem involves other assets such as stocks. As discussed in Section 2.2, high transaction costs and high volatility lower the Sharpe ratio of individual houses and thereby reduce their attractiveness. In the absence of more attractive assets, a household problem with low risk aversion around 5 will have reasonable implications for optimal portfolios. When the problem allows households to invest in more attractive assets such as stocks, low risk aversion will typically lead to extreme optimal portfolios that exploit the equity premium. To explain observed household portfolios, higher risk aversion or higher perceived risk about stock returns are needed, or high participation costs in the stock market.

4.1.2 Shocks

Exogenous moving shocks capture reasons for moving that are exogenous to the model. The probability of such shocks can be estimated from the American Housing Survey which asks households about their reasons for moving. Roughly a third of movers provide reasons that are unrelated to the economic reasons for moving captured in the models. Examples are disasters such as fires or floods, marriage, divorce, death of spouse, etc. This 1/3 probability is multiplied by the overall probability of moving which is roughly 1/10 per year, resulting in a 1/30 probability for an exogenous move per year.

Households face exogenous survival probability that depend on age. These survival probabilities can be taken from life tables published by the National Center of Health Statistics.[z]

The volatility of individual house prices has a large idiosyncratic component. As discussed in Section 2.2, the volatility of exogenous idiosyncratic shocks is around 9–15% per year. A small component of individual house prices also correlates with aggregate income and other asset prices (such as stock prices). This component can be estimated by assuming that house prices grow at the aggregate growth rate of the economy.

[z] Their website is http://www.cdc.gov/nchs/products/life_tables.htm

A common specification for individual income is

$$\log y_{it} = f(t, Z_{it}) + v_{it} + \varepsilon_{it} \tag{33}$$

where $f(t, Z_{it})$ is a deterministic function of age and a vector of other individual characteristics Z_{it}, ε_{it} is an idiosyncratic temporary shock distributed $N(0, \sigma_\varepsilon^2)$ and permanent income v_{it} is given by

$$v_{it} = v_{i,t-1} + u_{it}$$

where u_{it} is distributed as $N(0, \sigma_u^2)$ and is uncorrelated with ε_{it}.

Individual log income is the sum of the age profile, the permanent component v_{it} and a transitory shock ε_{it}. The deterministic age profile is a third-order polynomial in age, which is estimated to match the observed hump-shaped life-cycle profile of income. Carroll (1997) and Gourinchas and Parker (2002) assume that the process for the persistent component v_{it} is a random walk as in the last equation. Hubbard et al. (1995) estimate an AR(1) for v_{it} and find that the autocorrelation coefficient is indeed close to one. Cocco et al. (2005) report estimates for the standard deviation σ_u of persistent shocks around 10–13% per year in table 3, depending on education. Their estimate for the standard deviation of σ_ε of transitory shocks is around 22–31%. It is common to somewhat reduce these numbers to account for measurement error in the PSID. For example, Campbell and Cocco (2003) use 2% for σ_u and 14% for σ_ε.

The transitory shock ε_{it} is uncorrelated across households. The persistent shock u_{it} can be decomposed into an aggregate component ξ_t and an idiosyncratic component ω_{it},

$$u_{it} = \xi_t + \omega_{it}.$$

The aggregate component ξ_t helps to introduce correlation between individual labor income and aggregate variables, such as aggregate income or asset prices.

The process (33) is specified for income received in periods t before retirement τ. After retirement, income may be a fraction λ of permanent labor income in the last working year

$$\log y_{it} = \log \lambda + f(\tau, Z_{i\tau}) + v_{i\tau} \quad \text{for } t > \tau.$$

This approach is taken in Cocco, Gomes, and Maenhout (2005) , who estimate λ as the ratio between the average income for retirees in a given education group to the average labor income in the year before retirement. The estimate is between 68% and 94% in their table 2.

With this specification of the income process, households do not face any further risks after they retire. This assumption abstracts from a number of risks that older households face, especially uncertain life spans and out-of-pocket medical expenses. Recent work has made progress to quantify such risks. For example, De Nardi et al. (2010) estimate large and volatile medical expenses for retired singles. Moreover, they find that the volatility of shocks to medical expenses increases with age and permanent income. In a life cycle model, the risk of living long and requiring expensive medical care is an important reason to save for many older high-income households. More empirical work is needed that quantifies these risks for nonsingle households as well as distinguishes the savings motives in the presence of these health risks from bequest motives. In the meantime, it seems reasonable to assume the individual income process (33) for $t > \tau$ and thereby to allow for shocks during retirement.

4.1.3 Other Housing Parameters

Houses depreciate at a rate between 1.5% and 3%, as discussed in Section 2.1. With the assumption of "essential maintenance," the depreciation rate is also the fraction of the house value that is spent on maintenance. Transaction costs vary across cities and states as well as the price spectrum within cities. They are between 6% in real estate fees (for example, in California) and 10% when moving costs are included.

4.2 Consumption, Savings, and Portfolio Choice

The literature on consumption-savings problems is concerned with the facts in Section 2.5. Empirical work documents cross-sectional patterns of consumption and portfolios, and measures properties of returns and income that are relevant for optimal portfolio choice. In order to confront theory and data, a common approach is to quantify a household problem with frictions as discussed in Section 3. Some studies impose equilibrium but nevertheless emphasize cross-sectional patterns. Much of the work discussed in this section precedes the financial crisis and focuses on cross-sectional patterns that are a key feature of any quantitative study.

We divide the literature into five groups. The first considers quantitative models with housing and one other asset. The focus then is on consumption and savings in housing vs other goods or assets, respectively The second group tackles explicitly the choice of equity portfolios with return properties as in the data. This is a more challenging problem since it requires not only matching facts on housing but household behavior toward equity, a well known puzzle in its own right. Third, we consider a set of papers that looks for reduced form evidence on specific mechanisms at work in portfolio choice models, especially the role of housing as a hedge. We then discuss the effects of more complex mortgage products, as well as the effects of house prices on consumption.

Quantitative models can successfully explain wealth and portfolio patterns over the life cycle. The models predict that wealth is positive and hump shaped, as discussed in Section 9. Moreover, the models imply that young households hold highly leveraged portfolios in housing, while older and richer households have positive positions in many assets, including bonds and stocks. According to the models, the hump shape in the wealth position will translate into hump-shaped positions in other assets such as houses and stocks. These age patterns are roughly consistent with the data, especially for housing. However, the models struggle to explain the high concentration in wealth we observe in the data, especially the extreme concentration in stock wealth. The extensive margins are also hard to match for these models. It remains a puzzle why so many middle-class households choose not to participate in the stock market. It is also difficult to quantitatively match the homeownership rate along various dimensions of heterogeneity such as income and wealth.

4.2.1 Housing and Savings Over the Life Cycle

Early work on housing choice over the life cycle considered savings via multiple capital goods without price risk. Households face a two asset special case of the problem with collateral constraints in Section 3.7. Fernandez-Villaverde and Krueger (2010) consider a household with a finite horizon who accumulates capital used in production as well as a stock of durables that enters the utility function. The income process has idiosyncratic shocks and a deterministic age profile. The collateral constraint is important to explain the accumulation of durables early in life, as described in Section 3.7.

Yang (2009) narrows the definition of durables to housing consumption and focuses on the accumulation of housing. The key new feature in her setup is transaction costs for adjusting the housing stock, as in Section 3.8. Those costs are shown to be important for matching the slow downsizing of the housing stock late in life that is observed in the data. Both papers conclude that a standard life cycle model is broadly successful at explaining the hump-shaped patterns in nondurable consumption, durables, and wealth by cohort.

Focus on cohort averages omits variation along the extensive margins, that is who owns and who rents. A number of papers explore the various determinants of tenure choice discussed in Section 3.6. A new feature in these paper is uninsurable house price risk that may correlate with income risk. Li and Yao (2007) study tenure decisions in an environment where renting is expensive; rents are a higher fraction of the house value than the sum of maintenance and mortgage rates. The paper confirms the earlier findings regarding hump-shaped patterns in nonhousing consumption. A new feature in the model is a hump-shaped homeownership rate, which is the overall pattern shown in their fig. 8(a). The homeownership rate in the model shown in fig. 7(a) is a more extreme function of age than the data: all households aged 30 years and below rent, while all households aged 40–80 years own. This discrepancy illustrate the difficulty associated with quantitatively accounting for the extensive margin. The paper studies a number of counterfactuals in which older households benefit from house price increases, while younger households loose.

While Li and Yao (2007) abstract from taxes and directly assume that renting is costly, Díaz and Luengo-Prado (2008) embed the US tax system into their model. The paper carefully compares housing costs for renters and homeowners. It finds that rental equivalence approach (as used in the NIPA tables) overestimates the costs of owner-occupied housing services by roughly 11%. Reasons include the differential tax treatment of renter-occupied vs owner-occupied housing services, the tax deductability of mortgage interest rates and transaction costs in housing markets.

Chambers et al. (2009a,b) study an equilibrium model with tenure choice. with long-term mortgage contracts. The model parameters are estimated with 1994 data. Table 2 in Chambers et al. (2009b) shows that the 1994 model predictions match the homeownership rate as a function of age quite well. Its predicts, however, that all households with income in the upper 40% of the income distribution should own—again, illustrating how difficult it is to match the extensive margin along observed dimensions of heterogeneity.

Attanasio et al. (2012) match the homeownership rate by age and education in a setup with two discrete house sizes: flats and houses. The paper documents that transaction costs are crucial for both homeownership and the property ladder. Lower transaction costs increase the homeownership rate because they increase the number of young households who find it optimal to buy a flat before upgrading to a house.

Home Equity as a Buffer Stock for Consumption Smoothing

Hurst and Stafford (2004) study a life cycle problem in which homeowners may want to use the equity from their house as a buffer stock to smooth their consumption. When homeowners with low savings in liquid assets (such as checking accounts or stocks) experience an adverse income shock, they may have to drastically lower their consumption. To avoid a painful cut in consumption, these homeowners may want to refinance into a mortgage with a larger principal. While refinancing might not necessarily lower the costs of their mortgage, it helps their desire to smooth their consumption.

Hurst and Stafford provide empirical support for this mechanism with micro data from the PSID. Households who were unemployed between 1991 and 1996, and who had zero liquid assets going into 1991, were 25% more likely to refinance than otherwise similar households. They also were more likely to extract equity during the refinancing process.

The life cycle problem in Hurst and Stafford is not designed to be quantitative. For example, it has constant house prices and a fixed house that cannot be sold, the income process is iid, and mortgages are interest-only. Chen et al. (2013) introduce a choice between renting and owning, house price risk, aggregate and idiosyncratic persistent income risk, long-term mortgages, and various frictions (such as loan-to-value and loan-to-income constraints.) When the observed historical paths for house prices, aggregate income and interest rates are taken as given, the model predicts a dramatic increase in mortgage debt during the 2000s house price boom. A significant portion of the debt increase is associated with home equity extraction in the model as well as in the data.

Mian and Sufi (2011) provide new empirical evidence on the importance of this mechanism with micro data from the recent housing boom. The paper documents that existing homeowners—households who already owned their home in 1997—started to borrow significantly more during the early 2000s. The tendency to extract equity was strongest among young homeowners with low credit scores and high credit card utilization rates, while homeowners with good credit scores did not extract more equity from their house.

4.2.2 Household Portfolio Choice

Accounting for risk in household portfolios requires combining the illiquidity and collateralizability of housing with a richer menu of securities. Portfolio choice then depends on risk in multiple tradable assets, as described in Section 3.2. Houses now enable households to borrow and invest their liquid funds in assets with more attractive return properties, such as stocks. When transaction costs are high, illiquid houses act as undiversifiable background risk (similar to nontradable labor income) in portfolio choice.

Myopic Investors

Early work focused on the risk return tradeoffs in models with myopic investors. Berkovec and Fullerton (1992) study a two period general equilibrium model in which households consume housing and choose a portfolio of owner-occupied housing, housing as an investment, stocks, and bonds. Ownership is attractive because of tax subsidies, but exposes owners to undiversifiable risk. Indeed, the paper estimates the variance of house prices as the sum of national, regional and intraregional effects on house prices, resulting in a volatility of 8.2% per year.

Starting from the current US tax system, the paper runs counterfactuals to eliminate subsidies, namely that owner-occupied housing services are not taxable, nominal mortgage interest is deductable, and that there is an extra deduction for property taxes. Starting from the current US tax system, the effects are a priori ambiguous: while abolishing subsidies lowers the average return on housing, it also reduces the variance of returns—the government becomes a silent partner who shares both gains and losses on the house. The overall effect on homeownership then depends on risk attitudes and tax brackets.

Flavin and Yamashita (2002) study portfolio choice by myopic investors with an emphasis on the illiquidity of housing. The setup resembles the second stage problem from Section 3.2: the position in housing is predetermined. Households have mean variance preferences and the focus is on the asset portfolio: there is no explicit consumption margin and no labor income. The portfolio share on housing thus acts as a constraint on the problem of choosing a portfolio of financial assets, namely short and long bonds, stocks and a mortgage. Bonds and stocks cannot be sold short and there is a collateral constraint: the mortgage cannot exceed the value of the house.

Flavin and Yamashita construct the returns on an individual house from PSID data. The housing return has a high volatility as in Table 1, and a zero correlation with financial returns. The solution to the portfolio choice problem that includes housing is an efficient frontier that achieves the minimum-variance portfolio for a given expected return subject to the housing constraint. The constraint is matched to average portfolio shares on housing for various cohorts in the PSID. As discussed in Section 2, these observed portfolio shares decline in age.

For households with a high portfolio share on housing, the optimal portfolio involves the maximum possible amount of mortgage borrowing. Since leverage is risky, any remaining funds are invested in a safer financial portfolio consisting of mostly bonds, while the shorting constraint binds for the short bond. For households with a lower portfolio share on housing, the position in housing is less leveraged. These households choose more risk in their remaining portfolio by increasing their portfolio weight on stocks. Higher risk aversion lowers the risk in the optimal portfolio by reducing leverage and shifting the remaining portfolio toward bonds.

A high portfolio share on housing is typical of young households, while middle-aged households have a lower portfolio share. By connecting the magnitude of the initial housing constraint with data on age profiles, the mean-variance benchmark provides intuition for why younger households hold a lower portfolio share in stocks than older households.

Housing and Other Assets Over the Life Cycle

Cocco (2005) studies the consumption–portfolio choice problem of an owner–occupier household with finite horizon. The household receives a nontradable income process (33) with both transitory and persistent shocks. The household can choose stocks, bonds, housing and a mortgage. The returns on stocks are iid and uncorrelated with aggregate income risk, while the price of the house is perfectly correlated with aggregate income risk. The real interest rate on bonds and the (higher) mortgage rate are constant.

The consumption–portfolio problem has several important constraints. The first two of these constraints are similar to those in Flavin and Yamashita (2002). First, bonds and stocks cannot be shorted. Second, there is a downpayment constraint; the mortgage cannot exceed a fraction of the house value. A new feature in Cocco's setup is that households choose the size of their house (while the house is fixed and acts as a constraint in Flavin and Yamashita). A third constraint is that houses have a minimum size. Together with the downpayment constraint, the minimum size creates a strong motive to save for young households, especially in the absence of a rental market. There are additional frictions in the form of transaction costs for housing and a one-time fixed cost to participate in the stock market.

The model generates low rates of stock market participation among poorer households—consistent with the data—who are not willing to pay the fixed costs to participate. Households with enough wealth get a large mortgage and invest most of their portfolio in housing; they still choose not to participate in the stock market. Richer households participate in the stock market and increase their portfolio share on stocks with wealth. Over the life cycle, the model with housing is successful at predicting that young households are house poor: they take a large mortgage and buy a house, while they do not participate in the stock market. As they grow older, they pay down their mortgage and invest more in the stock market, as in the data.

Yao and Zhang (2005) study a life cycle problem in which households can choose between owning and renting a house. The possibility to rent is important for younger and poorer households who do not have enough savings to afford the downpayment.

Older, wealthier households choose to own a house. The downpayment is equity in the house, which acts as a buffer against income shocks. Once they own a house, households have riskier portfolios because of the leveraged position in housing. But homeowners still invest a larger fraction of their (nonhousing) portfolio in stocks for diversification reasons, because of the low correlation between stock and housing returns.

4.2.3 Housing as a Hedge

Section 3.2 emphasizes that once risk is explicitly taken into account, the attractiveness of housing depends on the covariance of housing returns with other random state variables in the future. Those state variables include, among others, (*i*) labor income, a component of future wealth (*ii*) the price of rental housing which affects continuation utility in a problem with a rental market and (*iii*) house prices in other markets if the household is subject to moving shocks or has the option to move across different markets. We now consider evidence on these effects.

Housing as a Hedge Against Income Risk

Housing is riskier for households whose incomes covary positively with house prices. For these households, housing is not as good a hedge against income risk. These households will thus tilt their portfolios away from housing toward other assets. Cocco (2005) shows that this effect is quantitatively small in his life cycle model. For example, raising the correlation coefficient between income and house prices from 0 to 0.33 lowers the portfolio share on housing by 1 percentage point. The effect is small because housing is not only an investment but also a consumption good.

Davidoff (2006) provides empirical evidence on the effect. The paper first uses time series data to estimate the covariance between income and house prices in various regions and industries. The paper then predicts the value of owner-occupied housing as well as tenure choices in the 1990 census with the estimated covariances. The results show that a one-standard deviation increase in income–price covariance is associated with a $7500 reduction in the value of the housing investment for owners. They also show that a higher income–price covariance has a negligible effect on the probability of renting.

Housing as a Hedge Against Rent Risk

A common and reasonable assumption is that every household needs to consume some housing services. In a setup with a rental market as in Section 3.6, those services can be obtained either in a rental market or by buying a housing asset that promises a stream of housing services. The rental market is a spot market, where housing services are sold at the current rental rate which fluctuates over time. By buying a house, households can lock in a known price for a stream of future housing services. They still face house price risk in case they need to sell the house later because, for example, they want to move to a new city.

Sinai and Souleles (2005) compare the two sources of risk in a simple spatial model with two locations. Households choose whether to rent or own a single housing unit to maximize their expected wealth net of the housing costs. There is a fixed number of housing units equal to the number of households. The stochastic processes for rents in the two locations are exogenous. Rents are AR(1) processes with correlated shocks. After a known number years, households move from one to the other location. The price of owner-occupied housing units is determined endogenously and clears the housing market. In the model, both the demand for homeownership and equilibrium price–rent ratios tend to increase with expected tenure, the volatility of rents and the correlation between rents across locations.

Table 1 of Sinai and Souleles documents a 2.9% volatility of real rents at the MSA level during the years 1990–98, almost half the volatility of MSA real house prices. Much of this volatility is variation across MSAs. For example, rent volatility ranges from 1.7% in Fort Lauderdale to 7.2% in Austin. Tables 2 and 3 documents that both the probability of owning estimated from a probit model and price–rent ratios are higher in areas with higher mean tenure rates and rent volatility.

Housing as a Hedge Against Future House Prices

Lu (2008) solves a life cycle problem with many locations. The problem assumes that households know that they will want to move in the future, sell their house in the current location and buy a house in the new location. Whether or not the current house can act as a hedge for the future house purchase depends on the correlation between house prices across locations. The conventional wisdom is that correlation in house prices across housing markets is low. Since house prices within MSA are more correlated than across MSAs, the hedging motive will be more important for moves within metropolitan areas. The paper documents some evidence on the importance of such within-MSA moves. It reports that among households in the PSID from 1968 to 1997, 62% of them traded up later by buying a more expensive house (in real terms). Among households who traded up, 71.3% of them moved within the same metropolitan area.

Sinai and Souleles (2013) document that the correlation of house prices across MSAs is indeed low. They estimate this correlation with annual observations on the OFHEO constant-quality MSA-level house price index over the years 1980–05. The simple unweighted median correlation in real house price growth across MSAs is 0.35. Sinai and Souleles argue that households do not move randomly across MSAs. Instead, households move between housing markets that are highly correlated. The paper computes the household's own expected correlation in house prices across MSAs by weighing each correlation with the probability that the household will move to that MSAs. The data for moving from one MSA to another MSA is from the US Department of the Treasury's County-to-County Migration Patterns. The resulting expected correlation is 0.60 for the median household.

4.2.4 Mortgage Choice and Refinancing

Mortgages are often modeled as short-term debt contracts, as we did in Section 3.7. In this case, the collateral constraints (28) can capture home equity lines of credit. Most mortgages are long-term debt contracts, however. Recent work has therefore started to incorporate longer maturities as well as other features, such as fixed vs floating mortgage rates, deferred amortization, prepayment penalties, etc. Much more work is needed in this area to understand the recent foreclosure crisis, the welfare losses associated with certain contracts more broadly, and their implications for financial regulation.

Fixed vs Adjustable Rates

Campbell and Cocco (2003) study the choice between a fixed-rate mortgage and an adjustable-rate mortgage in a life cycle model. The household receives the nontradable real labor income process (33). The growth rate in house prices experiences iid shocks. The only other asset is a short-term real bond with an interest rate that is also hit by iid shocks. Expected inflation is an AR(1), so that inflation is an ARMA(1,1). The nominal short-term interest rate is the sum of expected inflation and the real rate. Longer-term nominal interest rates are determined with the expectations hypothesis. Adjustable mortgage rates include a constant default premium, while fixed rates include a default premium as well as a compensation for prepayment risk, both are constant as well.

The household buys a house with a minimum downpayment and finances the remaining balance with either an adjustable or fixed rate mortgage. A nominal fixed-rate mortgage without prepayment option is a highly risky contract, because the real present value of its future payments is sensitive to inflation. The prepayment option insures households against a surprising fall in nominal interest rates, because they can refinance at the lower rate. The option is not free, however—it is priced into a higher fixed rate. During times of low inflation and low real rates, the fixed rate mortgage is thus an expensive form of borrowing. An adjustable-rate mortgage is safe because the real present value of its future payments is unaffected by inflation. However, it comes with real payments that vary over time with expected inflation and real rates. These high payments may coincide with adverse income shocks and low house prices, so that homeowners may not be able to borrow more to meet these payments.

The optimal choice between the two mortgage contracts compares the expected costs for the homeowner over the life of the mortgage with the risks associated with higher or lower realizations of these costs. The expected costs for the homeowner are either the expected adjusted rate over the life of the mortgage or the known fixed rate. The risks associated with higher cost realizations matter more for homeowners who are either risk averse or close to their borrowing constraints. These homeowners tend to have low savings, large houses relative to their income and volatile incomes. The horizon matters for computing the expected adjustable rate over the life of the mortgage. For homeowners who are likely to move in the near future, the current adjustable rate matters more. These homeowners will compare the current adjustable rate with the fixed rate and opt for the rate that is currently cheaper. Since fixed rates include the cost of the prepayment option and are longer maturities interest rates, the cheaper rate will on average be the adjustable 1-year rate.

More generally, the difference between the fixed rate and the expected adjustable rate over the life of the contract is determined by risk premia (as well as the cost of the prepayment option). These risk premia vary over time. Koijen et al. (2009) compute a time series these risk premia and show that they highly correlate with the actual share of adjustable-rate mortgages among newly originated mortgages. The expected adjustable rate can be computed, for example, with survey data on interest rate forecasts, VARs or some other estimated time series process, or under the assumption that beliefs are extrapolative. Badarinza, Campbell, and Ramadorai (2016) investigate the share of adjustable-rate mortgages in cross-country data. They find that low expected adjustable rates over short horizons, such as a year or a few years, relative to fixed rates are associated with a high share of adjustable rate mortgages.

Deferred Amortization Contracts

Piskorski and Tchistyi (2010) is a theoretical study of optimal mortgage design in a setup in which income by an impatient household is stochastic and unobservable by the lender. The household needs to borrow from the lender to be able to buy a house. The paper shows that the optimal contract is a combination of an interest-only mortgage and an equity line of credit—an alternative mortgage product that offers deferred amortization. The intuition behind the result is that deferred amortization helps borrowing-constrained households to smooth their consumption.

Chambers et al. (2009a,b) study mortgage choice in a quantitative general equilibrium model with tenure choice and long-term mortgage contracts. The model parameters are estimated with 1994 data. The model is recomputed with 2005 data by offering households a range of mortgage contracts with lower downpayment constraints, other forms of deferred amortization, and lower closing costs. The paper finds that these new mortgage contracts have enabled many borrowing-constrained renters to buy a house. It concludes that these mortgage innovations can explain around 70% of the large increase in the homeownership rate from 1994 to 2005.

Cocco (2013) provides empirical evidence that supports this consumption-smoothing mechanism with data from the British Household Panel Survey. The survey collects detailed housing information from a group of households over time (for example, about the type of mortgage, the year the mortgage began, the amount borrowed, monthly payments, etc.) The paper documents that, at least since 2001, households who choose alternative mortgages are better educated and have higher subsequent income growth These mortgages are used to buy expensive houses with high loan-to-value ratios. Amromin et al. (2013) document similar evidence for alternative mortgages in the United States.

Suboptimal Borrower Behavior in Mortgage Markets

Mortgages are complex products. Mortgage lenders do not have the incentives to make these contracts comparable, unless forced by regulation. Households make their mortgage choice infrequently and cannot learn much from their past mistakes. In this situation, it is not surprising that mortgage choices are not made optimally.

Woodward and Hall (2012) show that new home buyers vastly overpay for their mortgages. They use data on a sample of 30-year fixed-rate mortgages insured by the Federal Housing Administration to show that borrowers do not push brokers toward competitive pricing. Most borrowers would benefit from comparing quotes from a larger number of brokers. Borrowers would also benefit from comparing quotes of mortgages that do not involve any up-front cash payments, such as points. These findings hold especially for less educated borrowers.[aa]

A large literature on mortgage-backed securities documents that households' refinancing behavior is suboptimal. For example, Schwartz and Torous (1989) and Stanton (1995) show that many households do not refinance their fixed-rate mortgage when market rates fall below their locked-in contact rate. Other households refinance even though market rates are above their locked-in contract rate. Agarwal et al. (2013) develop a formula for the (S,s) inaction range for refinancing in the presence of fixed costs. Anderson et al. (2015) document suboptimal refinancing behavior in Denmark, especially for older, less educated and lower income households.

4.2.5 Consumption Response to Higher House Prices

House price booms are often associated with higher aggregate household consumption. What are the mechanisms that explain the consumption increase? There are two related issues. The first is to measure the marginal propensity to consume (MPC) out of housing wealth for different groups of consumers. The second is to identify what exogenous shocks might have given rise to the joint movement in house prices and consumption. For example, was the boom generated by changes in financial conditions, or rather by an increase in household income.

The household problem from Part 2 suggests potential determinants of the MPC out of housing wealth. Consider first the frictionless problem from Section 3.2. Here, an increase in the house price has three possible effects: it changes the relative price of housing services, it may change expectations of returns on housing or other assets in the future (that is, it may change the continuation utility V_{t+1}), and it changes current wealth (or cash on hand). Only the first effect is unique to housing which has a nontradable dividend—the latter two effects are shared by any other security.

[aa] Woodward and Hall (2012) also show that minority households overpay more for their mortgages. Important early work on discrimination in mortgage markets is the paper by Munnell et al. (1996). These authors show that minorities are more than twice as likely to be denied a mortgage as whites.

Berger et al. (2015) provide conditions on the problem such that only the last effect prevails. In particular, they consider permanent price changes that do not alter return expectations, and they assume Cobb–Douglas felicity so that income and substitution effects of the relative price of housing services cancel. They also point out that the result does not depend on the presence of incomplete markets, a rental market or a collateral constraint of the type (28)—indeed, the result is due to the fact that cash in hand is the only state variable that house prices affect which is true in all of these cases. The result does not hold once transaction costs for either housing or mortgages are added.

Consumption and House Prices in Life Cycle Models

Many studies have analyzed aggregate data on consumption during housing booms. For example, Muellbauer and Murphy (1990) argue that in the UK house prices in the 1980s generated a wealth effect on aggregate consumption that was enabled by financial liberalization. The liberalization allowed households to extract more wealth from the value increase in their illiquid housing investment. King (1990) and Pagano (1990) question the importance of wealth effects in accounting for the high correlation between house prices and consumption in the United Kingdom. They argue that higher income growth expectations account for the increase in consumption and also for higher house prices.

Micro evidence on household consumption helps distinguish between competing mechanism. For example, Attanasio and Weber (1994) argue that in basic life cycle models (with a single good and a single asset), wealth effects vs higher income expectations have different predictions for the consumption of younger vs older households. Older households have more wealth and have shorter horizons over which to spread an increase in their wealth. Therefore older households will increase their consumption more than younger households in response to a 1% increase in wealth. Younger households respond more to income shocks, because they have more human wealth. The paper uses micro data from the UK family expenditure survey (FES) to document that the 1980s consumption boom was driven in large part by strong consumption by young households.

In a life cycle model with housing as a collateral asset, the predictions of these mechanisms are less obvious. In this setup, higher house prices not only increase the wealth of homeowners as in the basic model, but also relax collateral constraints and thereby enable the young to consume more. Attanasio et al. (2011) solve such a life cycle problem with exogenous house price and income processes in which the collateral constraint (28) is only imposed in the period when a house is bought or the mortgage amount changes. They use the observed aggregate time series for house prices and income to extract two shock series. The paper then feeds the two shocks separately into the model and analyzes how various age cohorts adjust their consumption to a particular shock. The quantitative results show that the intuition from the basic life cycle model carries over to this model with housing: higher house prices lead to stronger consumption responses by older households, while higher income causes stronger consumption responses by young households. The paper again concludes that the evidence suggests that higher income expectations or common shocks that affect both income and house prices are more important than pure wealth effects.

Homeowners who want to consume more in response to higher house values need to adjust their portfolio position either by selling their house or by borrowing more against their house. When transaction costs are high in housing and mortgage markets, the costs of adjusting the portfolio may not make it optimal to cash out. Indeed, Berger et al. (2015) show that their theoretical result does not apply in the presence of transaction costs. They find that a model with transaction costs around 5% has approximately the same consumption elasticities than a model without transaction costs. With higher transaction costs, around 10%, the MPC × house value formula overstates consumption elasticities, especially for younger households.[ab]

Models with short-term mortgages make it easy for households to extract cash from their house and may overstate consumption elasticities. Gorea and Midrigan (2015) consider a model in which long-term mortgages are costly to refinance. When these costs are selected to match the share of mortgages that are refinanced, consumption elasticities are substantially lower.

Reduced Form Estimates and Housing Supply Elasticities as Instruments

Are consumption elasticities for individual households large enough for house price increases to generate quantitatively big effects on consumption? Reduced form estimates of the consumption elasticity vary widely across studies. Case et al. (2005) provide reduced-form evidence on the consumption elasticity to house price changes with aggregate data from many countries. Their estimate ranges from 0.02 to 0.17. Carroll et al. (2011) also use aggregate data and estimate an immediate (next quarter) consumption elasticity of 0.02, with an eventual elasticity of 0.09. Attanasio et al. (2009) use micro data from the UK family expenditure survey to estimate consumption elasticities across households. They regress the level of consumption on changes in house price and other demographic variables. The paper obtains an average elasticity of 0.15 and higher elasticities for young households. Campbell and Cocco (2007) also use the FES and obtain a much larger average elasticity of around 1.2 and higher elasticities for older homeowners than for younger renters.[ac]

Reduced form regressions should ideally have exogenous variations in house prices on their right hand side. The identification of such variation is tricky. Even if we had a good identification strategy to isolate such exogenous variation in the data, it is not possible to directly compare the observed consumption responses in the regressions with

[ab] Kaplan and Violante (2014) show that households who invest a large fraction of their wealth in an asset with high transaction costs have high MPCs.

[ac] Lustig and Van Nieuwerburgh (2010) collect data from various US metro areas to document that risk sharing between regions is reduced when the value of housing is low. More specifically, regional consumption is more sensitive to regional income when housing collateral is scarce.

those implied by a life cycle model. The response of consumption to an exogenous increase in house prices includes any general equilibrium effects of higher house prices on consumption. For example, higher house prices may encourage more residential investment and employment, and thereby increase consumption. These GE effects are typically not included in the model.

To address the identification problem, Mian et al. (2013) use IV regressions with the local housing supply elasticity index constructed by Saiz (2010) discussed in Section 2.1 as instrument. The instrument provides a source of exogenous variation in the exposure of different geographical areas to a common aggregate house price shock. Intuitively, areas with an inelastic housing supply (such as San Francisco and Boston) should experience larger house price changes than areas with a highly elastic housing supply (such as Omaha and Kansas City). The paper estimates high consumption elasticities between 0.34 and 0.38 with Mastercard data that measures credit-card spending on nondurables and house prices from Core Logic. Kaplan et al. (2016a) confirm these estimates with store-level sales from the Kilts-Nielsen Retail Scanner Dataset and Zillow house prices.

To interpret the causality of these elasticities, the housing supply elasticity instrument has to be valid. The instrument satisfies the first-stage requirement: areas with steep slopes, bodies of water and zoning restrictions tend to experience higher house price growth during booms, so that the instrument is correlated with house prices (for example, table 2 in Saiz, 2010 and table A2 in Stroebel and Vavra, 2015). A more difficult requirement to satisfy is the second stage exclusion restriction. In a standard IV approach, the supply constraints must be uncorrelated with omitted demand factors. Saiz (2010), however, documents that supply constraints are associated with high demand. For example, land-constrained areas have higher incomes, are more creative (in the sense of more patents per capita) and attract more tourists (measured by tourist visits per person.) As skilled workers sort into more attractive areas, their productivity/income growth will increase the demand for amenities and house prices. A detailed exposition of this argument is in Davidoff (2016).[ad]

4.3 Housing Over the Business Cycle

The literature on housing over the business cycle is concerned with the facts presented in Fig. 2. As for quantities, residential investment and consumption of housing services are procyclical, with residential investment being more volatile than other investment and leading the cycle. Moreover, the price of housing is procyclical and comoves positively

[ad] Various papers try to provide more direct evidence on the exclusion restriction. For example, table 5 in Mian and Sufi (2011) shows that IRS per capita wage growth is negatively correlated with supply elasticities, indicating that more constrained areas such as San Francisco experience higher wage growth. Other measures of income growth, however, appear uncorrelated with supply elasticities. Stroebel and Vavra (2015) provide additional evidence that measures of income growth are not correlated with supply constaints.

with housing investment. At the same time, the literature aims to account at least as well, or possibly better, for standard business cycle facts such as the volatility, cyclicality, autocorrelation of GDP, nonresidential investment, consumption, hours worked, as well as wages and interest rates and possibly mortgage debt.

We divide the literature into two parts. Early work focused on frictionless representative agent models; the key difference to the standard real business cycle model (RBC) is the presence of two final goods, one of which is either housing services or a "home good" with a large housing component. Papers in this line of research differ mostly in the production technology, and the bulk of the empirical work provides new evidence on technology by sector. More recent work has emphasized simple heterogeneity of agents, usually a borrower and a lender type, as well as nominal rigidities. These papers differ mostly in the asset structure and empirical work often provides new evidence on financial variables.

Models in this section are quantified using a mix of parameters from earlier literature and new estimates. Some papers work with observable shock processes—for example, sectoral TFP—which allows them to estimate the shock process in a first step before computing an equilibrium of the model. Other papers jointly estimate parameters of preferences, technology and the shock processes using GMM or maximum likelihood approaches. In all cases, model performance is assessed by comparing the empirical distribution of a set of observables to the joint distribution of those variables implied by a stationary rational expectations equilibrium of the model economy.

While much progress has been made in understanding the role of different shocks and model ingredients, the basic facts of housing over the business cycle remain puzzling. In particular, we do not yet have a joint account of the volatility and lead-lag behavior of residential investment together with the volatility of house prices. This is in part by design: most models in this section are solved by linearization around a balanced growth path and do not allow for changes in uncertainty. As a result, asset prices move only with changes in expected cash flow or interest rates which limits the scope for volatility. A promising area for future work is to place more emphasize on mechanisms for price volatility and draw tighter connections to the micro evidence on portfolios discussed in Section 4.2.

4.3.1 Home Production

Home productions models are two sector stochastic growth models. The "market" sector produces a market good using business capital and "market labor"—identified with hours as conventionally measured. The "home" sector produces a home good using home capital—identified with housing and consumer durables—together with home labor. Market output is used to make either type of capital—we can define technology as in (21) and assume that construction capital equals home capital that directly provides utility. In line with the features of housing stressed earlier, home capital is an asset with a nontradable dividend, the return of which is difficult to measure directly. Of course, nontradability and indivisibility of housing have no bite for model properties when there is only one agent.

In a frictionless two-good model, a positive TFP shock to sector A induces reallocation of labor away from sector B and hence lower output in that sector. Hours worked and output in sector B are thus more volatile than what one would expect if there were only TFP shocks to sector B itself. The home production literature exploits this mechanism to generate more volatile market hours and output than a standard RBC model. Suppose sector B is the business or market sector, then hours worked and GDP—the series usually targeted by business cycle models—correspond to hours and output in sector B. Sector A is the household sector which produces home goods using housing capital and work at home. Home good TFP shocks can then help increase the volatility of hours and output, especially if they are imperfectly correlated with business TFP.[ae]

At the same time, sectoral reallocation gives rise to a "comovement puzzle." Indeed if sectoral TFP shocks were uncorrelated, output, labor and investment would all be negatively correlated across sectors: it makes sense to move both labor and capital toward the most productive sector. The home production literature shows that this force helps make hours and GDP more volatile. However, it also makes home and business investment negatively correlated, whereas residential and nonresidential investment are both procyclical in the data. There is therefore a tension between the promise of the mechanism for labor reallocation and its implications for capital reallocation. A second puzzle follows from the input–output structure of the models. A typical assumption is that capital for both home and business use is produced by the business sector only. Consider now the response to a perfectly correlated shock to both sectors: it makes sense to shift factors to the business sector in order to build capital before increasing investment and production in the home sector. This force make home investment lag the business cycle, again contrary to the data. If the effect is strong enough, such as when the elasticity of substitution between home capital and labor is high, we can further have negative correlation of investment across sectors even with strong positive correlation of TFP.

Progress in the literature has been to compare specifications of technology that might overcome these two puzzles. Roughly, comovement obtains more easily if shocks affect sectors similarly and there are reasons not to move capital. Greenwood and Hercowitz (1991) consider highly correlated shocks together with a low elasticity of substitution between capital and labor in the home sector. In Hornstein and Praschnik (1997), all capital is produced by a durables good sector that uses nondurables as an intermediate input. Gomme et al. (2001) assume time to build in the business sector. Chang (2000) studies capital adjustment costs. The upshot of this literature is that comovement can be obtained with technologies justified by standard input–output matrices.

[ae] See Benhabib, Rogerson and Wright (1991) and Greenwood et al. (1995) for an exposition of the main mechanism.

The lead-lag pattern of residential investment has been a tougher nut to crack. Fisher (2007) proposes a model in which home capital serves as an input into business production. The idea is that workers who live in better houses are more rested and deliver higher quality work. This type of technology cannot be justified by standard NIPA input–output accounting; it is instead motivated by a regional level estimation of a production function that takes the new effect of home capital on business output into account. With that effect and appropriate elasticities, it can make sense to build home capital first in response to productivity shocks.[af]

4.3.2 Land and House Prices

Davis and Heathcote (2005) incorporate land into a two-sector stochastic growth model. Their setup is more directly geared toward housing than the typical home production model—the "home good" is explicitly identified with housing services. A simplified version of their technology is given by (21)–(22): housing services are provided by a housing asset produced by a construction sector, and while there are no adjustment costs to capital, a limited flow of new land induces an adjustment cost to housing. These assumptions on technology have been adopted by a number of later papers. The paper itself has a richer structure with input–output links between construction and other sectors via intermediate goods derived from NIPA sectoral accounts.

The model is driven by sectoral TFP shocks and produces comovement of residential and business investment, where the former is substantially more volatile, but does not lead the cycle. At the heart of the model is the construction sector, which is labor intensive and subject to particularly volatile TFP shocks. A positive construction TFP shock is amplified by hiring and generates a lot of construction output. The response of residential investment is larger than for business investment since housing is more construction intensive and depreciates more slowly. At the same time, the input–output structure ensures that comovement still obtains, but it does not allow for residential investment to lead the cycle.

The model-implied house price is procyclical but negatively correlated with residential investment. Its volatility is less than one third of that in the data. The key effect here is that a positive construction TFP shock not only increases residential investment, but also makes housing cheaper. At the same time, TFP shocks to other sectors can make the prices of all long lived assets, including the housing asset, procyclical. Put together, these results again illustrate the promise and limitations of sectoral productivity shocks as a driving force of housing. It is tricky to come up with input–output structures that generate the right quantity dynamics. Once prices are explicitly considered, further challenges arise.

[af] Recently Kydland et al. (2012) have shown that both the lead-lag behavior of residential investment and the prevalence of long term fixed rate mortgages are special features of US data. They provide a model in which residential investment leads the cycle because the cost of housing depends on forward looking long term yields in the United States, but less so in other countries.

4.3.3 Household Debt and Nominal Rigidities

A number of papers in the early 2000s extended New Keynesian models to allow for housing and collateral constraints along the lines of Section 3.7. Early work was concerned with the response to monetary policy, described further below.[ag] To illustrate the business cycle properties of such frameworks, we focus below on the results of Iacoviello and Neri (2010) (IN). On the firm side, that paper combines nominal rigidities, the technology of Davis and Heathcote, capital adjustment costs and free linear conversion of houses. There are two types of households who differ in discount factors and no rental markets so that housing dividends are nontradable. The model features many shocks and is estimated using consumption, house prices, inflation, the nominal interest rate as well as housing and nonhousing investment, hours and wages.

Heterogeneity in discount factors gives rise to a borrower-saver household sector. Impatient borrower households borrow and run into collateral constraints, whereas patient saver households are unconstrained in equilibrium. Borrowers are always constrained near the steady state, which allows for linear solutions.[ah] The assumption of linear conversion implies that there is a per-unit price of housing. Nontradability of dividends nevertheless allows for a steady state in which both types of households own housing. This makes the model different from linear two-agent models of equity pricing, in which the agent with the highest valuation of a tradable asset is typically the only owner.

Three key features distinguish New Keynesian borrower-saver models from the models discussed so far. First, a "housing preference shock" increases the felicity from housing. Together with the shock to construction productivity, it is the most important driver of house prices and residential investment. Since it increases housing demand rather than supply, it also makes those two variables move together. At the same time, it lowers comovement of business and residential investment. This tension implies that the model has trouble matching jointly the volatility and cyclicality of house prices and investment, as well as the lead-lag behavior of investment, even though it does generate volatile house prices (IN table 4).

Second, the models feature nominal rigidities which amplify "demand" shocks such as those to housing preference. In particular, with sticky wages, housing preference shocks have much larger effects on residential investment (IN fig. 2). Stickiness of prices is less

[ag] A related early borrower-saver business cycle model is Campbell and Hercowitz (2005) who study borrowing collateralized by durables with a constant price. They show that lower downpayment requirements allow borrower households to better smooth labor supply and hence lower the volatility of aggregate hours. They use this effect to relate financial innovation in the early 1980s to the Great Moderation of the US economy.

[ah] Linear solution imply symmetric business cycles. Allowing for occasionally binding constraints could alter dynamics by introducing nonlinear dynamics; for example, the response to shocks could be stronger in downturn when constraints bind, giving rises to asymmetric cycles as in the data. Guerrieri and Iacoviello (2015) develop a model to study such effects.

relevant since house prices are flexible. At the same time, however, sticky wages worsen the comovement problem: a construction productivity shock no longer increases business investment (IN fig. 4). Nominal rigidities also matter in that they allow other shocks—eg, to monetary policy and markups—to feed through to the housing sector.

The heterogeneity of households is not particularly important for the behavior of investment and house prices (IN figs. 2–4). Linear conversion of housing matters here: housing (as well as other capital) satisfies the Euler equation of the patient unconstrained investor, so investment and price dynamics look much like in a representative agent model. At the same time, the presence of collateral constrained households matters for the response of consumption to shocks. In particular, changes in housing wealth will affect aggregate consumption more. In a model with nominal rigidities, this also translates into effects on output.

4.3.4 Financial Frictions in the Business Sector

With financial frictions in the household sector, house prices can matter for output through their effects on demand. If businesses face collateral constraints, then real estate values can also affect firms' cost. For example, in Iacoviello (2005), entrepreneurs borrow using housing as collateral. Liu et al. (2013) estimate a model in which firms borrow against land as collateral. Housing or land preference shocks can then serve as a driver of the business cycle together with the price of land and the level of business debt.

In the wake of the financial crisis, it has become common to introduce shocks that directly change borrowing or intermediation costs. Some papers have studied such shocks in models with housing. For example, Dorofeenko et al. (2014) add financing constraints as well as risk shocks to the construction sector. Risk shocks then increase the volatility of house prices although this comes at the cost of overstating the volatility of residential investment. Gerali et al. (2010) estimate a model of the Euro Area. Their estimation backs out an important role for shocks to a frictional banking sector that lends to households and firms against collateral.

4.3.5 Effects of Monetary Policy

A growing literature studies the effect of monetary policy shocks in New Keynesian models with heterogenous households, following Aoki et al. (2004) and Iacoviello (2005). The goal is to match the impulse response to a change in the short term nominal interest rate obtained from structural VARs. A stylized fact is that an expansionary monetary policy shock—a decline in the short rate—increases house prices and residential investment along with output (see, for example, Calza et al. (2013) for evidence for a cross section of countries). The goal of the literature is to account for this fact as well as to show whether the presence of heterogenous agents and housing is an important force behind impulse responses for other variables.

As a benchmark, consider the response to a monetary policy shock in a New Keynesian model with a representative agent. With sticky prices, a decline in the nominal short rate generates a decline in the real short rate. From the Euler equation, the representative agent would like to substitute away from expensive future consumption and increase current consumption. Since firms are on their labor demand curve, hiring and output increase to provide extra consumption—monetary policy stimulates the economy. The Euler equation also says that the return on housing should decline—this can happen either through a drop in the dividend (an increase in the relative consumption of housing) or a drop in house prices. Finally, the return on investment declines and so does residential investment.

Suppose now instead that the short rate declines in model with heterogeneous agents and collateral constraints. Assume also that housing is priced linearly. A change in the real rate directly affects the Euler equation of unconstrained agents. Again the return on housing has to decline as well and this happens in part via an increase in housing consumption by the unconstrained—which decreases dividends—and in part through a drop in the house price. The quantity adjustment is not very large, so that the price response typically looks similar to a representative agent model. A key difference to the representative agent model is that the price change now tightens the collateral constraint and lowers consumption of constrained agents. As a result, the consumption and output responses are typically much larger than in a representative agent model, and they are driven to a much smaller extent by intertemporal substitution.

Iacoviello (2005) studies the above mechanism in a two-agent model with borrower and saver households. Aoki et al. (2004) consider savers and an entrepreneurial housing sector. Monacelli (2009) compares the implications of models with and without collateral constraints with evidence on the consumption response for durables and nondurables. Rubio (2011) introduces long term debt in a model without capital and shows that effects of monetary policy are stronger with variable rate mortgages, since real interest rate movements have larger effects. Calza et al. (2013) present SVARs evidence that monetary policy has larger effects in countries with more variable mortgages; as well as a model with capital that generates qualitatively similar effects. Garriga et al. (2013) consider a flexible price model and emphasize that variable-rate mortgages generate important nominal rigidities in their own right.

4.3.6 Rich Household Sector

Much of the literature on business cycles and monetary policy has built on traditional macro models with limited heterogeneity. In light of results on portfolio choice discussed earlier, there is considerable promise in models that allow for richer heterogeneity in both households and houses as well as for aggregate risk. Early work in this direction abstracted from house price risk. Silos (2007) studies a model with two capital stocks that also accounts for the cross section and time series properties of housing and wealth positions.

Iacoviello and Pavan (2013) allow for a rental market and emphasize the procyclicality of debt. Another interesting direction is to explicitly incorporate geography (for example, Van Nieuwerburgh and Weill (2010)).

The literature on monetary policy shocks has also been moving toward models with a richer household sector. For example, Kaplan et al. (2016b) consider a perpetual youth model with borrowing constraints and a subset of illiquid assets (including housing). Wong (2016) considers an overlapping generations model with long term mortgages and highlights the role of heterogeneity by age for the transmission mechanism. All of these papers show that the details of how the household sector is modeled matter for the strength of impulse responses.

4.4 Asset Pricing with Housing

This section summarizes work that studies regular patterns in asset prices implied by models with housing. Similarly to the business cycle analysis in the previous section, model exercises compare empirical distributions in the data to stationary equilibria implied by the model. The key difference is that explicit nonlinear solutions allow for time variation in risk implied by the role of housing as a consumption good or collateral asset. Changes in the risk return tradeoff then affect the pricing of all assets including housing.

The upshot from this literature is that the presence of housing introduces slow movement in the stochastic discount factor that lines up with observed movements in risk premia. At the same time, rational expectations versions of the models here do not generate sufficient volatility to price risky assets, unless risk aversion is large. It is an open question how much the channels described here can contribute once they are combined with less restrictive assumptions on expectations.

4.4.1 Housing as a Consumption Good

The standard consumption-based asset pricing model focuses on consumption risk: the value of an asset depends on the comovement of its return with a single factor, aggregate consumption growth. In a model with housing, households worry not only about future consumption growth, but also about the future composition of the consumption bundle (c_t, s_t). With frictionless rental markets, composition risk can be measured by the expenditure share of housing in the overall consumption bundle. Assets are then valued also for whether they provide a hedge against this second risk factor.

More formally, Piazzesi et al. (2007) assume a power utility function $U(C) = C^{1-1/\sigma}/(1-1/\sigma)$ over the CES aggregator $C = g(c, s)$, and assume a frictionless rental market. The pricing kernel (10) can then be written as

$$M_{t+1} = \beta \left(\frac{c_{t+1}}{c_t} \right)^{-1/\sigma} \left(\frac{1-x_{t+1}}{1-x_t} \right)^{\frac{\varepsilon-\sigma}{\sigma(\varepsilon-1)}}. \tag{34}$$

where x_t is the expenditure share on housing consumption. If $\sigma < 1 < \varepsilon$, then households worry both generally about recessions—low consumption growth—and in particular about "severe recessions" in which the expenditure share on housing consumption is low.

The pricing kernel (34) is observable since the housing expenditure share x_t is available in the NIPA tables. Use of expenditure shares avoids reliance on problematic measures of housing quantities s_t. Fig. 1 shows movements in x_t over the postwar period. The key feature is that the expenditure share contains a slow moving component that lines up with the low frequency component in the price dividend ratio on equity: both series are high in the 1960s, low in the 1970s and recover in the 1980s. These movements are predictable and occur at frequencies that are much lower than business cycle frequencies.

Low frequency movements in the housing share induce movements in stock prices that are in line with the data. For example, agents' concern with severe recessions increases risk premia on assets that pay off little when the expenditure share on housing drops, and more so when the housing share is already currently low. Comovement of x_t with the price dividend ratio implies that equity is such an asset. Since the expenditure share is stationary, the price-dividend ratio on stocks is persistent but mean reverting. This mean reversion explains why the price-dividend ratio forecasts excess returns on stocks. The model also implies that the housing expenditure share should predict excess stock returns, which it does in the data.

Since movements in the housing share are small, large risk premia obtain only if the exponent in (34) is sufficiently large. This can happen for two reasons. On the one hand, suppose that the intratemporal elasticity of consumption is close to one. Since household desire constant expenditure shares, the prospect of a drop in the housing share causes them large discomfort and requires high risk premia on equity even when risk aversion is low. At the same time, however, the intratemporal Euler equation (24) implies very large volatility in rents. On the other hand, high risk premia obtain without high rent volatility when risk aversion is large. In this case, the role of housing is still important in generating time variation in risk premia.

4.4.2 Adjustment Costs and Production

The same asset-pricing implications continue to hold when housing is costly to adjust (see, for example, Stokey, 2009). Adjustment costs typically alter the optimal consumption allocation—for example, consumption is constant or depreciates at a constant rate as long as there is no adjustment. At the same time, the Euler equation (8) for other securities still holds. The pricing kernel (34) continues to be observable with quantity data on housing and nonhousing consumption, or with data on nonhousing consumption and the expenditure share x_t. The argument further extends to setups with preferences over consumption and housing that deviate from expected utility. In this case, the pricing kernel has to be evaluated with continuation utility (10).

Flavin and Nakagawa (2008) measure the pricing kernel with quantity data on housing consumption. More specifically, they use square footage to measure housing consumption. An important disadvantage of this quantity-based measure is that it does not capture quality differences that would be reflected in dollar expenditures and therefore the expenditure share x_t as in (34). For example, a 2000 square foot house with a view will provide more utility than the same square footage without view. This quality difference will be reflected in a higher rent for the house with a view and an associated higher expenditure share x_t on housing.

Jaccard (2011) studies a two-sector model with production of housing. There is habit formation over the consumption bundle $g(c_t, s_t)$ and leisure. The presence of adjustment costs in housing production together with habit formation helps to generate volatile house prices. Habit formation also helps to generate a sizable equity premium as well as comovement between hours worked and output.

4.4.3 Housing as a Collateral Asset

Lustig and Van Nieuwerburgh (2005) consider the asset pricing implications of a heterogenous agents model with uninsurable idiosyncratic income risk and collateral constraints. The collateral constraint is similar to (28), except that the set of securities is a complete set of one-period-ahead contingent claims and any state contingent promise must be backed by the value of housing in the relevant state of nature next period. The presence of contingent claims allows an aggregation result that expresses the pricing kernel M_{t+1} as an aggregate consumption term as in (34) multiplied by a term that depends on the "housing collateral ratio," that is, the ratio of housing wealth relative to human wealth.

The housing collateral ratio now serves as a second state variable that describes variation in investors' required compensation for an additional risk factor. Indeed, investors perceive recessions as particularly severe when the housing collateral ratio is low and collateral constraints are more likely to bind. Moreover, if the current collateral ratio is already low, opportunities for smoothing uninsurable income shocks through collateralized borrowing are poor and required risk premia are high. Empirically, measures of the housing collateral ratio predict stock returns and also help explain the cross section of stock returns.

The paper further shows that large movements in risk premia are associated with large movements in the riskless interest rate. Indeed, if opportunities to borrow are currently low, then the supply of all contingent claims falls, which drives up the prices of all claims as well as their sum, the price of a riskless bond. This logic is not limited to models with collateral constraints on contingent claims but also applies with the constraint (28) or when default is punished by autarky. Quantitatively, it prevents rational expectations models with borrowing constraints from generating high and volatile equity premia without excessively volatile interest rates, unless risk aversion is high.

4.5 Housing Boom–Bust Cycles

A growing body of work tries to understand the mechanisms behind large house price swings and their quantitative importance. Two boom–bust episodes stand out—they were associated with high nationwide house prices both in the United States as well as many other industrialized countries. The first occurred during the Great Inflation of the 1970s, as documented in fig. 2 of Piazzesi and Schneider (2008). During the boom, houses in high quality segments of the US housing market appreciated by 11% *more* than low quality segments, as shown in table 2 of Poterba (1991). The second boom happened during the 2000s, when many countries experienced large increases in mortgage debt together with large house prices increases, as documented by Tsatsaronis and Zhu (2014). During this boom, houses in high quality segments of US housing markets appreciated *less* than low quality segments, as discussed in Section 2.2.

The typical account of a boom episode consists of one or more "shocks," that is, changes in the economic environment, together with a mechanism for how the economy responds to the shocks. Broadly, candidate shocks are changes to macroeconomic conditions that affect income and assets other than housing, changes in financial conditions that affect the ability to borrow given house prices, as well as changes in government policy and expectations about future house values. How exactly the shocks and the mechanism are modeled depends on how much of the response of the economy is endogenous in a given model exercise.

4.5.1 Overview of the Results

Studies of the 1970s housing boom show that the Great Inflation depressed user costs, especially for richer households. The lower user costs can quantitatively account for both higher overall house values as well as higher house values in high quality segments. Higher mortgage interest-rate tax deductions increased the attractiveness of homeownership. They can explain a large share of the overall increase in real house prices. The higher deductions especially benefitted households in higher tax brackets, which accounts for the higher appreciation of high quality segments. In surveys, young households reported to have higher inflation expectations and thereby lower perceived real rates than older households. This disagreement about inflation expectations and real rates across generations is consistent with the increase in credit during this episode. As a consequence, young households borrowed more at rates that they perceived as low and bought houses, pushing up prices.

User costs were again low during the 2000s boom. Credit was easy to get—with low interest rates and relaxed downpayment constraints, enabled partly by an inflow of foreign savings as well as an increase in securitization. The lower interest rates raised the present value of future housing services and thus house values across the board. The relaxation of downpayment constraints mattered mostly for poor households who were able to borrow more and buy houses, driving up house prices especially in the low quality segments of the housing market. Richer households increasingly bought low quality houses and neighborhoods gentrified, further pushing up prices in these low quality segments. All studies, however, find it difficult to quantitatively account for the entire increase in house prices. This suggests that expectations played a role during the 2000s boom. As long as households were expecting house prices to grow at trend together with income (instead of mean reverting to lower levels), it is possible to quantitatively explain the boom.

There has been much progress in our understanding of boom–bust episodes. Micro data—including on household behavior and survey expectations—have helped sort out the importance of competing mechanisms. At the same time, the nature of the shock that started the housing boom is yet not well understood. Changes in housing preferences, expectations, foreign capital inflows or downpayment constraints are essentially standins for changes in various market participants' attitudes toward housing and housing credit. To understand what generates these changes requires theories of expectation formation, financial innovation as well as international capital market integration.

Another open question is the precise role of the US government during the recent boom. It is clear that many policies (for example, associated with the 1994 National Homeownership Strategy developed by the US Department of Housing and Urban Development) encouraged poor households to take out large mortgages and buy houses, especially in low quality segments. How much did these policies contribute to the boom? A related question is whether the government should promote homeownership in the first place, given that it involves a large undiversified investment and potential welfare costs in default.

4.5.2 The 1970s Boom

Poterba (1984) investigates the user cost equation with Census data from the 1970s housing boom. His findings show that high expected inflation substantially lowered the user costs of housing. High expected inflation pushes up mortgage rates and thereby increases the mortgage tax subsidy. This mechanism is able to explain a 30% increase in real house prices during the 1970s.

Poterba (1991) calculates that user costs dropped especially for households in high tax brackets. The reason is that high mortgage rates translate into a larger mortgage tax subsidy for households who earn high incomes that are taxed at higher rates. Lower user costs for richer households increase the demand for more expensive houses. Tables 3 and 4 in the paper indeed find higher capital gains for more expensive houses during the 1970s boom, while cheaper houses appreciated less percentage-wise.

Piazzesi and Schneider (2009a) study an equilibrium model with three assets—houses, stocks, and nominal bonds. Households solve life cycle consumption–portfolio choice problems with an exogenous nontradable income process (33). The paper computes temporary equilibria as described in Section 3.4 in this model. The benchmark household beliefs about future returns and income dynamics are estimated with historical data. Moreover, these dynamics feature a large idiosyncratic component in house price volatility. When the model is evaluated during the 1970s, the temporary equilibrium concept is useful for exploring the implications of higher expected inflation as well as higher inflation volatility.

When evaluated with 1990s data on income and asset endowments, the temporary equilibrium of this model is successful at matching observed asset prices as well as life cycle patterns in wealth and portfolio weights on houses, stocks and nominal bonds. In particular, the model predicts that young households borrow to buy a house and do not participate in the stock market. As they get older, households pay down their mortgage and start saving in nominal bonds and stocks.

The model is evaluated with endowment data from the Survey of Consumer Finances in the 1960s, 1970s, and 1990s. The model predicts a 25% dip in aggregate wealth during the 1970s—which is exactly the pattern we observe in the data. There are three separate mechanisms that contribute to the drop in household wealth during the Great Inflation in the model. First, the 1970s experienced a demographic shift toward more young households—the Baby Boomers—who have lower savings rates. Second, capital losses from realized inflation lowered wealth and hence savings, especially for older households. Third, lower savings were not counteracted by a large increase in interest rates, because the outside supply of bonds to the household sector also fell.

While aggregate household wealth dropped, the portfolio composition looks similar across all three periods at benchmark beliefs which do not take into account higher expected inflation rates during the 1970s. When all households believe in the high expected inflation rate from the 1970 consensus forecasts in the Michigan survey, they increase their portfolio away from stocks toward housing. This shift happens because high expected inflation generates tax effects that favor housing investments: the returns on housing are essentially untaxed, while mortgage interest rate payments are tax deductable. Disagreement about inflation expectations shifts the portfolio further toward housing. The reason is that young households expect high inflation and perceive a low real interest rate. They therefore borrow more and buy housing. The two inflation mechanisms—higher mean inflation and disagreement across cohorts—explain roughly half of the portfolio shift toward housing observed in the data.[ai] The remaining shift can be attributed to lower stock return expectations in times with high expected inflation, which lead to a large decline in price–dividend ratios for stocks while the

[ai] Relatedly, Piazzesi and Schneider (2008) study a model in which some households suffer from inflation illusion. These households confuse changes in the nominal interest rates with changes in real interest rates, while smart households understand the Fisher equation. The model predicts a nonmonotonic relationship between the price–rent ratio and nominal interest rates: house prices are high when nominal rates are either particularly high (as in the 1970s) or low (as in the 2000s).

housing market was booming. The resulting negative comovement of house and stock prices is an important step toward our understanding of the 1970s.

4.5.3 The 2000s Boom

We divide studies of the 2000s boom into three groups by the type of exercise they undertake. One set of papers evaluates versions of the user cost equation (27): it asks whether reasonable scenarios for interest rates and housing payoff expectations—as well as other parameters of the user cost equation such as taxes—are consistent with high house prices. Second, studies that employ small open economy models take securities prices—in particular interest rates—from the data and endogenously determine only house prices and allocations. Finally, papers that work with closed economy models jointly determine house prices and the prices of other assets.

As usual these three approaches are complementary. Indeed, user cost studies (as well as more generally studies of Euler equations) or small open economy models do not explain why interest rates move. At the same time, they evaluate a given model mechanism without taking a stand on the explicit shock structure as well as the details of who participates in securities markets. They thus generate conclusions that are robust to those details. While a closed economy exercise is in principle more ambitious as it makes those details explicit and takes a stand on the nature of the shocks, it is also more prone to misspecification.

User Cost Calculations

Himmelberg et al. (2005) study user costs leading up to the recent housing boom. Their approach assumes that future payoffs can be discounted at the current long-term interest rate. They conclude that the large decline in long rates during the early 2000s can explain the house price boom during that time. Glaeser et al. (2013) show that discounting all future payoffs at the low 2000s long rate is crucial for this quantitative result. In an environment in which low current rates are allowed to mean revert in the future, the magnitude of the boom is significantly reduced. They conclude that optimistic expectations played an important role in the housing boom.

House Prices in Small Open Economies

Kiyotaki et al. (2011) study a small open economy in which households solve life cycle problems, choose between renting and owning, and face collateral constraints as in Section 3.7. In the model, housing is a capital stock that is produced with land and capital. The paper compares steady states with looser collateral constraints: downpayment constraints that range from 10% to 100%. The findings in their table 3 show that varying the downpayment constraint has large quantitative effects on the homeownership rate: while only 46% of households own a home when it is not possible to borrow against housing, 90% of households own a house when the downpayment constraint is 10%.

Despite their large effects on extensive margins, Kiyotaki et al. show that lower downpayment constraints have negligible effects on house prices. The price–rent ratio is essentially constant across all steady states in table 3. This outcome is intuitive, because all homeowners in the model are marginal investors and determine the per unit price of housing. With looser collateral constraints, there is an inflow of new home buyers. However, these new buyers do not affect the per unit price of housing because the Euler equations also hold for rich households. Their table 4 studies how these results quantitatively depend on the scarcity of land for the production of housing. Kermani (2012) studies these mechanisms in a continuous-time model with a representative agent.

Sommer et al. (2013) solve a similar model but without production. Instead, the overall housing supply is fixed and there is free conversion of housing units. The paper also finds small quantitative effects of looser collateral constraints and lower interest rates, and considers higher income expectations. Chu (2014) assumes that the supply of rental housing and the supply of owner-occupied housing are fixed separately and cannot be converted into each other. As a result, looser collateral constraints have larger effects on house prices. In particular, the value of owner-occupied housing appreciates more than rental housing. The bond market in the model clears; income shocks are assumed to be more volatile in 2005 which keeps the equilibrium interest rate constant over time (instead of matching the lower interest rates observed in the data.)

Landvoigt et al. (2015) study an assignment model with indivisible and illiquid houses in a metro area. The housing demand of movers is derived from a life cycle consumption and portfolio choice problem with transaction costs and collateral constraints. As in Section 3.4, house prices are determined in a temporary equilibrium to induce households with lower demand for housing services to move into lower quality houses. The distribution of equilibrium prices thus depends on the distribution of mover characteristics as well as the distribution of house qualities. While the market for all house qualities clears, the metro area is a small open economy.

Landvoigt (2015) measure continuous distributions for movers and house qualities with micro data from San Diego County for 2 years: 2000 and 2005, the peak of the boom. The distribution of mover characteristics—age, income, and wealth—is measured with data from the American Community Survey. The 2005 distribution shows that movers were richer than in 2000. Moreover, the 2005 house-quality distribution has fatter tails than in 2000—relatively more houses traded at the low and the high end of the quality spectrum than in intermediate ranges.

To measure the distribution of house qualities, Landvoigt (2015) assume that house quality is a one-dimensional index. Therefore, house quality can be measured by price in the base year 2000. The paper documents that 2005 house prices are strictly increasing in 2000 prices. This monotonicity implies that for every 2005 quality level, there is a unique initial 2000 quality level so that the average house of that initial quality resembled the given house in 2005. The 2000 distribution of house qualities is simply the distribution of transaction prices in that year. The 2005 distribution of house qualities can be constructed from the 2005 distribution of transaction prices using the monotonicity of the map from 2000 qualities to 2005 prices.[aj]

The paper compares the predictions for equilibrium prices in both years and derives the cross section of capital gains by quality. These predictions are compared to capital gains by quality in the data. Two key mechanisms allow the model to quantitatively match the observed cross section of capital gains by quality from 2000 to 2005. The first mechanism is cheaper credit: looser collateral constraints and lower mortgage rates in 2005 allowed poorer households to borrow more and increase their demand for housing, especially at the low end of the quality spectrum. Richer households were not affected much by lower downpayment constraints. But these richer households are not marginal investors for low quality houses (as discussed in Section 3.3.1). Therefore, the higher housing demand by poor households translates into higher prices of low quality houses.

The second mechanism is that more low quality houses transacted in 2005. When the distribution of movers is assigned to the distribution of houses in 2005, the marginal buyer of a low quality house is richer compared to 2000 and pushes up low-end prices. Both mechanisms generate capital gains that monotonically decline in house quality.

Whether or not the model quantitatively matches capital gains by quality depends on expectations. An advantage of temporary equilibria is that we can find out how much expectations matter. Under the assumption that households in 2005 were expecting house prices to continue to grow at the same rate as labor income and easy credit conditions to remain, the model implies the same capital gains from 2000 to 2005 as in the data. Under the assumption that households in 2005 foresee that future house prices and credit conditions will return to their 2000 values, capital gains as a function of house quality are shifted down until expensive houses do not appreciate in value. Our conclusion is that easy credit and fatter tails in the house quality distribution predict a monotonically declining pattern in capital gains. To quantitatively explain capital gains, expectations are important. In particular, expectations in 2005 cannot be pessimistic about the future.

Closed Economy and the Determination of Interest Rates

The closed economy models considered in the literature all assume costless conversion and thus linear pricing, which by design reduces the quantitative importance of looser collateral constraints on house prices. The per-unit price of housing enters the Euler equations of all investors, which includes rich investors for whom collateral constraints do not matter. Therefore, any change in collateral constraints will have small effects on

[aj] Epple et al. (2015) also assume house quality is a one-dimensional index. They estimate house prices as well as rental values as nonlinear functions of quality with data for various metro areas using a new structural estimation approach.

per-unit house prices. A key contribution of these models is to make the point that looser collateral constraints tend to push up equilibrium interest rates, so that a major force is needed to keep rates low during the boom.

Garriga et al. (2012) study a closed economy with production in two sectors without aggregate shocks and a representative household. The production of housing involves land and irreversible investment in structures. Foreign lenders determine mortgage rates. The collateral constraint is selected to match aggregate mortgage debt to housing wealth. Under the assumption of perfect foresight about looser collateral constraints and low mortgage rates in the future, the model is able to explain roughly half of the observed increase in national house price indices. Since housing and nonhousing consumption are strong complements, higher house prices do not lead to a large consumption increase (which would be counterfactual). The paper attributes the other half of the increase to expectations.

Favilukis et al. (2016) study a closed economy with households who solve life cycle problems with uninsurable labor income shocks as well as aggregate shocks. There is no rental market, so households have to buy in order to consume housing services. The paper solves the model under the assumption that foreigners bought more bonds during the 2000s and thereby increase the mortgage supply. The paper carefully measures the size of these bond purchases and quantifies their effect on equilibrium mortgage rates. Looser collateral constraints lower risk premia and thereby increase house prices by roughly half of the observed increase in national house prices.

The asset "house" in Favilukis et al. is a claim to the national housing stock. As discussed in Section 2.2, households in the data hold individual houses instead of such diversified claims. In fact, a diversified claim has much more attractive return properties than individual houses because national house price indices are not volatile. In panel B of table 5 of Favilukis et al., the equilibrium Sharpe ratio of the national housing stock is an impressive 0.82 compared a less attractive 0.37 Sharpe ratio for stocks. To better capture the Sharpe ratios of individual houses that households face in the data, the paper considers small idiosyncratic shocks to housing depreciation. These idiosyncratic shocks increase precautionary savings and thereby depress the equilibrium risk-free rate, but they do not match the high idiosyncratic component in the variance decomposition of individual housing returns.

Justiniano et al. (2015a) consider a closed economy in which patient households lend to impatient households until their lending reaches an exogenous supply limit. There are collateral constraints, so impatient households borrow to buy houses. The paper shows that looser collateral constraints increase the demand for houses and mortgages by the impatient households. As a consequence, both house prices and mortgage rates increase in equilibrium—contrary to what we saw in the data, where mortgage rates fell. The paper then argues that an exogenous increase in the credit supply limit increases borrowing while keeping mortgage rates low in equilibrium. It is important for this argument to assume that patient household have a fixed housing demand or buy houses in a different segment from impatient households, otherwise their Euler equation would determine house prices and thereby keep house prices low. Justiniano et al. (2015b) add poorer (subprime) borrowers to the model. They show that subprime borrowers increase their mortgage borrowing more than richer borrowers in response to an increase in the credit supply.

Landvoigt (2015) endogenizes the supply of mortgages in a closed-economy model with banks and aggregate risk. Households differ in their patience as well as their risk aversion. Banks issue deposits and equity to make mortgages. Looser collateral constraints increase the demand for housing and mortgage borrowing, but increase mortgage rates—contrary to what we observed in the data. Landvoigt introduces securitization, which allows banks to sell mortgages directly to risk-tolerant savers. The boom is initiated when banks underestimate the riskiness of new borrowers during the early 2000s and collateral constraints are relaxed. As banks securitize their mortgages and sell them as MBS to savers, risk premia decline, the supply of lending increases, and the model generates a boom–bust in house prices.

Expectations

The broad conclusion from existing studies of the 2000 boom is that expectations played a quantitatively important role. This conclusion is consistent with survey expectations about future house prices. For example, table 9 in Case and Shiller (2003) reports that homebuyers in 2003 were expecting house prices to appreciate between 9% and 15% each year over the next decade. Piazzesi and Schneider (2009b) document that at the peak of the recent boom, the fraction of households who believed that house prices would continue to increase doubled.

Recent research has started to capture such house price expectations. Piazzesi and Schneider (2009b) show that since only a small fraction of houses trade every year, a few exuberant households are enough to push up equilibrium house prices in these transactions. Barlevy and Fisher (2011) assume that a stream of new households enters every period with a certain probability and the stream may stop. Burnside et al. (2011) assume infectious-disease dynamics for expectations. Adam et al. (2012) study learning dynamics that temporarily decouple house prices from fundamentals.

Landvoigt (2016) uses micro data to estimate beliefs that rationalize the consumption–portfolio decisions of households in the SCF. The estimation finds that an important feature of beliefs is higher uncertainty about future house prices which increases the option value of default and thereby leverage during the housing boom.

ACKNOWLEDGMENTS

We thank Alina Arefeva, Eran Hoffmann, Amir Kermani, Moritz Lenel, Sean Myers, Alessandra Peter, John Taylor, Harald Uhlig, and conference participants at Stanford for comments.

REFERENCES

Adam, K., Kuang, P., Marcet, A., 2012. House price booms and the current account. NBER Macroeconomics Annual, vol. 26(1), University of Chicago Press, Acemoglu and Woodford, pp. 77–122.

Adelino, M., Schoar, A., Severino, F., 2015. Loan originations and defaults in the mortgage crisis: the role of the middle class. NBER Working Paper No. 20848.

Agarwal, S., Driscoll, J., Laibson, D., 2013. Optimal mortgage refinancing: a closed form solution. J. Money Credit Bank. 45, 591–622.

Amromin, G., Huang, J., Sialm, C., Zhong, E., 2013. Complex mortgages. Working Paper, University of Texas at Austin.

Anderson, S., Campbell, J.Y., Nielsen, K.M., Ramadorai, T., 2015. Inattention and inertia in household finance: evidence from the danish mortgage market. Working Paper, Harvard.

Andrews, D., Sánchez, A.C., Johansson, Å., 2011. Housing markets and structural policies in OECD countries. OECD Economics Department Working Papers.

Aoki, K., Proudman, J., Vlieghe, G., 2004. House prices, consumption, and monetary policy: a financial accelerator approach. J. Financ. Intermed. 13, 414–435.

Artle, R., Varaiya, P., 1978. Life cycle consumption and homeownership. J. Econ. Theory 18, 38–58.

Attanasio, O.P., Weber, G., 1994. The UK consumption boom of the late 1980s: aggregate implications of microeconomic evidence. Econ. J. 104 (427), 1269–1302.

Attanasio, O.P., Blow, L., Hamilton, R., Leicester, A., 2009. Booms and busts: consumption, house prices and expectations. Economica 76, 20–50.

Attanasio, O.P., Leicester, A., Wakefield, M., 2011. Do house prices drive consumption growth? The coincident cycles of house prices and consumption in the u.k. J. Eur. Econ. Assoc. 9 (3), 399–435.

Attanasio, O.P., Bottazzi, R., Low, H.W., Nesheim, L., Wakefield, M., 2012. Modelling the demand for housing over the life cycle. Rev. Econ. Dyn. 15, 1–18.

Bachmann, R., Cooper, D., 2014. The ins and outs in the U.S. housing market. Working Paper, Notre Dame.

Badarinza, C., Campbell, J.Y., Ramadorai, T., 2016. International comparative household finance. Annu. Rev. Econ. Forthcoming.

Barlevy, G., Fisher, J.D., 2011. Mortgage choices and housing speculation. Working Paper, Federal Reserve Bank of Chicago.

Benhabib, J., Rogerson, R., Wright, R., 1991. Homework in macroeconomics: household production and aggregate fluctuations. J. Polit. Econ. 99, 1166–1187.

Berger, D., Veronica, G., Guido, L., Joseph, V., 2015. House prices and consumer spending. NBER Working Paper 21667.

Berkovec, J., Fullerton, D., 1992. A general equilibrium model of housing, taxes, and portfolio choice. J. Polit. Econ 100 (2), 390–429.

Braid, R., 1981. The short-run comparative statics of a rental housing market. J. Urban Econ. 10, 280–310.

Braid, R., 1984. The effects of government housing policies in a vintage filtering model. J. Urban Econ. 16, 272–296.

Burnside, C., Eichenbaum, M., Rebelo, S., 2011. Understanding booms and busts in housing markets. J. Polit. Econ. Forthcoming.

Calza, A., Monacelli, T., Stracca, L., 2013. Housing finance and monetary policy. J. Eur. Econ. Assoc. 11 (1), 101–122.

Campbell, J.Y., Cocco, J.F., 2003. Household risk management and optimal mortgage choice. Q. J. Econ. 118 (4), 1449–1494.

Campbell, J.R., Hercowitz, Z., 2005. The role of collateralized household debt in macroeconomic stabilization. NBER Working Paper No. 11330.

Campbell, J.Y., Cocco, J.F., 2007. How do house prices affect consumption? Evidence from micro data. J. Monet. Econ. 54 (3), 591–621.

Campbell, S., Davis, M., Gallin, J., Martin, R.F., 2009. What moves housing markets: a variance decomposition of the rent-price ratio. J. Urban Econ. 66 (2), 90–102.

Caplin, A., Leahy, J.V., 2014. A graph theoretic approach to markets for indivisible goods. J. Math. Econ. 52, 112–122.

Cardarelli, R., Igan, D., Rebucci, A., 2008. The Changing Housing Cycle and Its Implications for Monetary Policy. IMF World Economic Outlook. International Monetary Fund, Washington, DC.

Carroll, C.D., 1997. Buffer-stock saving and the life-cycle/permanet income hypothesis. Q. J. Econ. 112, 1–55.

Carroll, C.D., Otsuka, M., Slacalek, J., 2011. How large are housing and financial wealth effects? A new approach. J. Money Credit Bank. 1 (43), 55–79.

Case, K.E., Shiller, R.J., 1989. The efficiency of the market for single-family homes. Am. Econ. Rev. 79 (1), 125–137.

Case, K.E., Shiller, R.J., 1990. Forecasting prices and exess returns in the housing market. Real Estate Econ. 18 (3), 253–273.

Case, K.E., Shiller, R.J., 2003. Is there a bubble in the housing market? Brook. Pap. Econ. Act. 2, 299–362.

Case, K.E., Quigley, J., Shiller, R.J., 2005. Comparing wealth effects: the stock market versus the housing market. In: Advances in Macroeconomics, vol. 5(1), Berkeley Electronic Press, pp. 1–34. http://dx.doi.org/10.2202/1534-6013.1235.

Chambers, M.S., Garriga, C., Schlagenhauf, D.E., 2009a. The loan structure and housing tenure decisions in an equilibrium model of mortgage choice. Rev. Econ. Dyn. 12, 444–468.

Chambers, M.S., Garriga, C., Schlagenhauf, D.E., 2009b. Accounting for changes in the homeownership rate. Int. Econ. Rev. 50 (3), 677–726.

Chang, Y., 2000. Comovement, excess volatility and home production. J. Monet. Econ. 46, 385–396.

Chen, H., Michaux, M., Roussanov, N., 2013. Houses as ATMs? Mortgage refinancing and macroeconomic uncertainty. NBER Working Paper No. 19421.

Chien, Y., Lustig, H., 2009. The market price of aggregate risk and the wealth distribution. Rev. Financ. Stud. 23 (4), 1596–1650.

Chu, Y., 2014. Credit constraints, inelastic supply, and the housing boom. Rev. Econ. Dyn. 17, 52–69.

Cocco, J.F., 2005. Portfolio choice in the presence of housing. Rev. Financ. Stud. 18, 535–567.

Cocco, J.F., 2013. Evidence on the benefits of alternative mortgage products. J. Financ. 68 (4), 1663–1690.

Cocco, J.F., Gomes, F.J., Maenhout, P.J., 2005. Consumption and portfolio choice over the life cycle. Rev. Financ. Stud. 18 (2), 491–533.

Cochrane, J.H., 2011. Discount rates. J. Financ. 66 (4), 1047–1108.

Cutler, D.M., Poterba, J.M., Summers, L.H., 1991. Speculative dynamics. Rev. Econ. Stud. 58, 529–546.

Davidoff, T., 2006. Labor income, housing prices, and homeownership. J. Urban Econ. 59, 209–235.

Davidoff, T., 2013. Supply elasticity and the housing cycle of the 2000s. Real Estate Econ. 41 (4), 793–813.

Davidoff, T., 2016. Supply constraints are not valid instrumental variables for home prices because they are correlated with many demand factors. Crit. Financ. Rev. Forthcoming.

Davis, M.A., Ortalo-Magné, F., 2011. Household expenditures, wages, rents. Rev. Econ. Dyn. 14 (2), 248–261.

Davis, M.A., Heathcote, J., 2005. Housing and the business cycle. Int. Econ. Rev. 46 (3), 751–784.

Davis, M.A., Heathcote, J., 2007. The price and quantity of residential land in the united states. J. Monet. Econ. 54 (8), 2595–2620.

Davis, M.A., Van Nieuwerburgh, S., 2015. Housing, finance and the macroeconomy. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), Handbook of Regional and Urban Economics, vol. 5. Elsevier, Amsterdam, pp. 813–886.

De Nardi, M., French, E., Jones, J., 2010. Why do the elderly save? The role of medical expenses. J. Polit. Econ. 118, 39–75.

Deaton, A., 1992. Understanding Consumption. Oxford University Press, Oxford, UK.

Detemple, J., Serrat, A., 2003. Dynamic equilibrium with liquidity constraints. Rev. Financ. Stud. 16 (2), 597–629.

Díaz, A., Luengo-Prado, M.J., 2008. On the user cost and homeownership. Rev. Econ. Dyn. 11 (3), 584–613.

Dorofeenko, V., Lee, G.S., Salyer, K.D., 2014. Risk shocks and housing supply: a quantitative analysis. J. Econ. Dyn. Control. 45, 194–219.

Englund, P., Ioannides, Y.M., 1997. House price dynamics: an international empirical perspective. J. Hous. Econ. 6, 119–136.

Epple, D., Quintero, L., Sieg, H., 2015. A new appproach to estimating hedonic equilibrium for metropolitan housing markets. Working Paper, University of Pennsylvania.

Epstein, L.G., Zin, S.E., 1989. Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Thoeretical Framework. Econometrica 57 (4), 937–969.

Evans, G.W., Honkapohja, S., 2009. Learning and macroeconomics. Annu. Rev. Econ. 1, 421–449.

Favilukis, J., Ludvigson, S.C., van Nieuwerburgh, S., 2016. The macroeconomic effects of housing wealth, housing finance and limited risk sharing in general equilibrium. J. Polit. Econ. Forthcoming.

Fernandez-Villaverde, J., Krueger, D., 2007. Consumption over the life cycle: facts from consumer expenditure survey data. Rev. Econ. Stat. 89 (3), 552–565.

Fernandez-Villaverde, J., Krueger, D., 2010. Consumption and saving over the life cycle: how important are consumer durables? Macroecon. Dyn. 15, 725–770.

Fisher, J.D., 2007. Why does household investment lead business investment over the business cycle? J. Polit. Econ. 115 (1), 141–168.

Flavin, M., Nakagawa, S., 2008. A model of housing in the presence of adjustment costs: a structural interpretation of habit persistence. Am. Econ. Rev. 98 (1), 474–495.

Flavin, M., Yamashita, T., 2002. Owner-occupied housing and the composition of the household portfolio. Am. Econ. Rev. 92 (1), 345–362.

Flemming, J.S., 1969. The utility of wealth and the utility of windfalls. Rev. Econ. Stud. 36, 55–66.

Fraumeni, B.M., 1997. The measurement of depreciation in the U.S. national income and product accounts. Surv. Curr. Bus. 77, 7–23.

Fudenberg, D., Levine, D., 1998. Learning in Games. M.I.T. Press, Cambridge, MA.

Garriga, C., Manuelli, R., Peralta-Alva, A., 2012. A model of price swings in the housing market. Working Paper, Federal Reserve Bank of St. Louis, 2012-022A.

Garriga, C., Kydland, F.E., Šustek, R., 2013. Mortgages and monetary policy. NBER Working Paper No. 19744.

Geanakoplos, J., 2011. What's missing from macroeconomics: endogenous leverage and default. In: Jarocinski, M., Smets, F., Thimann, C. (Eds.), Approaches to Monetary Policy Revisited–Lesson from the Crisis, vol. 2011, pp. 220–238.

Gerali, A., Neri, S., Sessa, L., Signoretti, F.M., 2010. Credit and banking in a DSGE model of the euro area. J. Money Credit Bank. 42 (6), 107–141.

Giglio, S., Maggiori, M., Stroebel, J., 2016. No-bubble condition: model-free tests in housing markets. Econometrica 84 (3), 1047–1091.

Glaeser, E.L., Gottlieb, J.D., Gyourko, J., 2013. Can cheap credit explain the housing boom? Housing and the Financial Crisis. National Bureau of Economic Research, Inc., pp. 301–359.

Glaeser, E.L., Gyourko, J., Morales, E., Nathanson, C.G., 2014. Housing dynamics: an urban approach. J. Urban Econ. 81, 45–56.

Gomme, P., Kyland, F.E., Rupert, P., 2001. Home production meets time to build. J. Polit. Econ. 109 (5), 1115–1131.

Gorea, D., Midrigan, V., 2015. Liquidity constraints in the U.S. housing market. Working Paper, NYU.

Gourinchas, P.O., Parker, J., 2002. Consumption over the life cycle. Econometrica 70, 47–91.

Grandmont, J.M., 1977. Temporary general equilibrium. Econometrica 45, 535–572.

Greenwood, J., Hercowitz, Z., 1991. The allocation of capital and time over the business cycle. J. Polit. Econ. 99 (6), 1188–1214.

Greenwood, J., Rogerson, R., Wright, R., 1995. Household production in real business cycle theory. In: Cooley, T.F. (Ed.), Frontiers of Business Cycle Research. Princeton University Press, Princeton, NJ, pp. 157–174.

Grossman, S.J., Laroque, G., 1990. Asset pricing and optimal portfolio choice in the presence of illiquid durable consumption goods. Econometrica 58, 22–51.

Guerrieri, L., Iacoviello, M., 2015. OccBin: a toolkit for solving dynamic models with occasionally binding constraints easily. J. Monet. Econ. 70 (C), 22–38.

Guerrieri, V., Hartley, D., Hurst, E., 2013. Endogenous gentrification and housing price dynamics. J. Public Econ. 103 (5), 1664–1696.

Gyourko, J., Linneman, P., 1997. The changing influences of education, income, family structure, and race on homeownership by age over time. J. Hous. Res. 8 (1), 1–25.

Gyourko, J., Saiz, A., Summers, A.A., 2008. A new measure of the local regulatory environment for housing markets. Urban Stud. 45 (3), 693–729.

Han, L., Strange, W., 2015. The microstructure of housing markets: search, bargaining, and brokerage. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), Handbook of Regional and Urban Economics, vol. 5. Elsevier, Amsterdam, pp. 813–886.

Henderson, J.V., Ioannides, Y.M., 1983. A model of housing tenure choice. Am. Econ. Rev. 73 (1), 98–113.

Himmelberg, C., Mayer, C., Sinai, T., 2005. Assessing high house prices: bubbles, fundamentals and misperceptions. J. Econ. Perspect. 19, 67–92.

Hornstein, A., Praschnik, J., 1997. Intermediate inputs and sectoral comovements in the business cycle. J. Monet. Econ. 40, 573–595.

Hubbard, R.G., Skinner, J., Zeldes, S.P., 1995. Precautionary savings and social insurance. J. Polit. Econ. 103, 360–399.

Hurst, E., Stafford, F., 2004. Home is where the equity is: mortgage refinancing and household consumption. J. Money Credit Bank. 36 (6), 985–1014.

Iacoviello, M., 2005. House prices, borrowing constraints, and monetary policy in the business cycle. Am. Econ. Rev. 95 (3), 739–764.

Iacoviello, M., Neri, S., 2010. Housing market spillovers: evidence from an estimated DSGE model. Am. Econ. Rev. 2, 125–164.

Iacoviello, M., Pavan, M., 2013. Housing and debt over the life cycle and over the business cycle. J. Monet. Econ. 60 (2), 221–238.

Jaccard, I., 2011. Asset pricing and housing supply in a production economy. J. Macroecon. 11 (1). Article 33.

Jordà, O., Schularick, M., Taylor, A.M., 2016a. The great mortgaging: housing finance, crises and business cycles. Econ. Policy 31 (85), 107–152.

Jordà, O., Schularick, M., Taylor, A.M., 2016b. Leveraged bubbles. J. Monet. Econ. Forthcoming.

Justiniano, A., Primiceri, G.E., Tambalotti, A., 2015a. Credit supply and the housing boom. Working Paper, Northwestern University.

Justiniano, A., Primiceri, G.E., Tambalotti, A., 2015b. A simple model fo subrprime borrowers and credit growth. Working Paper, Northwestern University.

Kaneko, M., 1982. The central assignment game and the assignment of markets. J. Math. Econ. 10, 205–232.

Kaplan, G., Violante, G., 2014. A model of the consumption response to fiscal stimulus. Econometrica 82 (4), 1199–1239.

Kaplan, G., Mitman, K., Violante, G., 2016a. Non-durable consumption and housing net worth in the great recession: evidence from easily accessible data. Working Paper, NYU.

Kaplan, G., Moll, B., Violante, G., 2016b. Monetary policy according to HANK. Working Paper, Princeton University.

Kathari, S., Saporta-Eksten, I., Yu, E., 2013. The (un)importance of geographical mobility in the great recession. Rev. Econ. Dyn. 16, 553–563.

Kermani, A., 2012. Cheap credit, collateral, and the boom-bust cycle. Working Paper, Berkeley.

Keys, B.J., Mukherjee, T., Seru, A., Vig, V., 2010. Did securitization lead to lax screening? Evidence from subprime loans. Q. J. Econ. 125 (1), 307–362.

King, M., 1990. Discussion. Econ. Policy 11, 383–387.

Kiyotaki, N., Moore, J., 1997. Credit Cycles. J. Polit. Econ. 105 (2), 211–248.

Kiyotaki, N., Michaelides, A., Nikolov, K., 2011. Winners and losers in housing markets. J. Money Credit Bank. 43 (2-3), 255–296.

Knoll, K., Schularick, M., Steger, T., 2014. No price like home: global house prices, 1870-2012. CEPR Working Paper No. 10166.

Koijen, R.S.J., Hemert, O., Van Nieuwerburgh, S., 2009. Mortgage timing. J. Financ. Econ. 93 (2), 292–324.

Krusell, P., Smith, T., 1998. Income and wealth heterogeneity in the macroeconomy. J. Polit. Econ. 106 (5), 867–896.

Kuminoff, N.V., Pope, J.C., 2013. The value of residential land and structures during the great housing boom and bust. Land Econ. 89 (1), 1–29.

Kydland, F.E., Rupert, P., Sustek, R., 2012. Housing dynamics over the business cycle. NBER Working Paper No. 18432.

Landvoigt, T., 2015. Financial intermediation, credit risk, and credit supply during the housing boom. Working Paper, University of Texas at Austin.

Landvoigt, T., 2016. Housing Demand During the Boom: The Role of Expectations and Credit Constraints. Forthcoming Review of Financial Studies.

Landvoigt, T., Piazzesi, M., Schneider, M., 2015. The housing market(s) of san diego. Am. Econ. Rev. 105 (4), 1371–1407.

Li, W., Yao, R., 2007. The life-cycle effects of house price changes. J. Money Credit Bank. 39 (6), 1375–1409.

Liu, Z., Wang, P., Zha, T., 2013. Land-price dynamics and macroeconomic fluctuations. Econometrica 81 (3), 1167–1184.

Lu, H., 2008. Hedging house price risk in the presence of lumpy transaction costs. J. Urban Econ. 64 (2), 270–287.

Lustig, H., Van Nieuwerburgh, S., 2005. Housing collateral, consumption insurance and risk premia: an empirical perspective. J. Financ. 60 (3), 1167–1219.

Lustig, H., Van Nieuwerburgh, S., 2010. How much does household collateral constrain regional risk sharing? Rev. Econ. Dyn. 13 (2), 265–294.

Mayerhauser, N., Reinsdorf, M., 2007. Housing services in the national economic accounts. Working Paper, Bureau of Economic Analysis.

Mian, A., Sufi, A., 2009. The consequences of mortgage credit expansion: evidence from the U.S. mortgage default crisis. Q. J. Econ. 124 (4), 1449–1496.

Mian, A., Sufi, A., 2011. House prices, home equity-based borrowing, and the U.S. household leverage crisis. Am. Econ. Rev. 101 (5), 2132–2156.

Mian, A., Sufi, A., 2015. Fraudulent income overstatement on mortgage applications during the credit expansion of 2002 to 2005. NBER Working Paper No. 20947.

Mian, A., Rao, K., Sufi, A., 2013. Household balance sheets, consumption, and the economic slump. Q. J. Econ. 128 (4), 1687–1726.

Monacelli, T., 2009. New keynesian models, durable goods, and collateral constraints. J. Monet. Econ. 56 (2), 242–254.

Muellbauer, J., Murphy, A., 1990. Is the UK balance of payments sustainable? Econ. Policy 5 (11), 347–395.

Munnell, A.H., Tootell, G.M.B., Browne, L.E., McEneaney, J., 1996. Mortgage lending in boston: interpreting HMDA data. Am. Econ. Rev. 86 (1), 25–53.

Nathanson, C.G., Zwick, E., 2015. Arrested development: theory and evidence of supply-side speculation in the housing market. Working Paper, Booth and Kellogg.

Olsen, E., Zabel, J., 2015. United States housing policies. In: Henderson, J.V., Strange, W.C. (Eds.), In: Handbook of Regional and Urban Economics, vol. 5. Elsevier, Amsterdam, pp. 887–986.

Ortalo-Magné, F., Rady, S., 1999. Boom in, bust out: young households and the housing price cycle. Eur. Econ. Rev. 43, 755–766.

Ortalo-Magné, F., Rady, S., 2006. Housing market dynamics: on the contribution of income shocks and credit constraints. Rev. Econ. Stud. 73, 459–485.

Pagano, M., 1990. Discussion. Econ. Policy 11, 387–390.

Piazzesi, M., Schneider, M., 2008. Inflation illusion, credit, and asset pricing. In: Campbell, J. (Ed.), Asset Pricing and Monetary Policy. Chicago University Press, Chicago, IL, pp. 147–181.

Piazzesi, M., Schneider, M., 2009a. Inflation and the price of real assets. Federal Reserve Bank of Minneapolis Research Department Staff Report 423.

Piazzesi, M., Schneider, M., 2009b. Momentum traders in the housing market: survey evidence and a search model. Am. Econ. Rev. 99 (2), 406–411.

Piazzesi, M., Schneider, M., Tuzel, S., 2007. Housing, consumption and asset pricing. J. Financ. Econ. 83, 531–569.

Piskorski, T., Tchistyi, A., 2010. Optimal mortgage design. Rev. Financ. Stud. 23, 3098–3140.

Poterba, J.M., 1984. Tax subsidies to owner-occupied housing: an asset-market approach. Q. J. Econ. 99, 729–752.

Poterba, J.M., 1991. House price dynamics: the role of tax policy and demography. Brook. Pap. Econ. Act. 2, 143–203.

Ríos-Rull, J.V., Sánchez-Marcos, V., 2010. An aggregate economy with different size houses. J. Eur. Econ. Assoc. 6 (2/3), 705–714.

Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. J. Polit. Econ. 82 (1), 34–55.

Rubio, M., 2011. Fixed- and variable-rate mortgages, business cycles, and monetary policy. J. Money Credit Bank. 43 (4), 657–688.

Saiz, A., 2010. The geographic determinants of housing supply. Q. J. Econ. 125 (3), 1253–1296.

Sargent, T.J., 1999. The Conquest of American Inflation. Princeton University Press, Princeton, NJ.

Schwartz, E.S., Torous, W.N., 1989. Prepayment and the valuation of mortgage-backed securities. J. Financ. 44 (2), 375–392.

Shleifer, A., Vishny, R.W., 1992. Liquidation Values and Debt Capacity: A Market Equilibrium Approach. J. Financ. 47 (4), 1343–1366.

Silos, P., 2007. Housing, portfolio choice and the macroeconomy. J. Econ. Dyn. Control 31, 2774–2801.

Sinai, T., Souleles, N., 2005. Owner-occupied housing as a hedge against rent risk. Q. J. Econ. 120 (2), 763–789.

Sinai, T., Souleles, N., 2013. Can owning a home hedge the risk of moving? Am. Econ. J. Econ. Pol. 5 (2), 282–312.

Slemrod, J., 1982. Down-Payment Constraints: Tax Policy Effects in a Growing Economy with Rental and Owner-Occupied Housing. Public Finance Quart. 10 (2), 193–217.

Sommer, K., Sullivan, P., Verbrugge, R., 2013. The equilibrium effect of fundamentals on house prices and rents. J. Monet. Econ. 60, 854–870.

Stanton, R., 1995. Rational prepayment and the valuation of mortgage-backed securities. Rev. Financ. Stud. 8 (3), 677–708.

Stein, J., 1995. Prices and trading volume in the housing market: a model with downpayment effects. Q. J. Econ. 110 (2), 379–406.

Stokey, N., 2009. Moving costs, nondurable consumption and portfolio choice. J. Econ. Theory 144 (6), 2419–2439.

Stroebel, J., Vavra, J., 2015. House prices, local demand, and retail prices. Working Paper, NYU.

Tsatsaronis, K., Zhu, H., 2014. What drives house price dynamics: cross-country evidence. BIS Q. Rev. 65–78. March.

Turner, T.M., Smith, M.T., 2009. Exits from homeownership: the effects of race, ethnicity, and income. J. Reg. Sci. 49 (1), 1–32.

Van Nieuwerburgh, S., Weill, P.O., 2010. Why has house price dispersion gone up? Rev. Econ. Stud. 77 (4), 1567–1606.

Wong, A., 2016. Population aging and the transmission mechanism of monetary policy. Working Paper, Northwestern.

Woodward, S.E., Hall, R.E., 2012. Diagnosing consumer confusion and sub-optimal shopping effort: theory and mortgage-market evidence. Am. Econ. Rev. 102 (7), 3249–3276.

Yang, F., 2009. Consumption over the life cycle: how different is housing? Rev. Econ. Dyn. 12, 423–443.

Yao, R., Zhang, H.H., 2005. Optimal consumption and portfolio choices with risky housing and borrowing constraints. Rev. Financ. Stud. 18 (1), 197–239.

CHAPTER 20

Term Structure of Uncertainty in the Macroeconomy

J. Borovička*, L.P. Hansen†
*New York University, New York, NY; NBER, Cambridge, MA, United States
†University of Chicago, Chicago, IL; NBER, Cambridge, MA, United States

Contents

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.06.005

Abstract

Dynamic economic models make predictions about impulse responses that characterize how macroeconomic processes respond to alternative shocks over different horizons. From the perspective of asset pricing, impulse responses quantify the exposure of macroeconomic processes and other cash flows to macroeconomic shocks. Financial markets provide compensations to investors who are exposed to these shocks. Adopting an asset pricing vantage point, we describe and apply methods for computing exposures to macroeconomic shocks and the implied compensations represented as elasticities over alternative payoff horizons. The outcome is a term structure of macroeconomic uncertainty.

Keywords

Asset pricing, Impulse response functions, Shock elasticities, Financing frictions, Martingales

JEL Classification Codes

C10, C32, C58, E44, G12, G32

1. INTRODUCTION

Impulse response functions quantify the impact of alternative economic shocks on future economic outcomes. In so doing, they provide a way to assess the importance of alternative sources of fluctuations. Building on the insights of Yule (1927) and Slutsky (1927), Frisch featured an important line of research on the "impulse and propagation problem" aimed at answering the question asking what are the sources of fluctuations and how they are propagated over time. An impulse, captured formally by the realization of a random shock, has an impact on an economic time series in all of the subsequent time periods. Response functions depict the intertemporal responses. Sims (1980) showed how to apply this approach in a tractable way to multivariate time series with a vector of underlying shocks, and he exposed the underlying challenges for identification. Subsequent research developed nonlinear counterparts to impulse response functions.

Macroeconomic shocks also play an important role in asset pricing. By their very nature, macroeconomic shocks cannot be diversified and investors exposed to those shocks require compensations. The resulting market-based remunerations differ depending on how cash flows are exposed to the alternative macroeconomic shocks. We call the compensations risk prices, and there is a term structure that characterizes these prices as a function of the investment horizon. In this chapter, we study methods for depicting this term structure and illustrate its use by comparing pricing implications across models. This leads us to formalize the exposure and pricing counterpart to impulse response functions familiar to macroeconomists. We call these objects shock-exposure and shock-price elasticities. Our calculations require either an empirical-based or model-based stochastic discount factor process along with a representation of how alternative cash flows with macroeconomic components respond to shocks.

There is an alternative way to motivate the calculations that we perform. A common characterization of risk aversion looks at local certainty equivalent calculations for small variance changes in consumption. We deviate in two ways. First, when making small changes, we do not use certainty as our benchmark but rather the equilibrium consumption from the stochastic general equilibrium model. This leads us to make more refined adjustments in the exposure to uncertainty. Second, movements in consumption at future dates could be induced by any of the macroeconomic shocks with occurrences at dates between tomorrow and this future date. Thus, similar to Hansen et al. (1999) and Alvarez and Jermann (2004), we have a differential measure depending on the specific shock and the dates of the impacts.

Empirical finance often focuses on the measurement of risk premia on alternative financial assets. In our framework, these risk premia reflect the exposure to uncertainty and the compensation for that exposure. Risk premia change when exposures change, when the prices of those exposures change, or both. We use explicit economic models to help us quantify these two channels by which risk premia are determined, but a more empirically based approach could also be applied provided that the uncertainty prices for shocks could be inferred. While there are interesting challenges in identification to explore, we will abstract from those challenges in this chapter.

Our chapter:

- defines and constructs a term structure of shock-exposure and shock-price elasticities applicable to nonlinear Markov models;

- compares these constructions to impulse response functions commonly used in macroeconomics;
- describes computational approaches pertinent for discrete-time and continuous-time models;
- applies the methods to continuous-time macroeconomic models with financing frictions proposed by He and Krishnamurthy (2013) and Brunnermeier and Sannikov (2014).

2. MATHEMATICAL FRAMEWORK

We introduce a framework designed to encompass a large class of macroeconomic and asset pricing general equilibrium models. There is an underlying stationary Markov model that is used to capture the stochastic growth of a vector of time series of economic variables. The Markov model emerges as the "reduced form" of a solution to a dynamic stochastic equilibrium model of the macroeconomy. Modeling stationary growth rates allows for inclusion of shocks that have permanent effects and nontrivial long-horizon implications for risk compensations. We provide a range of illustrative applications of this framework throughout the chapter, and we devote Section 7 to a more extensive exploration of nonlinear continuous-time models with financial constraints.

We start with a probability space $(\Omega, \mathcal{F}, P)$. On this probability space, there is an n-dimensional, stationary and ergodic Markov process $X = \{X_t : t \in \mathbb{N}\}$ and a k-dimensional process W of independent and identically distributed shocks. Unless otherwise specified, we assume that each W_t is a multivariate standard normal random variable. We will have more to say about discrete states and shocks that are not normally distributed in Section 3.5.

The Markov process is initialized at X_0. Denote $\mathfrak{F} = \{\mathcal{F}_t : t \in \mathbb{N}\}$ the completed filtration generated by the histories of W and X_0. We suppose that X is a solution to a law of motion

$$\begin{aligned} X_{t+1} &= \psi(X_t, W_{t+1}) \\ Y_{t+1} - Y_t &= \phi(X_t, W_{t+1}). \end{aligned} \tag{1}$$

The state vector X_t contains both exogenously specified states and endogenous ones. We presume full information in the sense that the shock W_{t+1} can be depicted in terms of $(X_t, Y_{t+1} - Y_t)$. In more general circumstances we would incorporate a solution to a filtering problem if we are to match an information structure to (X, Y), a filtering problem that is perhaps solved by economic agents.

Consistent intertemporal pricing together with the Markov property leads us to use a class of stochastic processes called multiplicative functionals. These processes are built from the underlying Markov process and will be used to model cash flows and stochastic discount factors. Since many macroeconomic time series grow or decay over time, we use the state vector X to model the growth rate of such processes. In particular, let the dynamics of a *multiplicative functional* M be defined as[a]

$$\log M_{t+1} - \log M_t = \kappa(X_t, W_{t+1}). \tag{2}$$

The components of Y are examples of multiplicative functionals. Since X is stationary, the process $\log M$ has stationary increments. A revealing example is the conditionally linear model

$$\kappa(X_t, X_{t+1}) = \beta(X_t) + \alpha(X_t) \cdot W_{t+1}$$

where $\beta(x)$ allows for nonlinearity in the conditional mean and $\alpha(x)$ introduces stochastic volatility.

We denote G a generic cash-flow process and S the equilibrium determined stochastic discount factor process, both modeled as multiplicative functionals. While we adopt a common mathematical formulation for both, G is expected to grow and S is expected to decay over time, albeit in stochastic manners.

Equilibrium models in macroeconomics and asset pricing build on the premise of utility-maximizing investors trading in arbitrage-free markets. Arbitrage-free pricing implies the existence of a strictly positive stochastic discount factor process S that can be used to infer equilibrium asset prices. Stochastic discount factors provide a convenient way to depict the observable implications of asset pricing models.[b] In this chapter, we consider a stochastic discount factor process that compounds the one-period stochastic discount factor increments in order to value multiperiod claims.

Definition 1 A stochastic discount factor S is a positive (with probability one) stochastic process such that for any $t, j \geq 0$ and payoff G_{t+j} maturing at time $t + j$, the time-t price is given by

$$\mathcal{Q}_t\left[G_{t+j}\right] = E\left[\left(\frac{S_{t+j}}{S_t}\right) G_{t+j} \mid \mathcal{F}_t\right]. \tag{3}$$

Notice that this definition does not restrict the date zero stochastic discount factor, S_0. This initialization may be chosen in a convenient manner. If markets are complete, then this stochastic discount factor is unique up to the initialization. Equations of the type (3) arise from investors' optimality conditions in the form of Euler equations. In an equilibrium model with complete markets, the stochastic discount factor is typically equated with the marginal rate of substitution of an unconstrained investor. The identity of such a person can change over time and across states. In some models with incomplete markets, the stochastic discount factor process ceases to be unique. There are different

[a] Multiplicative functionals are often initialized at one, or equivalently $\log M_0 = 0$. We will abuse this jargon a bit by allowing ourselves other possible initiations.

[b] See Hansen and Richard (1987) for an initial discussion of stochastic discount factors.

shadow prices for nontraded risk exposures but a common pricing of the exposures with explicit compensations in financial markets. With other forms of trading frictions, the pricing equalities can be replaced by pricing inequalities, still expressed using a stochastic discount factor.

In our framework, we will suppose that equilibrium stochastic discount factors inherit the multiplicative functional structure. Market frictions, portfolio constraints, and other types of market imperfections will then introduce distortions into formula (3). We will study such distortions in models with financial constraints in Section 7.

Notice that definition (3) of the stochastic discount factor involves an expectations operator. This expectations operator in general represents investors' beliefs about the future. Here, we have imposed rational expectations by assuming that investors' beliefs are identical to the data-generating probability measure P. This measure is that implied by historical evidence or by the fully specified model. Investors' beliefs, however, may differ from P and there exists alternative approaches to modeling these deviations in interesting ways. While the modeling of investors' beliefs is an important building block of the asset pricing framework, in this chapter we abstract from these considerations and impose rational expectations throughout the text.

3. ASSET PRICING OVER ALTERNATIVE INVESTMENT HORIZONS

We price cash flows exposed to macroeconomic uncertainty and modeled as multiplicative processes. Consider a generic cash flow process G, say the dividend process or an equilibrium consumption process. We start with a baseline payoff G_t maturing in individual periods $t = 0, 1, 2, \ldots$ and parameterize stochastic perturbations of this process. In particular, we derive measures that capture the sensitivity of expected payoff to exposure to alternative macroeconomic shocks, and the sensitivity of the associated risk compensations. We follow the convention in empirical finance by depicting compensations in terms of expected returns per unit of some measure of riskiness. The compensations differ depending upon which shock we target when we construct stochastic perturbations. The method relies on a comparison of the pricing of payoff G_t relative to another payoff that is marginally more exposed to risk in a particular way.

The cash flows G arising from equilibrium models will often have the form of multiplicative processes (2). A special case of such cash flows are payoffs that are positive functions of the Markov state, $\psi(X_t)$. These payoffs will be featured prominently in our subsequent analysis.

3.1 One-Period Pricing

We are interested in the pricing of payoffs maturing at different horizons, but we start with a simple one-period conditionally lognormal environment. This environment will provide an explicit link to familiar calculations in asset valuation. Suppose that

$$\log G_1 = \beta_g(X_0) + \alpha_g(X_0) \cdot W_1$$
$$\log S_1 - \log S_0 = \beta_s(X_0) + \alpha_s(X_0) \cdot W_1$$

where G_1 is the payoff to which we assign values and S_1 is the one-period stochastic discount factor used to compute these values. The one-period return on this investment is the payoff in period one divided by the period-zero price:

$$R_1 \doteq \frac{G_1}{Q_0[G_1]} = \frac{\left(\frac{G_1}{G_0}\right)}{E\left[\left(\frac{S_1}{S_0}\right)\left(\frac{G_1}{G_0}\right) \mid X_0\right]}.$$

The logarithm of the expected return can then be calculated explicitly as:

$$\begin{aligned} \log E[R_1 \mid X_0 = x] &= \log E\left[\left(\frac{G_1}{G_0}\right) \mid X_0 = x\right] - \log E\left[\left(\frac{S_1}{S_0}\right)\left(\frac{G_1}{G_0}\right) \mid X_0 = x\right] \\ &= \underbrace{-\beta_s(x) - \frac{|\alpha_s(x)|^2}{2}}_{\text{risk-free rate}} \underbrace{-\alpha_s(x) \cdot \alpha_g(x)}_{\text{risk premium}}. \end{aligned} \tag{4}$$

This compensation is expressed in terms of expected returns as is typical in asset pricing. Notice that we are using logarithms of proportional risk premia as a starting point.

Imagine applying this calculation to a family of such payoffs parameterized in part by α_g. The vector α_g defines a vector of exposures to the components of the normally distributed shock W_1. Then $-\alpha_s$ is the vector of shock "prices" representing the compensation for exposure to these shocks.

The risk prices in this conditionally lognormal model have a familiar conditional linear structure known from one-period factor models. In these models, the so-called factor loadings α_g on the individual shocks W_1 are multiplied by factor prices $-\alpha_s$. The total compensation in terms of an expected return is thus the product of the quantity of risk (risk exposure) and the price per unit of this risk. There are analogous simplifications for continuous-time diffusion models since the local evolution in such models is conditionally normal.

In a nonlinear multiperiod environment, this calculation ceases to be straightforward. We would, however, still like to infer measures of the quantity of risk and the associated price of the risk. We therefore explore a related derivation that will yield the same results in this one-period lognormal environment but will also naturally extend to a nonlinear setup and multiple-period horizons.

3.1.1 One-Period Shock Elasticities

We parameterize a family of random variables $H_1(\mathsf{r})$ indexed by r using

$$\log H_1(\mathsf{r}) = \mathsf{r}\nu(X_0) \cdot W_1 - \frac{\mathsf{r}^2}{2}|\nu(X_0)|^2 \tag{5}$$

where r is an auxiliary scalar parameter. The vector of exposures $\alpha_h(X_0)$ is normalized to

$$E\left[|\nu(X_0)|^2\right] = 1.$$

With this normalization,

$$E[H_1(\mathsf{r})|X_0] = 1.$$

Even when shocks are not normally distributed, we shall find it convenient to construct $H_1(\mathsf{r})$ to have a unit conditional expectation.

Given the baseline payoff G_t, form a parameterized family of payoffs $G_1H_1(\mathsf{r})$ given by

$$\log G_1 - \log G_0 + \log H_1(\mathsf{r}) = \underbrace{[\alpha_g(X_0) + \mathsf{r}\nu(X_0)]}_{\text{new shock exposure}} \cdot W_1 + \beta_g(X_0) - \frac{\mathsf{r}^2}{2}|\nu(X_0)|^2.$$

The new cash flow $G_1H_1(\mathsf{r})$ has shock exposure $\alpha_g(X_0) + \mathsf{r}\nu(X_0)$ and is thus more exposed to the vector of shocks W_1 in the direction $\nu(X_0)$. By changing r, we alter the magnitude of the exposure in direction $\nu(X_0)$. By choosing different vectors $\nu(X_0)$, we alter the combinations of shocks whose impact we want to investigate. A typical example of an $\nu(X_0)$ would be a coordinate vector e_j with a single one in jth place. In that case, we infer the pricing implications of the jth component of the shock vector W_1. In some applications it may be convenient to make $\nu(X_0)$ explicitly depend on X_0. For instance, Borovička et al. (2011) propose scaling of ν with X_0 in models with stochastic volatility.

The payoffs $G_1H_1(\mathsf{r})$ imply a corresponding family of logarithms of expected returns as in Eq. (4):

$$\log E[R_1(\mathsf{r}) \mid X_0 = x] = \log E\left[\left(\frac{G_1}{G_0}\right) H_1(\mathsf{r}) \mid X_0 = x\right] \\ - \log E\left[\left(\frac{S_1}{S_0}\right)\left(\frac{G_1}{G_0}\right) H_1(\mathsf{r}) \mid X_0 = x\right].$$

We are interested in comparing the expected return of the payoff $G_1H_1(\mathsf{r})$ relative to $G_1 = G_1H_1(0)$. Since our exposure direction $\nu(X_0)$ has a unit standard deviation, by differentiating with respect to r we compute an elasticity

$$\frac{d}{d\mathsf{r}} \log E[R_1(\mathsf{r}) \mid X_0 = x]|_{\mathsf{r}=0} \\ = \frac{d}{d\mathsf{r}} \log E\left[\left(\frac{G_1}{G_0}\right) H_1(\mathsf{r}) \mid X_0 = x\right]\Bigg|_{\mathsf{r}=0} - \frac{d}{d\mathsf{r}} \log E\left[\left(\frac{S_1}{S_0}\right)\left(\frac{G_1}{G_0}\right) H_1(\mathsf{r}) \mid X_0 = x\right]\Bigg|_{\mathsf{r}=0}.$$

This elasticity measures the sensitivity of the expected return on the payoff G_1 to an increase in exposure to the shock in the direction $\nu(x)$. The calculation leads us to define the counterparts of quantity and price elasticities from microeconomics:

1. The one-period *shock-exposure elasticity*

$$\varepsilon_g(x,1) = \frac{d}{d\mathsf{r}} \log E\left[\left(\frac{G_1}{G_0}\right) H_1(\mathsf{r}) \mid X_0 = x\right]\Bigg|_{\mathsf{r}=0} = \alpha_g(x) \cdot \nu(x)$$

measures the sensitivity of the expected payoff G_1 to an increase in exposure in the direction $\nu(x)$.

2. The one-period *shock-price elasticity*

$$\begin{aligned}\varepsilon_p(x,1) \;&= \frac{d}{d\mathsf{r}} \log E\left[\left(\frac{G_1}{G_0}\right) H_1(\mathsf{r}) \mid X_0 = x\right]\Bigg|_{\mathsf{r}=0} - \frac{d}{d\mathsf{r}} \log E\left[\left(\frac{S_1}{S_0}\right)\left(\frac{G_1}{G_0}\right) H_1(\mathsf{r}) \mid X_0 = x\right]\Bigg|_{\mathsf{r}=0} \\ &= -\alpha_s(x) \cdot \nu(x)\end{aligned}$$

measures the sensitivity of the compensation, in units of expected return, for this exposure.

Notice that the shock-exposure elasticity recovers the exposure vector $\alpha_g(x)$, and individual components of this vector can be obtained by varying the choice of the direction of the perturbation $\nu(x)$. Similarly, the shock-price elasticity recovers the vector of prices $-\alpha_s(x)$ associated with the risks embedded in the shock W_1.

In this one-period case, we replicated a straightforward decomposition of the expected return (4) into quantities and prices of risk. Now we move to the characterization of the asset pricing implications over longer horizons.

3.2 Multiperiod Investment Horizon

Consider the parameterized payoff $G_t H_1(\mathsf{r})$ with a date-zero price $E[S_t G_t H_1(\mathsf{r}) \mid X_0 = x]$. This is a payoff maturing at time t that has the same growth rate as payoff G_t except period one when the growth rate is stochastically perturbed by $H_1(\mathsf{r})$. The logarithm of the expected return (yield to maturity) is

$$\begin{aligned}\log E[R_{0,t}(\mathsf{r}) \mid X_0 = x] \doteq \log E\left[\left(\frac{G_t}{G_0}\right) H_1(\mathsf{r}) \mid X_0 = x\right] \\ - \log E\left[\left(\frac{S_t}{S_0}\right)\left(\frac{G_t}{G_0}\right) H_1(\mathsf{r}) \mid X_0 = x\right].\end{aligned}$$

Following our previous analysis, we construct two elasticities:

1. *shock-exposure elasticity*

$$\varepsilon_g(x,t) = \frac{d}{d\mathsf{r}} \log E\left[\left(\frac{G_t}{G_0}\right) H_1(\mathsf{r}) \mid X_0 = x\right]\Bigg|_{\mathsf{r}=0}$$

2. *shock-price elasticity*

$$\varepsilon_p(x,t)=\frac{d}{d\mathsf{r}}\log E\left[\left(\frac{G_t}{G_0}\right)H_1(\mathsf{r})\mid X_0=x\right]\Bigg|_{\mathsf{r}=0}-\frac{d}{d\mathsf{r}}\log E\left[\left(\frac{S_t}{S_0}\right)\left(\frac{G_t}{G_0}\right)H_1(\mathsf{r})\mid X_0=x\right]\Bigg|_{\mathsf{r}=0} \tag{6}$$

These elasticities are functions of the investment horizon t, and thus we obtain a term structure of elasticities. The dependence on the current state $X_0 = x$ incorporates possible time variation in the sensitivity of expected returns to exposure to shocks.

3.3 A Change of Measure and an Impulse Response for a Multiplicative Functional

Notice that the shock elasticities defined in the previous section have a common mathematical structure expressed using the multiplicative functionals $M = S$ and $M = SG$. Given a multiplicative functional M, we define

$$\varepsilon(x,t)=\frac{d}{d\mathsf{r}}\log E\left[\left(\frac{M_t}{M_0}\right)H_1(\mathsf{r})\mid X_0=x\right]\Bigg|_{\mathsf{r}=0}. \tag{7}$$

Taking the derivative in (7), we obtain

$$\varepsilon(x,t)=\nu(x)\cdot\frac{E\left[\left(\frac{M_t}{M_0}\right)W_1\mid X_0=x\right]}{E\left[\left(\frac{M_t}{M_0}\right)\mid X_0=x\right]}. \tag{8}$$

Thus a major ingredient in the computation is the covariance between $\left(\frac{M_t}{M_0}\right)$ and W_1 conditioned on X_0.

The random variable $H_1(r)$ given by (5) is positive and has expectation equal to unity conditioned on X_0. Multiplication by this random variable has the interpretation of changing the probability distribution of W_1 from having mean zero to having a mean given by $\mathrm{r}\nu(X_0)$. Thus given a multiplicative process M

$$\begin{aligned}E\left[\left(\frac{M_t}{M_0}\right)H_1(\mathsf{r})\mid X_0=x\right]&=E\left(H_1(\mathsf{r})E\left[\left(\frac{M_t}{M_0}\right)\mid X_0,W_1\right]\mid X_0=x\right)\\&=\widetilde{E}\left(E\left[\left(\frac{M_t}{M_0}\right)\mid X_0,W_1\right]\mid X_0=x\right)\end{aligned}$$

where $\widetilde{E}$ presumes that the random vector W_1 is distributed as a multivariate normal with mean $\mathrm{r}\nu(x)$ consistent with our multiplication by $H_1(\mathsf{r})$.

3.4 Long-Horizon Pricing

Shock elasticities depict the term structure of risk as we change the maturity of priced payoffs. To aid our understanding of the overall shape of the term structure of elasticities, we characterize the long-horizon limits of these shock elasticities. We provide a characterization for a general multiplicative process that takes the form of a factorization. A multiplicative process is a product of a geometric constant growth or decay process, a positive martingale, and a ratio of a function of the Markov state in date zero and date t. Since the factorization is applicable to any member of a general class of multiplicative processes, we apply it to both stochastic discount factor processes and positive cash flow processes.

As in Hansen and Scheinkman (2009) and Hansen (2012), we use Perron–Frobenius theory to provide a factorization of multiplicative processes. Given a multiplicative process M, solve the equation

$$E\left[\left(\frac{M_t}{M_0}\right)e(X_t) \mid X_0 = x\right] = \exp(\eta t)e(x) \tag{9}$$

for an unknown function $e(x)$ that is strictly positive and an unknown number η. The solution is independent of the choice of the horizon t.

Consider the pair (e, η) that solves (9) and form

$$\frac{\widetilde{M}_t}{\widetilde{M}_0} \doteq \exp(-\eta t)\frac{e(X_t)}{e(X_0)}\left(\frac{M_t}{M_0}\right). \tag{10}$$

The stochastic process $\widetilde{M}$ is a martingale under P, since

$$\begin{aligned} E\left[\widetilde{M}_{t+1} \mid \mathcal{F}_t\right] &= \frac{\exp[-\eta(t+1)]}{e(X_0)}\frac{M_t}{M_0}\widetilde{M}_0\, E\left[\frac{M_{t+1}}{M_t}e(X_{t+1}) \mid \mathcal{F}_t\right] \\ &= \exp(-\eta t)\frac{e(X_t)}{e(X_0)}\frac{M_t}{M_0}\widetilde{M}_0 = \widetilde{M}_t. \end{aligned}$$

Consequently, expression (10) can be reorganized as

$$\frac{M_t}{M_0} = \exp(\eta t)\frac{e(X_0)}{e(X_t)}\frac{\widetilde{M}_t}{\widetilde{M}_0}. \tag{11}$$

This formula provides a multiplicative decomposition of the multiplicative functional M into a deterministic drift $\exp(\eta t)$, a stationary function of the Markov state $e(x)$, and a martingale $\widetilde{M}$. This martingale component will be critical in characterizing long-term pricing implications.

Associated with the martingale $\widetilde{M}$ is a probability measure $\widetilde{P}$ such that for every measurable function Z of the Markov process between dates zero and t,

$$E\left(\widetilde{M}_t Z \mid X_0 = x\right) = \widetilde{E}\,(Z \mid X_0 = x)$$

where $\widetilde{E}\,(\cdot \mid X_0 = x)$ is the conditional expectation operator under the probability measure $\widetilde{P}$.[c]

In finite state spaces, Eq. (9) can be posed as a matrix problem with a solution that is an eigenvector with positive entries.

Example 3.1 In a finite-state Markov chain environment, Eq. (9) is a standard eigenvalue problem. Let realized value of the X_t be represented as alternative coordinate vectors. Suppose the ratio $\frac{M_{t+1}}{M_t}$ satisfies

$$\frac{M_{t+1}}{M_t} = (X_{t+1})'\mathbf{M}X_t$$

for some square matrix $\mathbf{M}$. In the same way, represent the one-period transition probabilities as a matrix $\mathbf{P}$. For $t = 1$, Eq. (9) becomes a vector equation

$$(\mathbf{P}^*\mathbf{M})\mathbf{e} = \exp(\eta)\mathbf{e}$$

where the operator * depicts elementwise multiplication, $(\mathbf{P}^*\mathbf{M})_{ij} = \mathbf{P}_{ij}\mathbf{M}_{ij}$. When

$$\sum_{j=0}^{\infty} \lambda^j (\mathbf{P}^*\mathbf{M})^j$$

has all strictly positive entries for some $0 < \lambda < 1$, the Perron–Frobenius theorem implies the existence of a unique normalized strictly positive eigenvector $\mathbf{e}$ associated with the largest eigenvalue $\exp(\eta)$ of the matrix $\mathbf{P}^*\mathbf{M}$. Then $e(X_t)$ in formula (9) is $\mathbf{e} \cdot X_t$.

In continuous state spaces, this factorization may not yield a unique strictly positive solution $e(x)$. Hansen and Scheinkman (2009) and Borovička et al. (2015) provide selection criteria based on the stochastic stability of the probability measure implied by the martingale component to guarantee uniqueness. Stochastic stability ensures that we have a valuable way to compute limiting approximations once we change measures. Here, we will assume that we have selected such a solution.[d]

[c] In order to completely define the measure $\widetilde{P}$, we also need to specify the unconditional probability distribution. For instance, $\widetilde{M}_0$ can be initiated to make $\widetilde{P}$ stationary. Since all pricing results in this chapter utilize conditional probability distributions, we abstract from these considerations here.

[d] Our formulation presumes an underlying Markovian structure. See Qin and Linetsky (2014b) for a more general starting point and an analogous factorization.

Factorization (11) leads to a characterization of long-horizon limits for the shock elasticities. Using this factorization in expression (7), we obtain[e]

$$\varepsilon(x,t) = \nu(x) \cdot \frac{\widetilde{E}[\hat{e}(X_t) W_1 \mid X_0 = x]}{\widetilde{E}\,[\hat{e}(X_t) \mid X_0 = x]}$$

where $\hat{e}(x) \doteq 1/e(x)$. Under technical assumptions the long-maturity limit for the shock elasticity is given by

$$\lim_{t\to\infty} \varepsilon(x,t) = \nu(x) \cdot \widetilde{E}\,[W_1 \mid X_0 = x].$$

The sensitivity of long-horizon payoffs to current shocks is therefore determined by the martingale components of the stochastic discount factor and the cash flow, and their implications for the expectations of shock W_1 as captured by the implied change in probability measures.

3.5 Non-Gaussian Frameworks

While we have made special reference to normally distributed shocks, our mathematical structure does not require this. We have featured perturbations $H_1(\mathsf{r})$ that are positive and expectations one. Risk prices in financial economics are denominated in terms of expected mean compensation per unit of risk. With normally distributed shocks, we measure risk in units of standard deviations. Provided that we adopt an interpretable way to denominate risk prices for other distributions, our methods continue to apply beyond the conditionally Gaussian framework. For instance, Zviadadze (2016) constructs shock elasticities in a stochastic environment with autoregressive gamma processes.

Another example are regime-shift models that may include both normally distributed shocks along with uncertain regimes. Exposure to macroeconomic regime-shift risk is of interest and can be characterized using shock elasticities by structuring appropriately the random variable $H_1(\mathsf{r})$. These switches can be exogenous (eg, exogenously modeled periods of low or high growth and volatility) or endogenous (eg, interest rate at the zero lower bound, financial sector in a period of binding financial constraints, or regime changes in government policies). We develop shock elasticities for regime-shift risk in Borovička et al. (2011).

For Markov chain models used to capture the regime shift dynamics of exogenous shocks see David (2008), Chen (2010), or Bianchi (2015) for some recent examples in the asset pricing literature and Liu et al. (2011) and Bianchi et al. (2013) in

[e] See Hansen and Scheinkman (2012) for a version of this result for a continuous-time diffusion model.

macroeconomic modeling. Regime switches are also utilized to model time variation in government policies, see Sims and Zha (2006), Liu et al. (2009), and Bianchi (2012) for regime switching in monetary policy rules, Davig et al. (2010, 2011) and Bianchi and Melosi (2016) for fiscal policy applications, and Chung et al. (2007) and Bianchi and Ilut (2015) for a combination of both. Farmer et al. (2011) and Foerster et al. (2014) analyze solution and estimation techniques in Markov chain models in conjunction with perturbation approximation methods. In Borovička and Hansen (2014), we introduce a tractable exponential-quadratic framework that permits semi-analytical formulas for shock elasticities and encompasses a large class of models solved using perturbation techniques.

4. RELATION TO IMPULSE RESPONSE FUNCTIONS

Impulse responses to specific structural shocks are a common way of representing the dynamic properties of macroeconomic models. As we mentioned previously, this idea goes back at least to Frisch (1933). Our elasticity computations change exposures of cash flows to shocks and explore the consequences for valuation. These constructs are closely related and in some circumstances are mathematically identical to impulse response functions. We explore these connections in this section.

To relate our elasticity calculation to an impulse response function, consider the conditional expectation

$$E\left[\left(\frac{M_t}{M_0}\right) \mid X_0, W_1 = w\right]$$

for alternative choices of w. Changing the value of w gives rise to the impulse response of M_t to a shock at date one. Instead of conditioning on alternative realized values of the shock at date one, as we have seen our computations are equivalent to changing the date zero distribution of W_1. A similarity in perspectives emerges because this distributional change could include a mean shift in the distribution for W_1. In practice, empirical macroeconomists typically study expectations of the logarithms of macroeconomic time series, often using linear models. For asset pricing it is important that we work with expectations of levels of macroeconomic quantities and cash flows, and account for nonlinearities. To compute shock *elasticities* we are lead to study the impact on the logarithm of the conditional expectation of M_t as developed in formula (7). In the remainder of this section, we consider two special cases in which the link to impulse functions is particularly close.

4.1 Lognormality

When M is a lognormal process, the impulse response functions for $\log M$ match exactly our shock elasticity as we will now see.

A linear vector-autoregression (VAR) model is a special case of the framework (1). Specifically X is a linear vector-autoregression with autoregression coefficient matrix $\bar{\mu}$ and shock-exposure matrix $\bar{\sigma}$:

$$X_{t+1} = \bar{\mu} X_t + \bar{\sigma} W_{t+1}. \tag{12}$$

We assume that the absolute values of eigenvalues of the matrix $\bar{\mu}$ are strictly less than one. Analogously, we introduce a multiplicative process M (constructed in general form in (2)) with evolution:

$$\log M_{t+1} - \log M_t = \bar{\beta} \cdot X_t + \bar{\alpha} \cdot W_{t+1}. \tag{13}$$

The shock W_{t+1} is distributed as a multivariate standard normal. With this construction of the multiplicative process M, we first study the responses of $\log M$.

4.1.1 Impulse Response Functions

Let $\nu(x) = \bar{\nu}$ where $\bar{\nu}$ is a vector with norm one. In typical applications, $\bar{\nu}$ is a coordinate vector. The impulse response function of $\log M_t$ for the linear combination of shocks chosen by the vector $\bar{\nu}$ is given by

$$E[\log M_t - \log M_0 \mid X_0 = x, W_1 = \bar{\nu}] - E[\log M_t - \log M_0 \mid X_0 = x, W_1 = 0] = \bar{\nu} \cdot \bar{\varrho}_t.$$

where the coefficients satisfy the recursions implied by (12) and (13). From (13), we have the recursion:

$$\bar{\varrho}_{t+1} - \bar{\varrho}_t = \left(\bar{\zeta}_t\right)' \bar{\beta} \tag{14}$$

with initial condition $\bar{\varrho}_1 = \bar{\alpha}$, and from (12):

$$\bar{\zeta}_{t+1} = \bar{\mu} \bar{\zeta}_t \tag{15}$$

with initial condition $\bar{\zeta}_1 = \bar{\sigma}$. Solving these recursions gives:

$$\begin{aligned} \bar{\zeta}_t &= \bar{\mu}^{t-1} \bar{\sigma} \\ \bar{\varrho}_t &= \bar{\alpha} + \left[(I - \bar{\mu})^{-1} \left(I - \bar{\mu}^{t-1} \right] \bar{\sigma} \right)' \bar{\beta}. \end{aligned} \tag{16}$$

The impulse response function in the linear model is thus a sequence of deterministic coefficients $\bar{\nu} \cdot \bar{\varrho}_t$. The first term, $\bar{\alpha} \cdot \bar{\nu}$, represents the immediate response arising from realization $\bar{\nu}$ of the current shock, while the remaining terms capture the subsequent propagation of the shock through the dynamics of state vector X as it influences $\log M$ in the future.

4.1.2 Shock Elasticities

Consider now our elasticity calculation. Write $\log M_t$ as its moving-average representation:

$$\log M_t = \sum_{j=0}^{t-1} \overline{\varrho}_j \cdot W_{t-j} + E(\log M_t \mid \mathcal{F}_0),$$

or equivalently

$$\begin{aligned}\log M_t - \log M_0 &= \sum_{j=1}^{t} \overline{\varrho}_j \cdot W_{t-j+1} + E(\log M_t - \log M_0 \mid X_0) \\ &= \sum_{j=1}^{t-1} \overline{\varrho}_j \cdot W_{t-j+1} + \overline{\varrho}_t \cdot W_1 + E(\log M_t - \log M_0 \mid X_0).\end{aligned}$$

Since the shocks W_t are independently distributed as a multivariate standard normals over time,

$$E\left[\left(\frac{M_t}{M_0}\right) \mid X_0 = x, W_1 = w\right] = \exp\left(\frac{1}{2}\sum_{j=1}^{t-1} \overline{\varrho}_j \cdot \overline{\varrho}_j\right) \exp\left(\overline{\varrho}_t \cdot W_1\right) \exp\left(E[\log M_t - \log M_0 \mid X_0]\right).$$

Using formula (8), we compute:

$$\varepsilon(x,t) = \frac{E\left[\left(\frac{M_t}{M_0}\right) W_1 \mid X_0 = x\right]}{E\left[\left(\frac{M_t}{M_0}\right) \mid X_0 = x\right]} = \frac{E[\exp(\overline{\varrho}_t \cdot W_1) W_1 \mid X_0 = x]}{E[\exp(\overline{\varrho}_t \cdot W_1)]} = \overline{\varrho}_t.$$

The second equality follows by observing that

$$\frac{\exp(\overline{\varrho}_t \cdot W_1)}{E[\exp(\overline{\varrho}_t \cdot W_1)]}$$

is strictly positive and has conditional expectation one. Multiplication by this random variable is equivalent to changing the distribution of W_1 from a multivariate standard normal to a multivariate normal with mean $\overline{\varrho}_t$. To summarize, in this lognormal case, the shock elasticities do not depend on the Markov state and they coincide with the impulse responses measured by $\overline{\nu} \cdot \overline{\varrho}_t$ for $t = 1, 2, \ldots$.

Consider in particular the shock-price elasticity (6). Notice that this shock-price elasticity consists of the difference of shock elasticities for G and SG, and thus we are lead to compute impulse response functions for $\log G$ and $\log S + \log G$. The additivity of the construction implies that the impulse response function coefficients for the latter are $\overline{\nu} \cdot \overline{\varrho}_{s,t} + \overline{\nu} \cdot \overline{\varrho}_{g,t}$, and thus the resulting shock-price elasticity corresponds to the impulse response function of $-\log S$, with coefficients $-\overline{\nu} \cdot \overline{\varrho}_{s,t}$.

4.1.3 Long-Term Pricing Revisited

In this example, as discussed in Hansen et al. (2008) there is a close link between the factorization described in Section 3.4 and the additive decompositions of linear time series. Beveridge and Nelson (1981) and Blanchard and Quah (1989) extracted a martingale component in linear models and used it to characterize the impact of permanent shocks.[f]

Consider solving

$$E\left[\left(\frac{M_1}{M_0}\right)e(X_1)\mid X_0=x\right]=\exp(\eta)e(x)$$

for the pair (e, η), where the evolution of M is given by (13). In this special case, a straightforward calculation using formulas for lognormals gives:

$$\begin{aligned}\log e(x) &= E\left(\sum_{j=0}^{\infty}\overline{\beta}\cdot X_{t+j}|X_t=x\right)\\ &=(\overline{\beta})'(I-\overline{\mu})^{-1}x,\end{aligned}$$

and[g]

$$\eta=\frac{1}{2}|\overline{\alpha}'+\overline{\beta}'(I-\overline{\mu})^{-1}\overline{\sigma}|^2.$$

Under the change of measure associated with the martingale $\widetilde{M}$ in the multiplicative factorization, W_1 has a mean equal to

$$\overline{\sigma}'(I-\overline{\mu}')^{-1}\overline{\beta}+\overline{\alpha}$$

which is independent of the state vector. Notice that this is also the limiting value of $\overline{\varrho}_t$ as given in (16). In this lognormal example

$$\log M_{t+1}-\log M_t+\log e(X_{t+1})-\log e(X_t)=\left[\overline{\beta}'(I-\overline{\mu}')^{-1}\overline{\sigma}+\overline{\alpha}'\right]W_{t+1}$$

where the right-hand side gives the permanent shock to $\log M$ as constructed in Beveridge and Nelson (1981) and Blanchard and Quah (1989). In VAR analyses, transitory shocks are typically constructed as linear combinations of W_{t+1} that are uncorrelated with this permanent shock. On the other hand $\log e(X_{t+1})$ and its innovation are typically correlated with the permanent shock.

This simple connection between permanent shocks and permanent components to pricing ceases to hold in more general nonlinear environments. Hansen (2012) has a more

[f] Hansen (2012) constructs an additive decomposition of $\log M$ in a continuous-time version of our nonlinear framework.

[g] If we were to include a constant included in the evolution of $\log M$, this would be added to η.

complete discussion of the relation between the permanent component to $\log M$ and the martingale component to M outside this lognormal specification.

4.2 Continuous-Time Diffusions

In this section, we focus on a framework with uncertainty modeled using Brownian shocks, and apply it to models with financial constraints in Section 7. While the Brownian information setup is not without loss of generality, it provides tools for a pedagogically transparent treatment and shows the close connection between shock elasticities and impulse responses. In Borovička et al. (2011) we also consider jumps in the form of regime shifts in continuous-time Markov chains and applications to consumption-based asset pricing models.

Let X be a Markov diffusion on $\mathcal{X} \subseteq \mathbb{R}^n$:

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t$$

with initial condition $X_0 = x$. Here, $\mu(x)$ is an n-dimensional vector and $\sigma(x)$ is an $n \times k$ matrix for each vector x in $\mathbb{R}^n$. In addition W is a k-dimensional Brownian motion. We use this underlying Markov process to construct a multiplicative process M via:

$$\log M_t = \log M_0 + \int_0^t \beta(X_u)du + \int_0^t \alpha(X_u) \cdot dW_u \tag{17}$$

where $\beta(x)$ is a scalar and $\alpha(x)$ is a k-dimensional vector, or, in differential notation,

$$d\log M_t = \beta(X_t)dt + \alpha(X_t) \cdot dW_t. \tag{18}$$

Thus M_t depends on the initial conditions $(X_0, M_0) = (x, m)$ and the innovations to the Brownian motion W between dates zero and t. Let $\{\mathcal{F}_t : t \geq 0\}$ be the (completed) filtration generated by the Brownian motion between time zero and time t along with any initial information captured by $\mathcal{F}_0$.

As before, stochastic discount factors and cash flows in this environment are specific versions of a multiplicative process M. This multiplicative process is exposed to two types of risk. The first source of risk exposure is the "local," or infinitesimal, risk in term $\alpha(X_u) \cdot dW_u$ in (17). The second source of risk comes from the time variation in X_t and the state dependence of coefficients $\beta(x)$ and $\alpha(x)$, and is manifested over longer horizons.

4.2.1 Haussmann–Clark–Ocone Formula

There is a natural counterpart to a moving-average representation for diffusions. Importantly, the moving-average coefficients are, in general, state dependent. They entail computing so-called Malliavin derivatives of the date-u shock to the process $\log M_t$ for $t \geq u$, denoted $\mathcal{D}_u \log M_t$. We do not develop Malliavin differentiation as a formal mathematical construct but instead proceed heuristically.[h] This calculation of a Malliavin derivative gives the random response to a shock at date-u and is only restricted to be t-measurable where $t \geq u$. By forming the date-u conditional expectation we get the expected response as of the date of the shock. The computation is localized by making the time interval over which the shock acts on the process $\log M_t$ arbitrarily small, which allows for the formal construction of a derivative.

The calculation of $\mathcal{D}_u \log M_t$ has two uses analogous to the lognormal example we examined earlier. First, the (random) impulse response function for $\log M$

$$\varrho_t(X_0) = \nu(X_0) \cdot E(\mathcal{D}_0 \log M_t \mid F_0) = \nu(X_0) \cdot E[\mathcal{D}_0(\log M_t - \log M_0) \mid X_0]$$

for $t \geq 0$ where $\nu(X_0)$ determines which conditional linear combination of the shocks is subject to an impulse. The resulting responses depend on conditioning information captured by X_0, in contrast to lognormal models in which responses depend only on the horizon $t \geq 0$. Relatedly we obtain the Haussmann–Clark–Ocone formula for the process $\log M$ that cumulates the impact shocks at various dates as a stochastic integral:

$$\log M_t = \int_0^t E(\mathcal{D}_u \log M_t \mid F_u) \cdot dW_u + E(\log M_t \mid \mathcal{F}_0),$$

where we may think of $E(\mathcal{D}_u \log M_t \mid F_u)$ as the counterpart to a coefficient vector in a moving-average representation. These random variables satisfy recursions analogous to (14) and (15). For a more detailed construction, see Borovička et al. (2014).

We use the rules of Malliavin differentiation (analogous to more familiar forms of differentiation):

$$\mathcal{D}_u M_t = M_t \mathcal{D}_u \log M_t,$$

implying that the impulse response function for the process M is

$$\begin{aligned}\nu(X_0) \cdot E(\mathcal{D}_0 M_t | \mathcal{F}_0) &= \nu(X_0) \cdot E(M_t \mathcal{D}_0 \log M_t | \mathcal{F}_0)\\ &= M_0 \nu(X_0) \cdot E\left[\left(\frac{M_t}{M_0}\right) \mathcal{D}_0(\log M_t - \log M_0) | X_0\right)\end{aligned}$$

for $t \geq 0$.

4.2.2 Shock Elasticities for Diffusions

The construction of shock elasticities in Section 3 perturbs the cash flow by exposing it to a specified shock in the next period. In the continuous-time model, we devise a perturbation of M over a short time interval $[0, \mathrm{r}]$ and then study the implications as $\mathrm{r} \searrow 0$. The resulting construction exploits the local linearity of continuous-time models with Brownian shocks.

[h] For a textbook treatment of Malliavin calculus see Di Nunno et al. (2009) or Nualart (2006).

Specifically, we construct the process H^{r} such that

$$\log H_t^{\mathrm{r}} = \int_0^{\mathrm{r}\wedge\mathrm{t}} \nu(X_u)\cdot dW_u - \frac{1}{2}\int_0^{\mathrm{r}\wedge\mathrm{t}} |\nu(X_u)|^2 du,$$

where $\mathrm{r}\wedge\mathrm{t} = \min\{\mathrm{r},\mathrm{t}\}$. Notice that this process is exposed to the Brownian shock on the time interval $[0,\mathrm{r}]$, with exposure vector $\nu(x)$, and stays constant after r. We assume that $\nu(x)$ is restricted so that the process H^{r} is a martingale. We use H^{r} to construct the perturbed process MH^{r}:

$$\log M_t + \log H_t^{\mathrm{r}} = \log M_0 + \int_0^t \beta(X_u)du - \frac{1}{2}\int_0^{\mathrm{r}\wedge\mathrm{t}} |\nu(X_u)|^2 du + \int_0^t \alpha(X_u)\cdot dW_u + \int_0^{\mathrm{r}\wedge\mathrm{t}} \nu(X_u)\cdot dW_u$$

Notice that on the interval $[0,\mathrm{r}]$, the exposure of the perturbed process to the Brownian shock is

$$[\alpha(X_u) + \nu(X_u)]\cdot dW_u.$$

As r ↘ 0, we are perturbing $\log M$ over an arbitrarily small interval.

As in Borovička et al. (2014), we define the shock elasticity for M at horizon t as

$$\varepsilon(x,t) = \lim_{\mathrm{r}\searrow 0} \frac{1}{\mathrm{r}} \log E\left[\left(\frac{M_t}{M_0}\right) H_t^{\mathrm{r}} \mid X_0 = x\right]$$

and show that this limit can be expressed as

$$\begin{aligned} \varepsilon(x,t) &= \nu(x)\cdot \frac{E\left(\mathcal{D}_0 \frac{M_t}{M_0} \mid X_0 = x\right)}{E\left[\left(\frac{M_t}{M_0}\right) \mid X_0 = x\right]} \\ &= \nu(x)\cdot \frac{E\left[\left(\frac{M_t}{M_0}\right)\mathcal{D}_0 \log M_t \mid X_0 = x\right)}{E\left[\left(\frac{M_t}{M_0}\right) \mid X_0 = x\right]}. \end{aligned} \tag{19}$$

The first equality in (19) is a limiting version of (8) divided by $E\left[\left(\frac{M_t}{M_0}\right) \mid X_0 = x\right]$ since the Haussmann–Clark–Ocone formula applied to $\frac{M_t}{M_0}$ has a contribution

$$E\left(\mathcal{D}_0 \frac{M_t}{M_0} \mid X_0 = x\right) dW_0$$

for the date zero increment. The limiting covariance between $\frac{M_t}{M_0}$ and dW_0 is therefore $E\left(\mathcal{D}_0 \frac{M_t}{M_0} \mid X_0 = x\right)$. From the second equality in (19), these elasticities coincide with the diffusion counterpart to impulse responses $\mathcal{D}_0(\log M_t - \log M_0)$ for $\log M_t - \log M_0$ weighted by

$$\frac{\left(\frac{M_t}{M_0}\right)}{E\left[\left(\frac{M_t}{M_0}\right) \mid X_0 = x\right]}$$

when averaging over future outcomes. For the lognormal model, the weighting is inconsequential. In Borovička et al. (2011), we provide details of this derivation and some related calculations including the following alternative formula relevant for computation:

$$\varepsilon(x,t) \doteq \nu(x) \cdot \left[\sigma(x)'\left(\frac{\partial}{\partial x}\log E\left[\left(\frac{M_t}{M_0}\right) \mid X_0 = x\right]\right) + \alpha(x)\right]. \tag{20}$$

The shock-elasticity formula (20) has a natural interpretation. The sensitivity of the multiplicative process M to a shock in the next instant consists of two terms. The term $\alpha(x)$ represents the direct impact of the Brownian shock on the evolution of M in expression (18). The partial derivative with respect to x captures the sensitivity of the conditional expectation to movements in the state vector, and it is multiplied by the exposure matrix $\sigma(x)$ to express the sensitivity with respect to the shock vector W. The use of the derivative of the logarithm in (18) justifies the term shock *elasticity*. The instantaneous short-term elasticity is $\alpha(x) \cdot \nu(x)$.[i]

5. DISCRETE-TIME FORMULAS AND APPROXIMATION

In the preceding sections, we developed formulas for shock-price and shock-exposure elasticities for a wide class of models driven by a state vector with Markov dynamics (1). We now present a tractable implementation that, when applicable, makes the computations straightforward to apply. The discussion draws on methods developed in Borovička and Hansen (2014).[j] We also provide Matlab software implementing the

[i] The instantaneous shock-price elasticity is $-\alpha_s(x) \cdot \nu(x)$ which coincides with the notion of a risk price vector that represents the compensation for exposure to Brownian increments.

[j] See Nakamura et al. (2016) for another discrete-time implementation of these methods.

solution methods described in this section including a toolkit that computes shock elasticities for models solved using Dynare.[k]

We start by introducing a convenient exponential-quadratic framework that we use for modeling the state vector X and the resulting multiplicative processes. In this framework, conditional expectations of multiplicative processes and the shock elasticities are available in a convenient functional form. We then consider a special class of approximate solutions to dynamic macroeconomic models constructed using perturbation methods. We show how to approximate the equilibrium dynamics, additive and multiplicative functionals, and the resulting shock elasticities. By construction, the dynamics of these approximate solutions will be nested within the exponential-quadratic framework.

5.1 Exponential-Quadratic Framework

We study dynamic systems for which the state vector can be partitioned as $X = \left(X_1', X_2'\right)'$ where the two components follow the laws of motion:

$$\begin{aligned} X_{1,t+1} &= \Theta_{10} + \Theta_{11} X_{1,t} + \Lambda_{10} W_{t+1} \\ X_{2,t+1} &= \Theta_{20} + \Theta_{21} X_{1,t} + \Theta_{22} X_{2,t} + \Theta_{23}(X_{1,t} \otimes X_{1,t}) \\ &\quad + \Lambda_{20} W_{t+1} + \Lambda_{21}(X_{1,t} \otimes W_{t+1}) + \Lambda_{22}(W_{t+1} \otimes W_{t+1}). \end{aligned} \tag{21}$$

We restrict the matrices Θ_{11} and Θ_{22} to have stable eigenvalues. Notice that the restrictions imposed by the triangular structure imply that the process X_1 is linear, while the process X_2 is linear conditional on the evolution of X_1.

The class of multiplicative functionals M that interest us satisfies, for $Y = \log M$, the restriction

$$\begin{aligned} Y_{t+1} - Y_t &= \Gamma_0 + \Gamma_1 X_{1,t} + \Gamma_2 X_{2,t} + \Gamma_3(X_{1,t} \otimes X_{1,t}) \\ &\quad + \Psi_0 W_{t+1} + \Psi_1(X_{1,t} \otimes W_{t+1}) + \Psi_2(W_{t+1} \otimes W_{t+1}). \end{aligned} \tag{22}$$

In what follows we use a $1 \times k^2$ vector Ψ to construct a $k \times k$ symmetric matrix $\text{sym}[\text{mat}_{k,k}(\Psi)]$ such that[l]

$$w'(\text{sym}[\text{mat}_{k,k}(\Psi)])w = \Psi(w \otimes w).$$

[k] Dynare is a freely available Matlab/Octave toolkit for solving and analyzing dynamic general equilibrium models (see http://www.dynare.org). Our software is available at http://borovicka.org/software.html.

[l] In this formula $\text{mat}_{k,k}(\Psi)$ converts a vector into a $k \times k$ matrix and the sym operator transforms this square matrix into a symmetric matrix by averaging the matrix and its transpose. Appendix A introduces convenient notation for the algebra underlying the calculations in this and subsequent sections.

This representation will be valuable in some of the computations that follow. We use additive functionals to represent stochastic growth via a technology shock process or aggregate consumption, and to represent stochastic discounting used in representing asset values.

The system (21)–(22) is rich enough to accommodate stochastic volatility, which has been featured in the asset pricing literature and to a lesser extent in the macroeconomics literature. For instance, the state variable $X_{1,t}$ can capture a linear process for conditional volatility, and $X_{2,t}$ the conditional growth rate of cash flows. The coefficient Ψ_1 in (22) then determines the time variation in the conditional volatility of the growth rate of M, while Λ_{21} in (21) impacts the conditional volatility of the changes in the growth rate. In Section 5.2, we will map the solution obtained using perturbation approximations into this framework as well.

A virtue of parameterization (21)–(22) is that it gives quasi-analytical formulas for our dynamic elasticities. The implied model of the stochastic discount factor has been used in a variety of reduced-form asset pricing models. Later we will use an approximation to deduce this dynamical system.

We illustrate the convenience of this functional form by calculating the logarithms of conditional expectations of multiplicative functionals of the form (22). Consider a function that is linear-quadratic in $x = (x_1', x_2')'$:

$$\log f(x) = \Phi_0 + \Phi_1 x_1 + \Phi_2 x_2 + \Phi_3 (x_1 \otimes x_1). \tag{23}$$

Then conditional expectations are of the form:

$$\begin{aligned} \log E\left[\left(\frac{M_{t+1}}{M_t}\right) f(X_{t+1}) \mid X_t = x\right] &= \log E[\exp(Y_{t+1} - Y_t) f(X_{t+1}) \mid X_t = x] \\ &= \Phi_0^* + \Phi_1^* x_1 + \Phi_2^* x_2 + \Phi_3^* (x_1 \otimes x_1) \\ &= \log f^*(x) \end{aligned} \tag{24}$$

where the formulas for Φ_i^*, $i = 0, \ldots, 3$ are given in Appendix A. This calculation maps a function f into another function f^* with the same functional form. Our multiperiod calculations exploit this link. For instance, repeating these calculations compounds stochastic growth or discounting. Moreover, we may exploit the recursive Markov construction in (24) initiated with $f(x) = 1$ to obtain:

$$\log E\left[\left(\frac{M_t}{M_0}\right) \mid X_0 = x\right] = \Phi_{0,t}^* + \Phi_{1,t}^* x_1 + \Phi_{2,t}^* x_2 + \Phi_{3,t}^* (x_1 \otimes x_1)$$

for appropriate choices of $\Phi_{i,t}^*$.

5.1.1 Shock Elasticities

To compute shock elasticities given in (8) under the convenient functional form, we construct:

$$\frac{E\left[\left(\frac{M_t}{M_0}\right)W_1 \mid X_0 = x\right]}{E\left[\left(\frac{M_t}{M_0}\right) \mid X_0 = x\right]} = \frac{E\left[\left(\frac{M_1}{M_0}\right)E\left[\left(\frac{M_t}{M_1}\right) \mid X_1\right]W_1 \mid X_0 = x\right]}{E\left[\left(\frac{M_1}{M_0}\right)E\left(\frac{M_t}{M_1} \mid X_1\right) \mid X_0 = x\right]}.$$

Notice that the random variable:

$$L_{1,t} = \frac{\left(\frac{M_1}{M_0}\right)E\left(\frac{M_t}{M_1} \mid X_1\right)}{E\left[\left(\frac{M_1}{M_0}\right)\left(\frac{M_t}{M_1} \mid X_1\right) \mid X_0 = x\right]} \tag{25}$$

has conditional expectation one. Multiplying this positive random variable by W_1 and taking expectations is equivalent to changing the conditional probability distribution and evaluating the conditional expectation of W_1 under this change of measure. Then under the transformed measure, using a complete-the-squares argument we may show that W_1 remains normally distributed with a covariance matrix that is no longer the identity and a mean conditioned on $X_0 = x$ that is affine in x_1. The formulas are given in Appendix B. Thus the shock elasticity function $\varepsilon(x, t)$ can be computed recursively using formulas that are straightforward to implement. We show in Appendix B that the resulting shock elasticity function is also affine in the state x_1.

5.2 Perturbation Methods

In macroeconomic models, the equilibrium Markov dynamics (1) is typically ex ante unknown and needs to be solved for from a set of equilibrium conditions. We now describe a solution method for dynamic general equilibrium models that yields a solution in the form of an approximate law of motion that is a special case of the exponential-quadratic functional form analyzed in Section 5.1. This solution method, based on Holmes (1995) and Lombardo and Uhlig (2014), constructs a perturbation approximation where the first- and second-order terms follow the restricted dynamics (21).

For the purposes of approximation, we consider a family of models parameterized by q and study first- and second-order approximations around this limit system in which $\mathsf{q} = 0$. For each q, we consider the system (equations

$$0 = E(g[X_{t+1}(\mathsf{q}), X_t(\mathsf{q}), X_{t-1}(\mathsf{q}), \mathsf{q}W_{t+1}, \mathsf{q}W_t, \mathsf{q}] \mid \mathcal{F}_t). \tag{26}$$

The $\mathsf{q} = 0$ equation system is one without shocks, and more generally small values q will make the shocks less consequential. There are well-known saddle-point stability conditions on the system (26) that lead to a unique equilibrium of the linear approximation (see Blanchard and Kahn, 1980 or Sims, 2002), and we assume that these are satisfied. Following Holmes (1995) and Lombardo and Uhlig (2014), we form an approximating system by deducing the dynamic evolution for the pathwise derivatives with respect to q and evaluated at $\mathsf{q} = 0$. Our derivation will be admittedly heuristic as is much of the related literature in macroeconomics.

To build a link to the parameterization in Section 5.1, we feature a second-order expansion:

$$X_t(\mathsf{q}) \approx X_{0,t} + \mathsf{q}X_{1,t} + \frac{\mathsf{q}^2}{2}X_{2,t},$$

where $X_{m,t}$ is the mth order, date t component of the stochastic process. We abstract from the dependence on initial conditions by restricting each component process to be stationary. Our approximating process will similarly be stationary.[m] The expansion leads to laws of motion for the component processes $X_{1,\cdot}$ and $X_{2,\cdot}$. The joint process $(X_{1,\cdot}, X_{2,\cdot})$ will again be Markov, although the dimension of the state vector under the approximate dynamics doubles.

5.2.1 Approximating State Vector Dynamics

While $X_t(\mathsf{q})$ serves as a state vector in the dynamic system (26), the state vector itself depends on the parameter q. Suppose that $\mathcal{F}_t$ is the σ-algebra generated by the infinite history of shocks $\{W_j : j \leq t\}$. For each dynamic system, we presume that the state vector $X_t(\mathsf{q})$ is $\mathcal{F}_t$ measurable and that in forecasting future values of the state vector conditioned on $\mathcal{F}_t$ it suffices to condition on X_t. Although $X_t(\mathsf{q})$ depends on q, the construction of $\mathcal{F}_t$ does not. We now construct the dynamics for each of the component processes. The result will be a recursive system that has the same structure as the triangular system (21).

Define $\bar{x}$ to be the solution to the equation:

$$\bar{x} = \psi(\bar{x}, 0, 0),$$

which gives the fixed point for the deterministic dynamic system. We assume that this fixed point is locally stable. That is $\psi_x(\bar{x}, 0, 0)$ is a matrix with stable eigenvalues, eigenvalues with absolute values that are strictly less than one. Then set

$$X_{0,t} = \bar{x}$$

for all t. This is the zeroth-order contribution to the solution constructed to be time invariant.

In computing pathwise derivatives, we consider the state vector process viewed as a function of the shock history. Each shock in this history is scaled by the parameter q, which results in a parameterized family of stochastic processes. We compute derivatives with respect to this parameter where the derivatives themselves are stochastic processes.

[m] As argued by Lombardo and Uhlig (2014), this approach is computationally very similar to the pruning approach described by Kim et al. (2008) or Andreasen et al. (2010).

Given the Markov representation of the family of stochastic processes, the derivative processes will also have convenient recursive representations. In what follows we derive these representations.

Using the Markov representation, we compute the derivative of the state vector process with respect to q, which we evaluate at $\mathsf{q}=0$. This derivative has the recursive representation:

$$X_{1,t+1} = \psi_q + \psi_x X_{1,t} + \psi_w W_{t+1}$$

where ψ_q, ψ_x, and ψ_w are the partial derivative matrices:

$$\psi_q \doteq \frac{\partial \psi}{\partial \mathsf{q}}(\bar{x},0,0), \quad \psi_x \doteq \frac{\partial \psi}{\partial x'}(\bar{x},0,0), \quad \psi_w \doteq \frac{\partial \psi}{\partial w'}(\bar{x},0,0).$$

In particular, the term $\psi_w W_{t+1}$ reveals the role of the shock vector in this recursive representation. Recall that we have presumed that $\bar{x}$ has been chosen so that ψ_x has stable eigenvalues. Thus the first derivative evolves as a Gaussian vector autoregression. It can be expressed as an infinite moving average of the history of shocks, which restricts the process to be stationary. The first-order approximation to the original process is:

$$X_t \approx \bar{x} + \mathsf{q} X_{1,t}.$$

In particular, the approximating process on the right-hand side has $\bar{x} + \mathsf{q}(I - \psi_x)^{-1}\psi_q$ as its unconditional mean.

We compute the pathwise second derivative with respect to q recursively by differentiating the recursion for the first derivative. As a consequence, the second derivative has the recursive representation:

$$\begin{aligned} X_{2,t+1} = {} & \psi_{qq} + 2\left(\psi_{xq} X_{1,t} + \psi_{wq} W_{t+1}\right) \\ & + \psi_x X_{2,t} + \psi_{xx}(X_{1,t} \otimes X_{1,t}) + 2\psi_{xw}(X_{1,t} \otimes W_{t+1}) + \psi_{ww}(W_{t+1} \otimes W_{t+1}) \end{aligned} \tag{27}$$

where matrices ψ_{ij} denote the second-order derivatives of ψ evaluated at $(\bar{x},0,0)$ and formed using the construction of the derivative matrices described in Appendix A.2. As noted by Schmitt-Grohé and Uribe (2004), the mixed second-order derivatives ψ_{xq} and ψ_{wq} are often zero using second-order refinements to the familiar log approximation methods.

The second-derivative process $X_{2,\cdot}$ evolves as a stable recursion that feeds back on itself and depends on the first derivative process. We have already argued that the first derivative process $X_{1,t}$ can be constructed as a linear function of the infinite history of the shocks. Since the matrix ψ_x has stable eigenvalues, $X_{2,t}$ can be expressed as a linear-quadratic function of this same shock history. Since there are no feedback effects from $X_{2,t}$ to $X_{1,t+1}$, the joint process $(X_{1,\cdot}, X_{2,\cdot})$ constructed in this manner is necessarily stationary.

The dynamic evolution for $(X_{1,\cdot}, X_{2,\cdot})$ is a special case of the triangular system (21) given in Section 5.1. When the shock vector W_t is a multivariate standard normal, we can utilize results from Section 5.1 to produce exact formulas for conditional expectations of exponentials of linear-quadratic functions in $(X_{1,t}, X_{2,t})$. We exploit this construction in the subsequent section. For details on the derivation of the approximating formulas, see Appendix A.

5.3 Approximating the Evolution of a Stationary Increment Process

Consider the approximation of a parameterized family of multiplicative processes with increments given by:

$$\log M_{t+1}(\mathsf{q}) - \log M_t(\mathsf{q}) = \kappa[X_t(\mathsf{q}), \mathsf{q} W_{t+1}, \mathsf{q}]$$

and an initial condition $\log M_0$. We use the function κ in conjunction with q to parameterize implicitly a family of additive functionals. We approximate the resulting additive functionals by

$$\log M_t \approx \log M_{0,t} + \mathsf{q} \log M_{1,t} + \frac{\mathsf{q}^2}{2} \log M_{2,t}$$

where the processes on the right-hand side have stationary increments.

Following the steps of our approximation of X, the recursive representation of the zeroth-order contribution to $\log M$ is

$$\log M_{0,t+1} - \log M_{0,t} = \kappa(\bar{x}, 0, 0) \doteq \bar{\kappa};$$

the first-order contribution is

$$\log M_{1,t+1} - \log M_{1,t} = \kappa_q + \kappa_x X_{1,t} + \kappa_w W_{t+1}$$

where κ_x and κ_w are the respective first derivatives of κ evaluated at $(\bar{x}, 0, 0)$; and the second-order contribution is

$$\begin{aligned}\log M_{2,t+1} - \log M_{2,t} = {} & \kappa_{qq} + 2\left(\kappa_{xq} X_{1,t} + \kappa_{wq} W_{t+1}\right) \\ & + \kappa_x X_{2,t} + \kappa_{xx}(X_{1,t} \otimes X_{1,t}) + 2\kappa_{xw}(X_{1,t} \otimes W_{t+1}) \\ & + \kappa_{ww}(W_{t+1} \otimes W_{t+1})\end{aligned}$$

where the κ_{ij}'s are the second derivative matrices constructed as in Appendix A.2. The resulting component additive functionals are special cases of the additive functional given in (22) that we introduced in Section 5.1.

5.3.1 Approximating Shock Elasticities

We could compute corresponding second-order approximations for the elasticities of multiplicative processes. Alternatively, since the approximating processes satisfy the structure given in Section 5.1, we have the formulas that we described earlier at our disposal and the supporting software. See Borovička and Hansen (2014) for further discussion.

5.4 Related Approaches

There also exist ad hoc approaches which mix orders of approximation for different components of the model or state vector. The aim of these methods is to improve the precision of the approximation along specific dimensions of interest, while retaining tractability in the computation of the derivatives of the function ψ. Justiniano and Primiceri (2008) use a first-order approximations but augment the solution with heteroskedastic innovations. Benigno et al. (2010) study second-order approximations for the endogenous state variables in which exogenous state variables follow a conditionally linear Markov process. Malkhozov and Shamloo (2011) combine a first-order perturbation with heteroskedasticity in the shocks to the exogenous process and corrections for the variance of future shocks. These solution methods are designed to produce nontrivial roles for stochastic volatility in the solution of the model and in the pricing of exposure to risk. The approach of Benigno et al. (2010) or Malkhozov and Shamloo (2011) gives alternative ways to construct the functional form used in Section 5.1.

5.5 Recursive Utility Investors

The recursive utility preference specification of Kreps and Porteus (1978) and Epstein and Zin (1989) warrants special consideration. By design, this specification of preferences avoids presuming that investors reduce intertemporal, compound consumption lotteries. Instead investors may care about the intertemporal composition of risk. It is motivated in part by an aim to allow for risk aversion to be altered without changing the elasticity of intertemporal substitution. Anderson et al. (2003), Maenhout (2004), and others extend the literature on risk-sensitive control by Jacobson (1973), Whittle (1990), and others and provide a "concern for robustness" interpretation of the utility recursion. Under this alternative interpretation the decision maker explores the potential misspecification of the transition dynamics as part of the decision-making process. This perspective yields a substantially different interpretation of the utility recursion. In establishing these connections in the control theory and economics literatures, it is sometimes advantageous to parameterize the utility recursion in a manner that depends explicitly on the parameter q. Borovička and Hansen (2013) and Bhandari et al. (2016) explore the resulting implications for approximations analogous to those studied here. Among other things, they provide a rationale for the first-order adjustments for recursive utility as suggested by Tallarini (2000), and they show novel ways in which higher-order adjustments are more impactful.

6. CONTINUOUS-TIME APPROXIMATION

Many interesting macroeconomic models specified in continuous time, including those we analyze in Section 7, require the application of numerical solution techniques. In the construction of shock elasticities, the central object of interest is the conditional expectation of M in (19). Consider the more general problem

$$\phi_t(x) \doteq E\left[\left(\frac{M_t}{M_0}\right)\phi_0(X_t) \mid X_0 = x\right] \tag{28}$$

with a given function ϕ_0. The conditional expectation of M is obtained by setting $\phi_0(x) \equiv 1$.

6.1 An Associated Partial Differential Equation

For the purposes of computation, we evaluate ϕ_t recursively. Given $\phi_{t-\Delta t}$ for small Δt, exploiting the time homogeneity of the underlying Markov process and applying the Law of Iterated Expectations gives:

$$\phi_t(x) = E\left[\left(\frac{M_{\Delta t}}{M_0}\right)\phi_{t-\Delta t}(X_{\Delta t}) \mid X_0 = x\right].$$

Itô's lemma applied to the product in the conditional expectation gives the linear, second-order partial differential equation:

$$\begin{aligned} \frac{\partial}{\partial t}\phi_t &= \left(\beta + \frac{1}{2}|\alpha|^2\right)\phi_t + \left[\frac{\partial}{\partial x}\phi_t\right] \cdot (\mu + \sigma\alpha) \\ &+ \frac{1}{2}\mathrm{tr}\left[\sigma'\left(\frac{\partial}{\partial x \partial x'}\phi_t\right)\sigma\right] \end{aligned} \tag{29}$$

with terminal condition ϕ_0 where tr(·) denotes the trace of the matrix argument. Eq. (29) is a generalization of the Kolmogorov backward equation for multiplicative processes of the type (17). The resulting partial differential equation can be solved using standard numerical techniques for differential equations.

6.2 Martingale Decomposition and a Change of Measure

To study the long-run implications for pricing, we proposed the extraction of a martingale component from the dynamics of the stochastic discount factors and cash flows by solving the Perron–Frobenius equation (9) for the strictly positive eigenfunction $e(x)$ and the associated eigenvalue η. In the Markov diffusion setup we localize this problem by computing

$$\lim_{t \to 0} \frac{E[M_t e(X_t) | X_0 = x] - \exp(\eta t) e(x)}{t} = 0.$$

Defining the infinitesimal operator

$$\mathbb{B}f(x) \doteq \frac{d}{dt} E[M_t f(X_t) | X_0 = x] \bigg|_{t=0}$$

we have

$$\mathbb{B}f = \left(\beta + \frac{1}{2}|\alpha|^2\right) f + (\sigma\alpha + \mu) \cdot \frac{\partial f}{\partial x} + \frac{1}{2} \mathrm{tr}\left(\sigma\sigma' \frac{\partial^2 f}{\partial x \partial x'}\right)$$

and we can write the limiting Perron–Frobenius equation as

$$\mathbb{B}e = \eta e \tag{30}$$

which is a second-order partial differential equation for the function $e(x)$ and a number η. Eq. (30) is known as the Sturm–Liouville equation. Notice that it is identical to the partial differential equation (29) when we are looking for an unknown discounted stationary function $\phi_t(x) = \exp(\eta t) e(x)$ with initial condition $\phi_0(x) = e(x)$. As before, there are typically multiple strictly positive solutions to this equation. Hansen and Scheinkman (2009) show that there is at most one such solution that preserves stochastic stability of the state vector X. We implicitly assume that we always choose such a solution.[n]

In line with the discussion from Section 3.4, we can now define the martingale $\widetilde{M}$ as[o]

$$\frac{\widetilde{M}_t}{\widetilde{M}_0} \doteq \exp(-\eta t) \frac{e(X_t)}{e(X_0)} \frac{M_t}{M_0}. \tag{31}$$

Applying Itô's lemma, we find that

$$d \log \widetilde{M}_t = \widetilde{\alpha}(X_t) \cdot dW_t - \frac{1}{2} |\widetilde{\alpha}(X_t)| dt$$

with

$$\widetilde{\alpha}(x) = \left[\sigma'(x) \frac{\partial}{\partial x} \log e(x) + \alpha(x)\right].$$

This implies that under the probability measure $\widetilde{P}$, the Brownian motion evolves as

$$dW_t = \widetilde{\alpha}(x) dt + d\widetilde{W}_t$$

where $\widetilde{W}$ is a Brownian motion under $\widetilde{P}$. It also implies that we can write the dynamics of the state vector under the change of measure as

[n] See also Borovička et al. (2015), Qin and Linetsky (2014a), Qin et al. (2016), Walden (2014), or Park (2015) for problems closely related to solving for the eigenvalue–eigenfunction pair (η, e).

[o] We note that the solution obtained using the localized version of the Perron–Frobenius problem may yield a process $\widetilde{M}$ that is only a local martingale. See Hansen and Scheinkman (2009) and Qin and Linetsky (2014b) for details and additional assumptions that assure $\widetilde{M}$ is a martingale. We will assume that such conditions are satisfied in the discussion that follows.

$$dX_t = \left[\mu(X_t) + \sigma(X_t)\,\tilde{\alpha}(X_t)\right]dt + \sigma(X_t)d\tilde{W}_t.$$

Inverting Eq. (31), we obtain the analog of the martingale decomposition in discrete time:

$$\frac{M_t}{M_0} = \exp(\eta t)\frac{e(X_0)}{e(X_t)}\frac{\tilde{M}_t}{\tilde{M}_0}. \tag{32}$$

To implement the factorization of the multiplicative functional M, we compute the strictly positive eigenfunction $e(x)$ and the associated eigenvalue η by solving the Perron–Frobenius problem (30). Since analytical solutions are often not available, we must rely on numerical methods. Pryce (1993) gives various numerical solution techniques for this problem. Notice that since there are typically infinitely many strictly positive solutions $e(x)$, it is necessary to determine which of these solutions is the relevant one.

An alternative approach is to utilize the time-dependent PDE (29) and exploit the fact that η is the principal eigenvalue, ie, one associated with the most durable component. In that case, one can start with an initial condition $\phi_0(x)$ that serves as a guess for the eigenfunction, and iterate on (29) to solve for $\phi_t(x)$ as $t \to \infty$. For large t, the solution should behave as

$$\phi_t(x) \approx \exp(\eta t)e(x)$$

and thus

$$\eta = \frac{\partial}{\partial t}\log\phi_t(x)\bigg|_{t\to\infty} \approx \frac{1}{\Delta t}\left[\log\phi_{t+\Delta t}(x) - \log\phi_t(x)\right]\bigg|_{t\to\infty}$$

and since the eigenfunction is only determined up to scale, we can use any proportional rescaling of ϕ_t as $e(x) \approx \exp(-\eta t)\phi_t(x)|_{t\to\infty}$.

6.3 Long-Term Pricing

We now apply the decomposition (32) in the shock elasticity formula (19) to obtain:

$$\varepsilon(x,t) \doteq \nu(x)\cdot\left[\sigma(x)'\left(\frac{\partial}{\partial x}\log e(x) + \frac{\partial}{\partial x}\log\tilde{E}\left[\frac{1}{e(X_t)}\,|\,X_0 = x\right]\right) + \alpha(x)\right].$$

Taking the limit as $t \to \infty$, the conditional expectation in brackets converges to a constant provided that we select a martingale that induces a probability measure under which X is stochastically stable. See Hansen and Scheinkman (2009) and Hansen (2012) for further discussion. Therefore,

$$\lim_{t\to\infty}\varepsilon(x,t) = \nu(x)\cdot\left[\sigma(x)'\frac{\partial}{\partial x}\log e(x) + \alpha(x)\right].$$

6.4 Boundary Conditions

The construction of shock elasticity functions requires solving the conditional expectations of M, for instance, by solving the partial differential equation (29). This requires proper specification of the boundary conditions not only in terms of the terminal condition $\phi_0(x)$ but also at the boundaries of the state space for the state vector X_t. The boundary behavior of the diffusion X is a central and often economically important part of the equilibrium, as we will see in the models with financial frictions discussed in Section 7. In those models, the state variable is a univariate diffusion and there are well understood characterizations of the boundary behavior based on the classical Feller boundary classification.[P] The textbook treatment of the boundary conditions for problem (28) typically abstracts from the impact of the multiplicative process M. While a detailed discussion of the boundary characterization is beyond the scope of this chapter, we briefly discuss how the inclusion of M can alter the analysis. In what follows, we utilize the martingale decomposition introduced in Section 3.4 and draw connections to the treatment of boundaries for scalar diffusions.

We represent the conditional expectation (32) using a Kolmogorov equation under the change of measure induced by $\widetilde{M}$. Using the martingale factorization (32) we write (28) as

$$\phi_t(x) \doteq E\left[\exp(\eta t)\frac{e(X_0)}{e(X_t)}\frac{\widetilde{M}_t}{\widetilde{M}_0}\phi_0(X_t) \mid X_0 = x\right].$$

Define

$$\psi_t(x) \doteq \exp(-\eta t)\frac{\phi_t(x)}{e(x)} = \widetilde{E}\left[\frac{\phi_0(X_t)}{e(X_t)} \mid X_0 = x\right] = \widetilde{E}\left[\psi_0(X_t) \mid X_0 = x\right] \tag{33}$$

with the initial condition $\psi_0(x) = \phi_0(x)/e(x)$. This converts the boundary condition problem into a standard Kolmogorov backward equation (Eq. (28) with $M \equiv 1$), albeit under the probability measure $\widetilde{P}$. Under $\widetilde{P}$, the diffusion X satisfies the law of motion

$$dX_t = \widetilde{\mu}(X_t)dt + \sigma(X_t)d\widetilde{W}_t,$$

$$\widetilde{\mu}(x) = \mu(x) + \sigma(x)\sigma'(x)\frac{\partial}{\partial x}\log e(x) + \sigma(x)\alpha(x)$$

and the associated generator

$$\widetilde{\mathbb{B}}f = \widetilde{\mu} \cdot \frac{\partial f}{\partial x} + \frac{1}{2}\mathrm{tr}\left(\sigma\sigma'\frac{\partial^2 f}{\partial x \partial x'}\right)$$

corresponds to the generator of a diffusion with infinitesimal variance $\sigma^2(x)$ and infinitesimal mean $\widetilde{\mu}(x)$ under $\widetilde{P}$.

[P] See the seminal work by Feller (1952) and Feller (1957). Karlin and Taylor (1981), Borodin and Salminen (2002), or Linetsky (2008) offer summarizing treatments.

The boundary characterization under $\widetilde{P}$ and the associated boundary conditions for $\psi_t(x)$ follow from formulas from Section 6.4. The character of the boundary can change under $\widetilde{P}$, although a reflecting boundary remains reflecting to preserve local equivalence of measures P and $\widetilde{P}$. Observe that Eq. (33) introduces a relationship between the conditional expectation given by $\phi_t(x)$ and the eigenfunction $e(x)$. For instance, when the boundary point x_b is reflecting, the appropriate boundary condition is[q]

$$\frac{\partial}{\partial x}\psi_t(x)\bigg|_{x=x_b}=0.$$

When both $\phi_t(x)$ and $e(x)$ are strictly positive at the boundary, this implies that

$$\frac{\partial}{\partial x}\log\phi_t(x)\bigg|_{x=x_b}=\frac{\partial}{\partial x}\log e(x)\bigg|_{x=x_b}$$

equalizing logarithmic slopes of the conditional expectation (28) and the eigenfunction $e(x)$ at the boundary.

7. MODELS WITH FINANCIAL CONSTRAINTS IN CONTINUOUS TIME

Recently, there has been renewed interest in nonlinear stochastic macroeconomic models with financing restrictions. The literature was initiated by Bernanke and Gertler (1989) and Bernanke et al. (1999), and it has been revived and extended since the advent of the financial crisis. Continuous-time models have been featured in He and Krishnamurthy (2013), Brunnermeier and Sannikov (2014), Di Tella (2015), Moreira and Savov (2016), Adrian and Boyarchenko (2012), or Klimenko et al. (2016). Differential equation methods give the equilibrium solutions, and the resulting dynamics exhibit quantitatively substantial nonlinearity. The nonlinearity emerges because of financing constraints that bind only in a specific part of the state space.[r]

To preserve tractability, models typically assume a low-dimensional specification of the state space. In this section, we analyze two such models, He and Krishnamurthy (2013) and Brunnermeier and Sannikov (2014). Both models utilize frameworks that are judiciously chosen to lead to a scalar endogenous state variable that follows the diffusion

$$dX_t=\mu(X_t)dt+\sigma(X_t)dW_t. \tag{34}$$

The endogenous state represents the allocation of wealth between households and financial experts, capturing the capitalization of the financial sector relative to the size of the

[q] This assumes that the so-called scale measure is finite at the boundary, see, eg, Borodin and Salminen (2002).

[r] See Bocola (2016) or Bianchi (2016) for discrete-time models solved using global to account for financing constraint that binds only occasionally.

economy. When the capitalization is low, the financial constraint is binding, and asset valuations are more sensitive to aggregate shocks.

Both papers also feature an exogenous process that introduces aggregate risk into their model economies. He and Krishnamurthy (2013) construct an endowment economy with a permanent shock to the aggregate dividend. On the other hand, Brunnermeier and Sannikov (2014) feature endogenous capital accumulation with a shock to the quality of the capital stock. In this section, we utilize the continuous-time tools developed in Section 6 to study the state dependence in asset pricing implications of the two models. We refer the reader to the respective papers for a detailed discussions of the underlying economic environments.

7.1 Stochastic Discount Factors

Stochastic discount factors and priced cash flows in the models we analyze can be written as special cases of multiplicative functionals introduced in Section 4.2:

$$d\log S_t = \beta(X_t)dt + \alpha(X_t) \cdot dW_t \tag{35}$$

with coefficients $\beta(x)$ and $\alpha(x)$ determined in equilibrium. In an arbitrage-free, complete market environment, there exists a unique stochastic discount factor that represents the prices of the traded securities.

In economies with financial market imperfections and constraints, this ceases to be true. There are two key features that are of interest to us. First, financial markets in these economies are segmented, and different investors can own specific subsets of assets. This implies the existence of alternative stochastic discount factors for individual investors that have to agree only on prices of assets traded between investors. Second, assets are valuable not only for their cash flows but also because their ownership can relax or tighten financing constraints faced by individual investors. Given the potential for these constraints to be binding, asset values include contributions from the shadow prices of these constraints.

7.2 He and Krishnamurthy (2013)

He and Krishnamurthy (2013) construct an economy populated by two types of agents, specialists and households. There are two assets in the economy, a safe asset earning an infinitesimal risk-free rate r_t and a risky asset with return R_t that is a claim on aggregate dividend

$$d\log D_t = \left(g_d - \frac{1}{2}\sigma_d^2\right)dt + \sigma_d dW_t \doteq \overline{\beta}_d dt + \overline{\alpha}_d dW_t. \tag{36}$$

7.2.1 Households and Specialists

Households have logarithmic preferences and therefore consume a constant fraction of their wealth, $C_t^h = \rho A_t^h$, where ρ is the time-preference coefficient. A fraction λ of households can only invest into the safe asset, while a fraction $1 - \lambda$ invests a share α_t^h of their wealth through an intermediary managed by the specialists who hold a portfolio with return $d\widetilde{R}_t$. Aggregate wealth of the households therefore evolves as

$$dA_t^h = \left(\ell D_t - \rho A_t^h\right)dt + A_t^h r_t dt + \alpha_t^h (1-\lambda) A_t^h \left(d\widetilde{R}_t - r_t dt\right),$$

where ℓD_t is households' income, modeled as a constant share ℓ of the dividend.

Specialists are endowed with CRRA preferences over their consumption stream C_t with risk aversion coefficient γ and trade both assets. Their stochastic discount factor is

$$\frac{S_t}{S_0} = e^{-\rho t}\left(\frac{C_t}{C_0}\right)^{-\gamma}. \tag{37}$$

This stochastic discount factor also prices all assets traded by specialists. The law of motion for their wealth is given by

$$dA_t = -C_t dt + A_t r_t dt + A_t\left(d\widetilde{R}_t - r_t dt\right).$$

The intermediary combines all wealth of the specialists A_t with the households' wealth invested through the intermediary $\alpha_t^h(1-\lambda)A_t^h$ and invests a share α_t of the combined portfolio into the risky asset. The return on the intermediary portfolio then follows

$$d\widetilde{R}_t = r_t dt + \alpha_t (dR_t - r_t dt).$$

The risky asset market clears, so that the wealth invested into the risky asset equals the market price of the asset, P_t

$$\alpha_t\left(A_t + \alpha_t^h(1-\lambda)A_t^h\right) = P_t.$$

7.2.2 Financial Friction

The critical financial friction is introduced into the portfolio choice of the household. Motivated by a moral hazard problem, the household is not willing to invest more than a fraction m of the specialists' wealth through the intermediary, which defines the *intermediation constraint*

$$\alpha_t^h(1-\lambda)A_t^h \le m A_t. \tag{38}$$

Because of logarithmic preferences, the portfolio choice α_t^h of the household is static. The household is also not allowed to sell short any of the assets, so that it solves

$$\max_{\alpha_t^h \in [0,1]} \alpha_t^h E\left[d\widetilde{R}_t - r_t dt \mid \mathcal{F}_t\right] - \frac{1}{2}\left(\alpha_t^h\right)^2 Var\left[d\widetilde{R}_t - r_t dt \mid \mathcal{F}_t\right]$$

subject to the intermediation constraint (38).

The parameter m determines the tightness of the intermediation constraint. This constraint will be endogenously binding when the wealth of the specialists becomes sufficiently low relative to the wealth of the household. In that case, risk sharing partially breaks down and the specialists will have to absorb a large share of the risky asset in their portfolio. As an equilibrium outcome, risk premia increase and the wealth of the specialists becomes more volatile, which in turn induces larger fluctuations of the right-hand side of the constraint (38). Without the intermediation constraint, the model reduces to an endowment economy populated by agents solving a risk-sharing problem with portfolio constraints.

7.2.3 Equilibrium Dynamics

The equilibrium in this model is conveniently characterized using the wealth share of the specialists, $X_t \doteq A_t / P_t \in (0,1)$, that will play the role of the single state variable with endogenously determined dynamics (34) where the coefficients $\mu(x)$ and $\sigma(x)$ are given by the relative wealth accumulation rates of households and specialists, and the equilibrium price of the claim on the risky cash flow. He and Krishnamurthy (2013) show that both boundaries $\{0,1\}$ are entrance boundaries.

Given the homogeneity in the model, we can write the consumption of the specialists as

$$\begin{aligned} C_t &= D_t(1+\ell) - C_t^h = D_t\left[(1+\ell) - \frac{C_t^h}{A_t^h}\frac{A_t^h}{P_t}\frac{P_t}{D_t}\right] \\ &= D_t[(1+\ell) - \rho(1-X_t)\pi(X_t)] \end{aligned}$$

where $\pi(x)$ is the price-dividend ratio for the claim on the dividend stream. The price-dividend ratio is determined endogenously as part of the solution to a set of differential equations. Given a solution for the price-dividend ratio $\pi(x)$, we construct the stochastic discount factor (37).

The top row of Fig. 1 shows the drift and volatility coefficients of the state variable process X, and the associated stationary density. When the specialists' wealth share X_t is low (below $x^* = 0.091$), the intermediation constraint binds. As $X_t \to 0$, the intermediation capacity of the specialists decreases, which increases the expected return on the risky asset, thereby increasing the rate of wealth accumulation of the specialists. On the other hand, when $X_t \to 1$, the economy is unconstrained, risk premia are low, and situation reverses. The drift coefficient $\mu(x)$ in the top left panel reflects these effects.

In the moment when the constraints start binding (to the left of the point $x^* = 0.091$), volatility $\sigma(x)$ of the experts' wealth share starts rising. Ultimately, this volatility has to decline to zero as $X_t \to 0$ to prevent the experts' wealth share from hitting the zero boundary with a positive probability, but the volatility of experts' wealth *level* keeps rising as we approach the boundary.

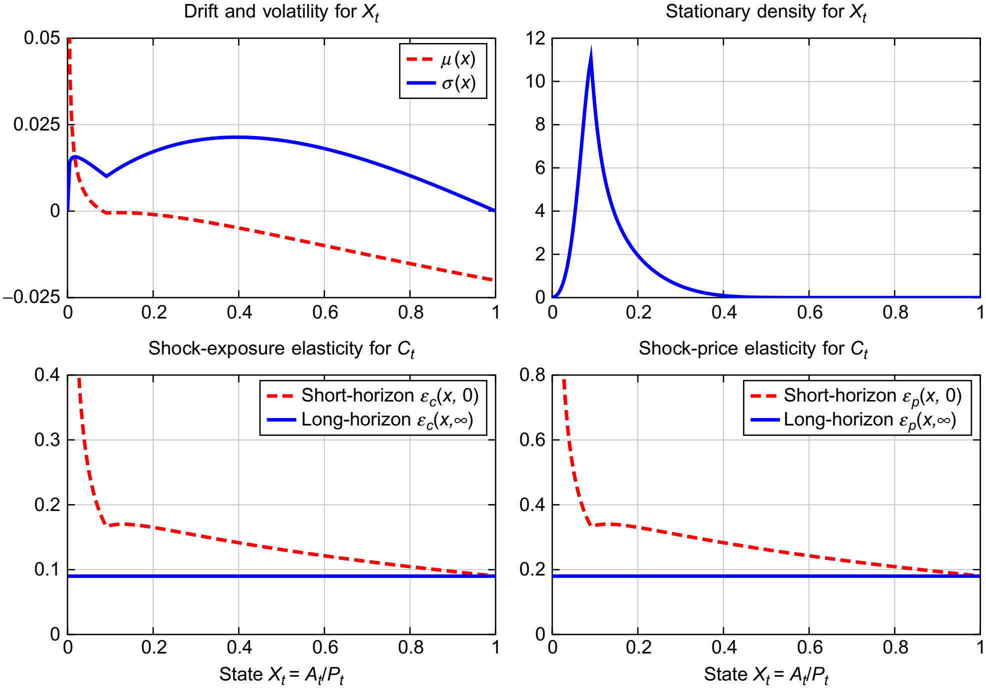


Fig. 1 Dynamics of the experts' wealth share $X_t = A_t/P_t$ (*horizontal axis*), shock-exposure and shock-price elasticities for the He and Krishnamurthy (2013) model. *Top left panel* shows the drift and volatility coefficients for the evolution of X_t, while *top right panel* the stationary density for X_t. *Panels in the bottom row* show the short- and long-horizon shock elasticity for the experts' consumption process C_t. The intermediation constraint (38) binds in the interval $X_t \in (0, 0.091)$, and $x^* = 0.091$ corresponds to the 35.3% quantile of the stationary distribution of X_t.

7.2.4 Stochastic Discount Factor and Cash Flows

Aggregate dividend D_t in (36) follows a geometric Brownian motion with drift. This directly implies a constant shock-exposure elasticity

$$\varepsilon_d(x,t) = \sigma_d.$$

Time variation in expected returns on the claim on the aggregate dividend thus must come solely from the time variation in prices of risk. In particular, the consumption process of specialists is:

$$\frac{C_t}{C_0} = \left(\frac{D_t}{D_0}\right)\left[\frac{(1+\ell) - \rho(1-X_t)\pi(X_t)}{(1+\ell) - \rho(1-X_0)\pi(X_0)}\right]. \tag{39}$$

Notice that the consumption of specialists has the same long-term stochastic growth as the aggregate dividend process. Since the dividend process D is a geometric Brownian motion, we immediately obtain the martingale factorization of C with

$$\begin{aligned} e_c(x) &= [(1+\ell) - \rho(1-x)\pi(x)]^{-1} \\ \eta_c &= g_d \\ \tilde{C}_t &= \exp(-\eta_c t) D_t \end{aligned}$$

where $\tilde{C}$ is the martingale component of C. Analogously, the stochastic discount factor of the specialists (37) is decomposed as

$$\begin{aligned} e_s(x) &= [(1+\ell) - \rho(1-x)\pi(x)]^{\gamma} \\ \eta_s &= -\rho - \gamma g_d + \frac{1}{2}\sigma_d^2 \gamma(\gamma+1) \\ \tilde{S}_t &= \exp[(-\eta_s - \rho)t](D_t)^{-\gamma} \end{aligned}$$

where $\tilde{S}$ is the martingale component.

These factorization results indicate a simple form for the long-horizon limits of the shock elasticities. The consumption and dividend processes share the same martingale component, and thus, assuming $\nu(x) = 1$, their shock-exposure elasticities imply

$$\lim_{t\to\infty} \varepsilon_c(x,t) = \lim_{t\to\infty} \varepsilon_d(x,t) = \sigma_d.$$

Similarly, the shock-price elasticities for the two cash-flow processes have the common long-horizon limit

$$\lim_{t\to\infty} \varepsilon_p(x,t) = \gamma\sigma_d.$$

As we have just verified, the intermediation constraint does not have any impact on prices of long-horizon cash flows. Long-horizon shock elasticities behave as in an economy populated only by unconstrained specialists with risk aversion γ who consume the whole dividend stream D_t. The intermediation constraint only affects the stationary part $e_s(x)$ of the stochastic discount factor.[s] As a consequence, long-term risk adjustments in this model are the same as those implied by a model with power utility function and consumption equal to dividends. The financing constraint induces deviations in short-term risk prices, which we now characterize.

7.2.5 Shock Elasticities and Term Structure of Yields

The blue solid lines in the bottom row of Fig. 1 represent the long-horizon shock-exposure and shock-price elasticities. These results are contrasted with the infinitesimal shock-exposure and shock-price elasticities, depicted with red dashed lines, that are equal to the volatility coefficients $\alpha_c(x)$ and $\alpha_s(x)$ in the differential representation (35) for the experts' consumption process (39) and stochastic discount factor process (37), respectively.

[s] Without the intermediation constraint and the debt constraint ($\lambda = 0$), the economy reduces to a complete-market risk-sharing problem between households and specialists and will converge in the long run to a homogeneous-agent economy populated only by households when $\gamma > 1$.

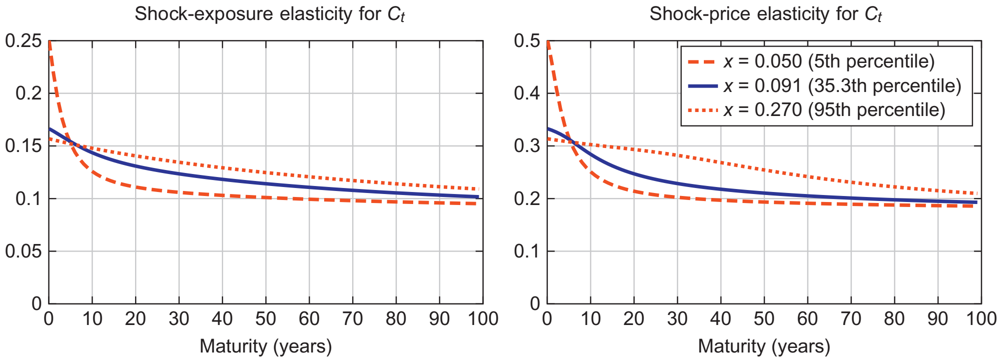


Fig. 2 Shock-exposure and shock-price elasticities for the He and Krishnamurthy (2013) model. *Individual lines* correspond to alternative choices of the current state, the experts' wealth share $X_0 = x$. The *solid line* represents the state in which the intermediation constraint (38) starts binding ($x = 0.091$), corresponding to the 35.3% quantile of the stationary distribution of X_t. The *dashed line* corresponds to the 5% quantile of the stationary distribution of X_t (intermediation constraint tightly binding), while the *dotted line* corresponds to the 95% quantile.

Fig. 2 depicts these shock elasticities evaluated at three different points in the state space. These elasticities were computed numerically.[t] A remarkable feature of the model is the following. The short-horizon consumption cash flows are more exposed to risk as revealed by a larger shock-price elasticity in the constrained region of the state space ($x = 0.05$). This finding is reversed for long-horizon cash flows, showing that the term structure of risk prices is much more strongly downward sloping for low values of the state variable. Since the state variable responds positively to shocks, low realizations of the state variable are the consequence of adverse shocks in the past.

Fig. 3 explores the implications for yields on dividends and experts' consumptions for alternative payoff horizons computed as logarithms of expected returns to the respective payoffs. While the yields on dividends and experts' consumption are initially increasing in maturity, this is all the more so when x is low. The yields are monotone over all horizons except when x is low, in which case the yields eventually decline a bit. The same effect is even more pronounced for the risk-free yield curve except the eventual decline is even slighter. Excess yields are therefore downward sloping for the experts' consumption process, and are lower for longer maturities for low values of x in contrast to high values.[u]

[t] We solved Eq. (29) for $M = C$ and $M = SC$, with $\phi_0(x) = 1$ using an implicit finite difference scheme. We used the solution for $\pi(x)$ constructed using the code from He and Krishnamurthy (2013).

[u] For empirical evidence and modeling of the downward sloping term structure of risky yields see van Binsbergen et al. (2012, 2013), Ai et al. (2013), Belo et al. (2015), Hasler and Marfè (2015), Lopez et al. (2015), or van Binsbergen and Koijen (2016).

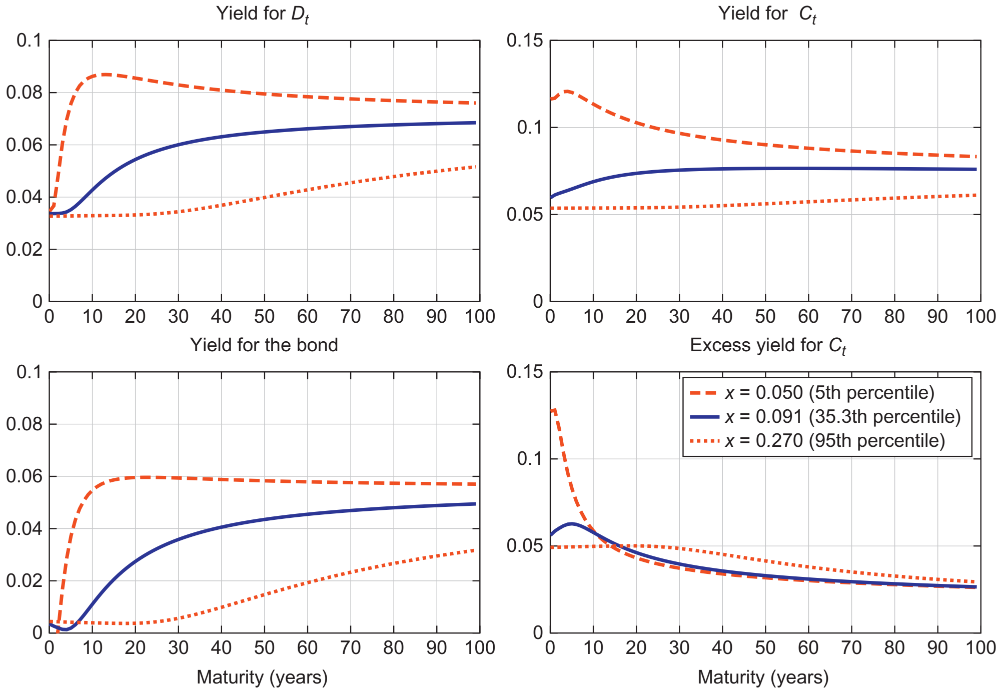


Fig. 3 Yields and excess yields for the He and Krishnamurthy (2013) model. Parameterization and description as in Fig. 2.

7.3 Brunnermeier and Sannikov (2014)

Brunnermeier and Sannikov (2014) construct a model with endogenous capital accumulation, populated by two types of agents, households and experts. The experts have access to a more productive technology for output and new capital than the households. The state variable of interest is the wealth share of experts, defined as

$$X_t = \frac{N_t}{Q_t K_t}$$

where N_t is the net worth of the experts and $Q_t K_t$ is the market value of capital. The equilibrium stock of capital evolves as

$$d \log K_t = \beta_k(X_t) dt + \overline{\alpha}_k dW_t$$

where the rate of accumulation of aggregate capital, $\beta_k(X_t)$, is determined by the wealth share of experts along with a standard local lognormal adjustment. The shock dW_t alters the quality of the capital stock.

7.3.1 Households and Experts

In the baseline model, both households and experts have linear preferences and differ in their time-preference coefficients, r and ρ, respectively, assuming that $\rho > r$. In particular, the preferences for experts are given by

$$E\left[\int_0^\infty e^{-\rho t} d\mathcal{C}_t \mid \mathcal{F}_0\right]$$

where Cu_t is the cumulative consumption and as such is restricted to be a nondecreasing process. In contrast, the cumulative consumption of the household can have negative increments. The linearity in their preferences implies a constant equilibrium rate of interest r.

7.3.2 Financial Friction

In the model, experts are better at managing the capital stock, making it more productive. This creates a natural tendency to move the capital from the hands of the households to the hands of the experts, who in turn issue financial claims on this capital to the households. Absent any financial frictions, the experts would instantly consume the total value of their own net worth (given their higher impatience and linear utility), and accept households' capital under management by issuing equity claims.

Brunnermeier and Sannikov (2014) assume that experts cannot issue any equity and have to finance all capital purchases using risk-free borrowing. This naturally creates a leveraged portfolio on the side of the experts. When the wealth share of experts X_t decreases, they can intermediate households' capital only by increasing their leverage, and the price of capital $Q(X_t)$ has to fall in order to generate a sufficiently high expected return on capital for the experts to hold this leveraged portfolio.

7.3.3 Equilibrium Dynamics

In equilibrium, the expected return on capital has to balance the hedging demand on the side of the experts with the supply of capital from households. Experts' hedging motive (limited willingness to hold a leveraged portfolio) arises from the fact that a leveraged portfolio generates a low return after an adverse realization of the shock dW_t which, at the same time, decreases X_t and therefore increases the future expected return on capital.

On the other hand, when the wealth share of experts X_t increases, the price of capital $Q(X_t)$ increases, and the expected return falls. Define the marginal value of experts' wealth $\Theta_t = \theta(X_t)$ through

$$\Theta_t N_t = E\left[\int_t^\infty e^{-\rho(s-t)} d\mathcal{C}_s \mid \mathcal{F}_t\right]$$

where $d\mathcal{C}$ is the cumulative consumption process of the experts. Linearity of preferences implies that experts' consumption is zero as long as $\Theta_t > 1$. As X_t increases, it reaches an endogenously determined threshold $\bar{x}$ for which $\theta(\bar{x}) = 1$. At this point, the marginal utility of wealth equals the marginal utility of consumption, and experts consume out of their wealth. Consequently, the equilibrium dynamics for the wealth share of experts is given by

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t - X_t d\zeta_t,$$

where $\mu(x)$ and $\sigma(x)$ are endogenously determined coefficients that depend on relative rates of wealth accumulation of experts and households, and the consumption rate of experts $d\zeta_t \doteq d\mathcal{C}_t / N_t > 0$ only if $X_t = \bar{x}$. Formally, the right boundary for the stochastic process X_t behaves as a reflecting boundary. See Brunnermeier and Sannikov (2014) for the construction of μ and σ.

7.3.4 Stochastic Discount Factor and Cash Flows

We now turn to the study of asset pricing implications in the model. To construct the shock elasticities, we construct the coefficients $\beta(x)$ and $\alpha(x)$ for the evolution of the stochastic discount factor and priced cash flows modeled as multiplicative functionals (35).

The marginal utility of wealth implies the following stochastic discount factor of the experts:

$$\frac{S_t}{S_0} = \exp(-\rho t)\frac{\theta(X_t)}{\theta(X_0)}.$$

The coefficients $\beta_s(x)$ and $\alpha_s(x)$ in the equation for the evolution of the stochastic discount factor functional can be constructed by applying Ito's lemma to this expression taking account of the functional dependence given by $\theta(x)$ and the evolution of X. Observe that this stochastic discount factor does not contain a martingale component. Nevertheless, since the equilibrium local risk-free interest rate is r,

$$\exp(rt)\frac{S_t}{S_0} = \exp[(r-\rho)t]\frac{\theta(X_t)}{\theta(X_0)}$$

must be a positive local martingale. As such, its expectation conditioned on date t information could decline in t implying that long-term interest rates could be higher and in fact converge to ρ. More generally, from the standpoint of valuation, the fat right tail of the process $\theta(X_t)$ could have important consequences for valuation even in the absence of a martingale component for the stochastic discount factor process.

As a priced cash flow, we consider the aggregate consumption flow process C^a given by

$$C_t^a = [a_e\psi(X_t) + a_h[1-\psi(X_t)] - \iota(X_t)]K_t \tag{40}$$

where $\iota(x)$ is the aggregate investment rate, $\psi(x)$ is the fraction of the capital stock owned by the experts, and $a_e > a_h$ are the output productivities of the experts and households, respectively. Thus C_t^a is equal to aggregate output net of aggregate investment. Aggregate consumption is therefore given as a stationary fraction of aggregate capital. Thus aggregate consumption flow and capital stock processes share a common martingale component.[v]

7.3.5 Shock Elasticities and Term Structure of Yields

The top left panel in Fig. 4 depicts the drift and volatility coefficients for the state variable X_t. At the right boundary $\bar{x}$, the experts accumulated a sufficiently large share of capital and start consuming. Given their risk neutrality, the boundary behaves as a reflecting boundary.

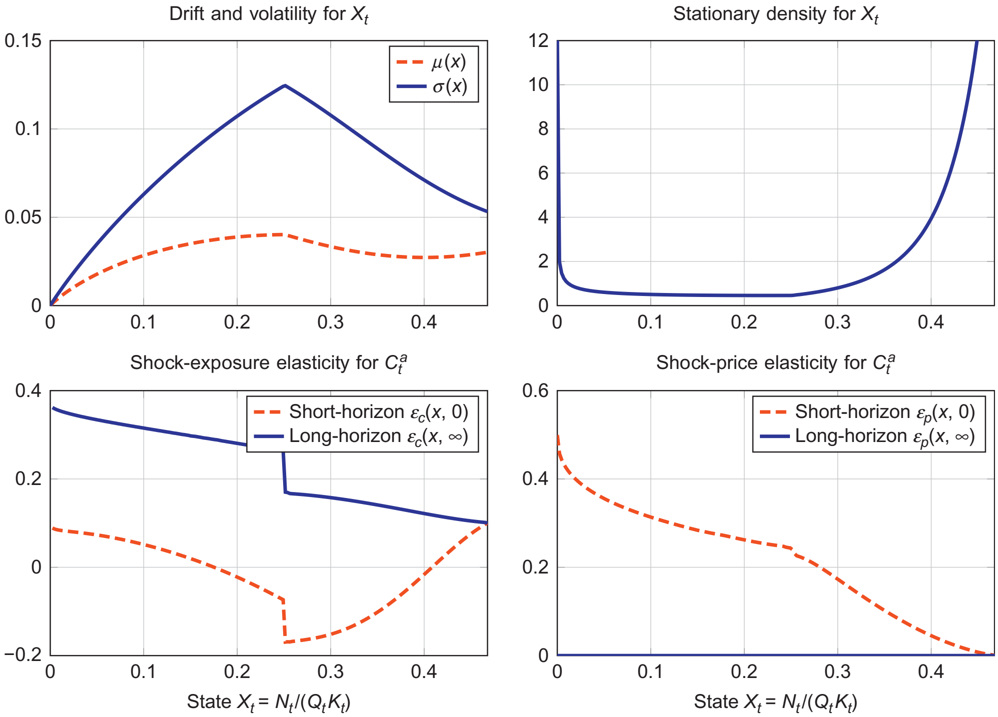


Fig. 4 Dynamics of the experts' wealth share $X_t = N_t/(Q_tK_t)$ (*horizontal axis*), shock-exposure and shock-price elasticities for the Brunnermeier and Sannikov (2014) model. *Top left panel* shows the drift and volatility coefficients for the evolution of X_t, while *top right panel* the stationary density for X_t. *Panels in the bottom row* show the short- and long-horizon shock elasticity for the aggregate consumption process C^a. The intermediation constraint binds in the interval $X_t \in (0, 0.25)$, and $x^* = 0.25$ corresponds to the 15% quantile of the stationary distribution of X.

[v] Brunnermeier and Sannikov (2014) also consider an extension where experts and households are endowed with logarithmic utilities. In that case consumption of both households and experts is given as constant fractions of their respective net worth, and the stochastic discount factor of the experts inherits the martingale component from the reciprocal of the aggregate capital process.

At the left boundary, the situation is notably different. Experts' ability to intermediate capital is limited by their own net worth, and hence their portfolio choice corresponds to an effectively risk averse agent. The left boundary is natural and nonattracting.

The existence of a stationary distribution, depicted in the second panel of Fig. 4, arises from a combination of two forces. Experts are more impatient, so whenever they accumulate a sufficient share of capital, they start consuming, which prevents them from taking over the whole economy. On the other hand, when their wealth share falls, their intermediation ability becomes scarce, the expected return on capital rises, and they use their superior investment technology to accumulate wealth at a faster rate than households.

The stationary density has peaks at each of the two boundaries. The positive drift coefficient $\mu(x)$ implies that there is a natural pull toward the right boundary, creating the peak in the density there. However, whenever a sequence of shocks brings the economy close to the left boundary, solvency constraints imply that it takes time for experts to accumulate wealth again, and the economy spends a long period time in that part of the state space. Economically, most times are "good" times when intermediation is fully operational, with rare periods of protracted "financial crises."

The bottom row of Fig. 4 plots the shock elasticities for the aggregate consumption process (40). Observe that the short-horizon exposure elasticity is negative in a part of the state space, making aggregate consumption countercyclical there. The long-horizon elasticities are noticeably higher, and particularly high when the intermediation constraint binds. The discontinuity at $X_t = x^*$ is caused by the change in consumption behavior in the moment when the intermediation constraint starts binding.

Given that the stochastic discount factor has no martingale component, the long-horizon shock-price elasticity is zero. On the other hand, the short-horizon price of risk varies strongly with the wealth share of the experts. This state dependence is also confirmed in Fig. 5 which plots the shock elasticity functions for selected points in the state space. Shock-exposure elasticities for the aggregate consumption process $\{C_t^a : t \geq 0\}$ increase with maturity, while the shock-price elasticities vanish as $t \to \infty$. Notice that there is a sign reversal in the exposure elasticities for aggregate consumption. The shock-exposure elasticities are initially negative but eventually become positive in the middle part of the state space, mirroring the bottom left panel of Fig. 4. This pattern emerges because the equilibrium investment responses over short horizons lead to more substantial longer-term consumption responses in the constrained states. Nevertheless, the shock-price elasticities are positive for all horizons and states that we consider.

Finally, Fig. 6 plots the yields on risk-free bonds and claims on horizon-specific cash flows from aggregate consumption. In line with the nonmonotonicity of the shock-exposure elasticities across states in Fig. 4, the short-maturity yields are also nonmonotonic, being lowest, and in fact lower than the risk-free rate, in the center of the distribution of the state X_t.

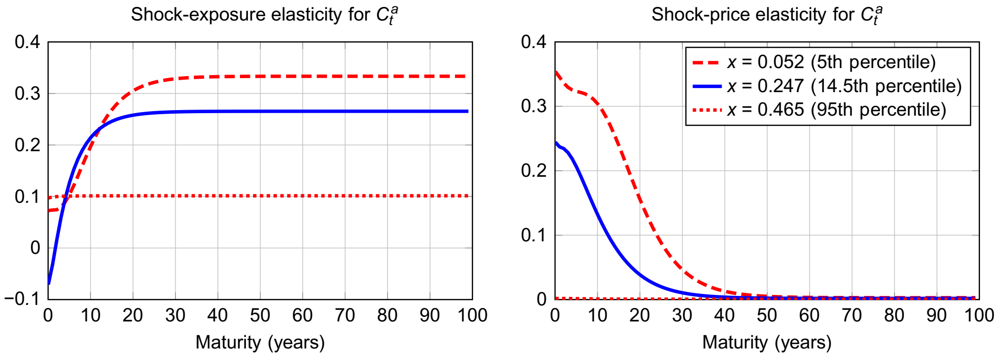


Fig. 5 Shock-exposure and shock-price elasticities for the Brunnermeier and Sannikov (2014) model. *Individual lines* correspond to alternative choices of the current state, the experts' wealth share $X_0 = x$. The *solid line* represents the state in which the intermediation constraint starts binding ($x = 0.247$), corresponding to the 14.5% quantile of the stationary distribution of X_t. The *dashed line* corresponds to the 5% quantile of the stationary distribution of X_t (intermediation constraint tightly binding), while the *dotted line* corresponds to the 95% quantile.

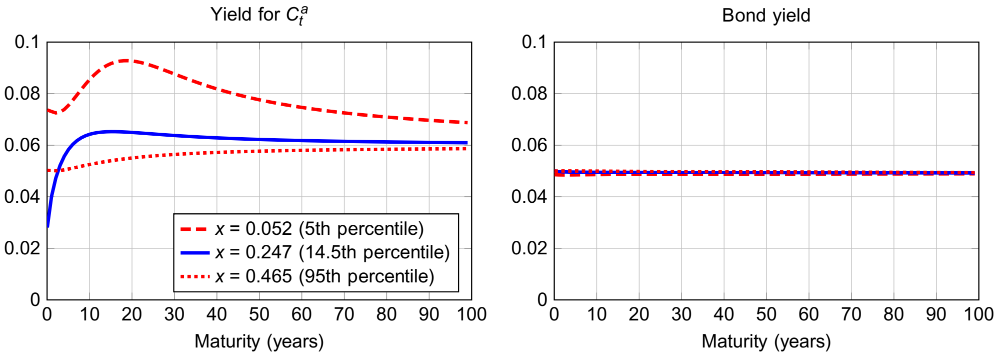


Fig. 6 Yields and excess yields for the Brunnermeier and Sannikov (2014) model. Parameterization and description as in Fig. 5.

8. DIRECTIONS FOR FURTHER RESEARCH

In this chapter, we developed dynamic value decompositions (DVDs) for the study of intertemporal asset pricing implications of dynamic equilibrium models. We constructed *shock elasticities* as building blocks for these decompositions. The DVD methods are distinct but potentially complementary to the familiar Campbell and Shiller (1988) decomposition. Campbell and Shiller use linear VAR methods to quantify the impact of (discounted) "cash flow shocks" and "expected return shocks" on price-dividend ratios. In general these shocks are correlated and are themselves combinations of shocks that are fundamental to structural models of the macroeconomy. Our aim is to explore pricing implications of models in which alternative macroeconomic shocks are identified and their impact quantified. We replaced linear approximation with local sensitivity analysis, and we characterized how cash flows are exposed to alternative macroeconomic shocks and what the corresponding price adjustments are for these exposures. We showed that shock elasticities are mathematically and economically related to impulse response functions. The shock elasticities represent sensitivities of expected cash flows to alternative macroeconomic shocks and the associated market implied compensations when looking across differing investment horizons.

We apply these DVD methods to a class of dynamic equilibrium models that feature financial frictions and segmented markets. The methods uncover the ways financial frictions contribute to pricing of alternative cash flows and to the shape of the term structure of macroeconomic risk prices.

There are two extensions of our analysis that require further investigation. First, risk prices are only well defined relative to an underlying probability distribution. In this chapter, we have not discussed the consequences for pricing when investors inside our models use different probability measures than the data-generating measure presumed by an econometrician. Typically, researchers invoke an assumption of rational expectations to connect investor perceptions with the data generation. More generally, models of investors that allow for subjective beliefs, learning, ambiguity aversion, or concerns about model misspecification alter how we interpret market-based compensations for exposure to macroeconomic fluctuations. For instance, see Hansen (2014) for further discussion. Incorporating potential belief distortions into the analysis should be a valuable extension of these methods.

Second, we left aside empirical and econometric aspects of the identification of shocks and measurement of risk premia. The empirical finance literature has made considerable progress in the characterization and measurement of the term structure of risk premia in various asset markets. The challenge for model building is to connect these empirical facts to specific sources of macroeconomic risks and financial market frictions of model economies. Our methodology suggests a way to make these connections, but further investigation is required.

Finally, we refrained from the discussion of implications for policy analysis. Financial frictions create economic externalities that can potentially be rectified by suitable policy actions. Since asset prices enter these financial constraints, understanding their behavior is an important ingredient to meaningful policy design. Forward looking asset prices provide both a source of information about private sector beliefs and an input into the regulatory challenges faced in the conduct of policy. Our methods can help to uncover asset pricing implications for alternative potential policies.

APPENDICES

Appendix A Exponential-Quadratic Framework

Let $X = (X_1', X_2')'$ be a $2n \times 1$ vector of states, $W \sim N(0, I)$ a $k \times 1$ vector of independent Gaussian shocks, and $\mathcal{F}_t$ the filtration generated by $(X_0, W_1, \ldots, W_t)$. In this appendix, we show that given the law of motion from Eq. (21)

$$\begin{aligned} X_{1,t+1} &= \Theta_{10} + \Theta_{11} X_{1,t} + \Lambda_{10} W_{t+1} \\ X_{2,t+1} &= \Theta_{20} + \Theta_{21} X_{1,t} + \Theta_{22} X_{2,t} + \Theta_{23} (X_{1,t} \otimes X_{1,t}) \\ &\quad + \Lambda_{20} W_{t+1} + \Lambda_{21} (X_{1,t} \otimes W_{t+1}) + \Lambda_{22} (W_{t+1} \otimes W_{t+1}) \end{aligned} \tag{A.1}$$

and a multiplicative functional $M_t = \exp(Y_t)$ whose additive increment is given in Eq. (22):

$$\begin{aligned} Y_{t+1} - Y_t &= \Gamma_0 + \Gamma_1 X_{1,t} + \Gamma_2 X_{2,t} + \Gamma_3 (X_{1,t} \otimes X_{1,t}) \\ &\quad + \Psi_0 W_{t+1} + \Psi_1 (X_{1,t} \otimes W_{1,t+1}) + \Psi_2 (W_{t+1} \otimes W_{t+1}), \end{aligned} \tag{A.2}$$

we can write the conditional expectation of M as

$$\log E[M_t \mid \mathcal{F}_0] = (\overline{\Gamma}_0)_t + (\overline{\Gamma}_1)_t X_{1,0} + (\overline{\Gamma}_2)_t X_{2,0} + (\overline{\Gamma}_3)_t (X_{1,0} \otimes X_{1,0}) \tag{A.3}$$

where $(\overline{\Gamma}_i)_t$ are constant coefficients to be determined.

The dynamics given by (A.1) and (A.2) embed the perturbation approximation constructed in Section 5.2 as a special case. The Θ and Λ matrices needed to map the perturbed model into the above structure are constructed from the first and second derivatives of the function $\psi(x, w, \mathsf{q})$ that captures the law of motion of the model, evaluated at $(\bar{x}, 0, 0)$:

$$\begin{array}{llll} \Theta_{10} = \psi_q & \Theta_{11} = \psi_x & \Lambda_{10} = \psi_w & \\ \Theta_{20} = \psi_{qq} & \Theta_{21} = 2\psi_{xq} & \Theta_{22} = \psi_x & \Theta_{23} = \psi_{xx} \\ \Lambda_{20} = 2\psi_{wq} & \Lambda_{21} = 2\psi_{xw} & \Lambda_{22} = \psi_{ww} & \end{array}$$

where the notation for the derivatives is defined in Appendix A.2.

A.1 Definitions

To simplify work with Kronecker products, we define two operators vec and $\text{mat}_{m,n}$. For an $m \times n$ matrix H, $\text{vec}(H)$ produces a column vector of length mn created by stacking the columns of H:

$$h_{(j-1)m+i} = [\text{vec}(H)]_{(j-1)m+i} = H_{ij}.$$

For a vector (column or row) h of length mn, $\text{mat}_{m,n}(h)$ produces an $m \times n$ matrix H created by "columnizing" the vector:

$$H_{ij} = [\mathrm{mat}_{m,n}(h)]_{ij} = h_{(j-1)m+i}.$$

We drop the m, n subindex if the dimensions of the resulting matrix are obvious from the context. For a square matrix A, define the sym operator as

$$\mathrm{sym}(A) = \frac{1}{2}(A + A').$$

Apart from the standard operations with Kronecker products, notice that the following is true. For a row vector $H_{1\times nk}$ and column vectors $X_{n\times 1}$ and $W_{n\times 1}$

$$H(X \otimes W) = X'[\mathrm{mat}_{k,n}(H)]' W$$

and for a matrix $A_{n\times k}$, we have

$$X'AW = (\mathrm{vec}A')'(X \otimes W). \tag{A.4}$$

Also, for $A_{n\times n}$, $X_{n\times 1}$, $K_{k\times 1}$, we have

$$\begin{aligned}(AX) \otimes K &= (A \otimes K)X\\ K \otimes (AX) &= (K \otimes A)X.\end{aligned}$$

Finally, for column vectors $X_{n\times 1}$ and $W_{k\times 1}$,

$$(AX) \otimes (BW) = (A \otimes B)(X \otimes W)$$

and

$$(BW) \otimes (AX) = \left[B \otimes A_{\bullet j}\right]_{j=1}^{n} (X \otimes W)$$

where

$$\left[B \otimes A_{\bullet j}\right]_{j=1}^{n} = [B \otimes A_{\bullet 1} \quad B \otimes A_{\bullet 2} \quad \dots \quad B \otimes A_{\bullet n}].$$

A.2 Concise Notation for Derivatives

Consider a vector function $f(x, w)$ where x and w are column vectors of length m and n, respectively. The first-derivative matrix f_i where $i = x$, w is constructed as follows. The kth row $[f_i]_{k\bullet}$ corresponds to the derivative of the kth component of f

$$[f_i(x,w)]_{k\bullet} = \frac{\partial f^{(k)}}{\partial i'}(x,w).$$

Similarly, the second-derivative matrix is the matrix of vectorized and stacked Hessians of individual components with kth row

$$\left[f_{ij}(x,w)\right]_{k\bullet} = \left(\mathrm{vec}\frac{\partial^2 f^{(k)}}{\partial j \partial i'}(x,w)\right)'.$$

It follows from formula (A.4) that, for example,

$$x'\left(\frac{\partial^2 f^{(k)}}{\partial x \partial w'}(x,w)\right)w = \left(\operatorname{vec}\frac{\partial^2 f^{(k)}}{\partial w \partial x'}(x,w)\right)'(x\otimes w) = [f_{xw}(x,w)]_{k\bullet}(x\otimes w).$$

A.3 Conditional Expectations

Notice that a complete-the-squares argument implies that, for a $1\times k$ vector A, a $1\times k^2$ vector B, and a scalar function $f(w)$,

$$\begin{aligned} &E[\exp(B(W_{t+1}\otimes W_{t+1}) + AW_{t+1})f(W_{t+1})\,|\,\mathcal{F}_t] \\ &\quad = E\left[\exp\left(\frac{1}{2}W'_{t+1}(\mathrm{mat}_{k,k}(2B))W_{t+1} + AW_{t+1}\right)f(W_{t+1})\,|\,\mathcal{F}_t\right] \\ &\quad = |I_k - \mathrm{sym}[\mathrm{mat}_{k,k}(2B)]|^{-1/2}\exp\left(\frac{1}{2}A(I_k - \mathrm{sym}[\mathrm{mat}_{k,k}(2B)])^{-1}A'\right)\widetilde{E}\,[f(W_{t+1})\,|\,\mathcal{F}_t] \end{aligned} \tag{A.5}$$

where $\widetilde{\cdot}$ is a measure under which

$$W_{t+1}\sim N\big((I_k - \mathrm{sym}[\mathrm{mat}_{k,k}(2B)])^{-1}A', (I_k - \mathrm{sym}[\mathrm{mat}_{k,k}(2B)])^{-1}\big).$$

We start by utilizing formula (A.5) to compute

$$\begin{aligned} \bar{Y}(X_t) &= \log E[\exp(Y_{t+1} - Y_t)\,|\,\mathcal{F}_t] = \Gamma_0 + \Gamma_1 X_{1,t} + \Gamma_2 X_{2,t} + \Gamma_3(X_{1,t}\otimes X_{1,t}) \\ &\qquad + \log E\left[\exp\left(\left[\Psi_0 + X'_{1t}[\mathrm{mat}_{k,n}(\Psi_1)]'\right]W_{t+1} + \frac{1}{2}W'_{t+1}[\mathrm{mat}_{k,k}(\Psi_2)]W_{t+1}\right)\,|\,\mathcal{F}_t\right] \\ &= \Gamma_0 + \Gamma_1 X_{1,t} + \Gamma_2 X_{2,t} + \Gamma_3(X_{1,t}\otimes X_{1,t}) \\ &\quad -\frac{1}{2}\log|I_k - \mathrm{sym}[\mathrm{mat}_{k,k}(2\Psi_2)]| + \frac{1}{2}\mu'(I_k - \mathrm{sym}[\mathrm{mat}_{k,k}(2\Psi_2)])^{-1}\mu \end{aligned}$$

with μ defined as

$$\mu = \Psi'_0 + [\mathrm{mat}_{k,n}(\Psi_1)]X_{1,t}.$$

Reorganizing terms, we obtain

$$\bar{Y}(X_t) = \overline{\Gamma}_0 + \overline{\Gamma}_1 X_{1,t} + \overline{\Gamma}_2 X_{2,t} + \overline{\Gamma}_3(X_{1,t}\otimes X_{1,t}) \tag{A.6}$$

where

$$\begin{aligned} \overline{\Gamma}_0 &= \Gamma_0 - \frac{1}{2}\log|I_k - \mathrm{sym}[\mathrm{mat}_{k,k}(2\Psi_2)]| + \frac{1}{2}\Psi_0(I_k - \mathrm{sym}[\mathrm{mat}_{k,k}(2\Psi_2)])^{-1}\Psi'_0 \\ \overline{\Gamma}_1 &= \Gamma_1 + \Psi_0(I_k - \mathrm{sym}[\mathrm{mat}_{k,k}(2\Psi_2)])^{-1}[\mathrm{mat}_{k,n}(\Psi_1)] \\ \overline{\Gamma}_2 &= \Gamma_2 \\ \overline{\Gamma}_3 &= \Gamma_3 + \frac{1}{2}\mathrm{vec}\left[[\mathrm{mat}_{k,n}(\Psi_1)]'(I_k - \mathrm{sym}[\mathrm{mat}_{k,k}(2\Psi_2)])^{-1}[\mathrm{mat}_{k,n}(\Psi_1)]\right]'. \end{aligned} \tag{A.7}$$

For the set of parameters $\mathcal{P} = (\Gamma_0, \ldots, \Gamma_3, \Psi_0, \ldots, \Psi_2)$, Eqs. (A.7) define a mapping

$$\bar{\mathcal{P}} = \bar{\mathcal{E}}(\mathcal{P}),$$

with all $\overline{\Psi}_j = 0$. We now substitute the law of motion for X_1 and X_2 to produce $\bar{Y}(X_t) = \widetilde{Y}(X_{t-1}, W_t)$. It is just a matter of algebraic operations to determine that

$$\begin{aligned}\widetilde{Y}(X_{t-1}, W_t) &= \log E[\exp(Y_{t+1} - Y_t) \mid \mathcal{F}_t] \\ &= \widetilde{\Gamma}_0 + \widetilde{\Gamma}_1 X_{1,t-1} + \widetilde{\Gamma}_2 X_{2,t-1} + \widetilde{\Gamma}_3 (X_{1,t-1} \otimes X_{1,t-1}) \\ &\quad + \widetilde{\Psi}_0 W_t + \widetilde{\Psi}_1 (X_{1,t-1} \otimes W_t) + \widetilde{\Psi}_2 (W_t \otimes W_t)\end{aligned}$$

where

$$\begin{aligned}\widetilde{\Gamma}_0 &= \overline{\Gamma}_0 + \overline{\Gamma}_1 \Theta_{10} + \overline{\Gamma}_2 \Theta_{20} + \overline{\Gamma}_3 (\Theta_{10} \otimes \Theta_{10}) \\ \widetilde{\Gamma}_1 &= \overline{\Gamma}_1 \Theta_{11} + \overline{\Gamma}_2 \Theta_{21} + \overline{\Gamma}_3 (\Theta_{10} \otimes \Theta_{11} + \Theta_{11} \otimes \Theta_{10}) \\ \widetilde{\Gamma}_2 &= \overline{\Gamma}_2 \Theta_{22} \\ \widetilde{\Gamma}_3 &= \overline{\Gamma}_2 \Theta_{23} + \overline{\Gamma}_3 (\Theta_{11} \otimes \Theta_{11}) \\ \widetilde{\Psi}_0 &= \overline{\Gamma}_1 \Lambda_{10} + \overline{\Gamma}_2 \Lambda_{20} + \overline{\Gamma}_3 (\Theta_{10} \otimes \Lambda_{10} + \Lambda_{10} \otimes \Theta_{10}) \\ \widetilde{\Psi}_1 &= \overline{\Gamma}_2 \Lambda_{21} + \overline{\Gamma}_3 \left(\Theta_{11} \otimes \Lambda_{10} + \left[\Lambda_{10} \otimes (\Theta_{11})_{\bullet j} \right]_{j=1}^{n} \right) \\ \widetilde{\Psi}_2 &= \overline{\Gamma}_2 \Lambda_{22} + \overline{\Gamma}_3 (\Lambda_{10} \otimes \Lambda_{10}).\end{aligned} \tag{A.8}$$

This set of equations defines the mapping

$$\widetilde{\mathcal{P}} = \widetilde{\mathcal{E}}(\bar{\mathcal{P}}).$$

A.4 Iterative Formulas

We can write the conditional expectation in (A.3) recursively as

$$\log E[M_t \mid \mathcal{F}_0] = \log E\left[\exp(Y_1 - Y_0) E\left[\frac{M_t}{M_1} \mid \mathcal{F}_1\right] \mid \mathcal{F}_0\right].$$

Given the mappings $\bar{\mathcal{E}}$ and $\widetilde{\mathcal{E}}$, we can therefore express the coefficients $\bar{\mathcal{P}}$ in (A.3) using the recursion

$$\bar{\mathcal{P}}_t = \bar{\mathcal{E}}\left(\mathcal{P} + \widetilde{\mathcal{E}}(\bar{\mathcal{P}}_{t-1})\right)$$

where the addition is by coefficients and all coefficients in $\bar{\mathcal{P}}_0$ are zero matrices.

A.5 Coefficients Φ_i^*

In the above calculations, we constructed a recursion for the coefficients in the computation of the conditional expectation of the multiplicative functional M. A single iteration of this recursion can be easily adapted to compute the coefficients $\mathbf{\Phi}_i^*$, $i=0, \ldots, 3,$ in the conditional expectation in Eq. (24) for an arbitrary function $\log f(x)$.

1. Associate $\log f(x_{t+1}) = \bar{Y}(x_{t+1})$ from Eq. (A.6), ie, set $\overline{\Gamma}_i$, $i=0, \ldots, 3$, in Eq. (A.6) equal to the desired $\mathbf{\Phi}_i$ from Eq. (23). These are the coefficients in set $\bar{\mathcal{P}}$.
2. Apply the mapping $\tilde{\mathcal{E}}\,(\bar{\mathcal{P}})$, ie, compute $\tilde{\Gamma}_i$, $i=0, \ldots, 3$, and $\tilde{\mathbf{\Psi}}_i$, $i=0, 1, 2$, using (A.8). This yields the function $\log \tilde{f}\,(x_t, w_{t+1}) \equiv \log f(x_{t+1})$, with coefficient set $\tilde{\mathcal{P}}$.
3. Add to these coefficients $\tilde{\Gamma}_i$ and $\tilde{\mathbf{\Psi}}_i$ the corresponding coefficients Γ_i and $\mathbf{\Psi}_i$ of $Y_{t+1} - Y_t$ from Eq. (A.2), ie, form coefficient set $\mathcal{P} + \tilde{\mathcal{E}}\,(\bar{\mathcal{P}})$.
4. Apply the mapping $\bar{\mathcal{E}}\left(\mathcal{P} + \tilde{\mathcal{E}}(\bar{\mathcal{P}})\right)$, ie, compute (A.7) where on the right-hand side the coefficients Γ_i and $\mathbf{\Psi}_i$ (coefficient set $\mathcal{P}$) are replaced with coefficients computed in the previous step, ie, set $\mathcal{P} + \tilde{\mathcal{E}}\,(\bar{\mathcal{P}})$.
5. The resulting coefficients $\overline{\Gamma}_i$, $i=0, \ldots, 3$, are the desired coefficients $\mathbf{\Phi}_i^*$.

Appendix B Shock Elasticity Calculations

In this appendix, we provide details on some of the calculations underlying the derived shock elasticity formulas for the convenient functional form from Section 5.1.1. In particular we show, using a complete-the-squares argument, that under the transformed measure generated by the random variable $L_{1,t}$ from (25) the shock W_1 remains normally distributed with a covariance matrix:

$$\tilde{\Sigma}_t = \left[I_k - 2\,\mathrm{sym}\left(\mathrm{mat}_{k,k} \left[\mathbf{\Psi}_2 + \mathbf{\Phi}_{2,t-1}^* \Lambda_{22} + \mathbf{\Phi}_{3,t-1}^* (\Lambda_{10} \otimes \Lambda_{10}) \right] \right) \right]^{-1},$$

where I_k is the identity matrix of dimension k.[w] We suppose that this matrix is positive definite. The conditional mean vector for W_1 under the change of measure is:

$$\tilde{E}\,[W_1 \mid X_0 = x] = \tilde{\Sigma}_t \left[\mu_{t,0} + \mu_{t,1} x_1 \right],$$

where $\tilde{E}$ is the expectation under the change of measure and the coefficients $\mu_{t,0}$ and $\mu_{t,1}$ are given in the following derivation.

Thus the shock elasticity is given by:

$$\begin{aligned} \varepsilon(x,t) &= \nu(x) \cdot E[L_{1,t} W_1 \mid X_0 = x] \\ &= \nu(x)' \tilde{\Sigma}_t \left[\mu_{t,0} + \mu_{t,1} x_1 \right]. \end{aligned}$$

The shock elasticity function in this environment depends on the first component, x_1, of the state vector. Recall from (21) that this component has linear dynamics. The

[w] This formula uses the result that $(\Lambda_{10} W_1) \otimes (\Lambda_{10} W_1) = (\Lambda_{10} \otimes \Lambda_{10})(W_1 \otimes W_1)$.

coefficient matrices for the evolution of the second component, x_2, nevertheless matter for the shock elasticities even though these elasticities do not depend on this component of the state vector.

B.1 Shock Elasticities Under the Convenient Functional Form

To calculate the shock elasticities in Section 5.1.1, utilize the formulas derived in Appendix A to deduce the one-period change of measure

$$\log L_{1,t} = \log M_1 + \log E\left(\frac{M_t}{M_1} \mid X_1\right) - \log E\left[M_1 E\left(\frac{M_t}{M_1} \mid X_1\right) \mid X_0 = x\right].$$

In particular, following the set of formulas (A.8), define

$$\begin{aligned}
\mu_{0,t} &= \left[\mathbf{\Psi}_1 + \mathbf{\Phi}^*_{1,t-1}\Lambda_{1,0} + \mathbf{\Phi}^*_{2,t-1}\Lambda_{20} + \mathbf{\Phi}^*_{3,t-1}(\mathbf{\Theta}_{10}\otimes\Lambda_{10} + \Lambda_{10}\otimes\mathbf{\Theta}_{10})\right]' \\
\mu_{1,t} &= \operatorname{mat}_{k,n}\left[\mathbf{\Psi}_1 + \mathbf{\Phi}^*_{2,t-1}\Lambda_{21} + \mathbf{\Phi}^*_{3,t-1}\left(\mathbf{\Theta}_{11}\otimes\Lambda_{10} + \left[\Lambda_{10}\otimes(\mathbf{\Theta}_{11})_{\bullet j}\right]_{j=1}^{n}\right)\right] \\
\mu_{2,t} &= \operatorname{sym}\left[\operatorname{mat}_{k,k}\left(\mathbf{\Psi}_2 + \overline{\Gamma}_2\Lambda_{22} + \overline{\Gamma}_3(\Lambda_{10}\otimes\Lambda_{10})\right)\right].
\end{aligned}$$

Then it follows that

$$\begin{aligned}
\log L_{1,t} \; &= (\mu_{0,t} + \mu_{1,t}X_{1,0})' W_1 + (W_1)'\mu_{2,t}W_1 \\
&\quad -\frac{1}{2}\log E\left[\exp\left((\mu_{0,t} + \mu_{1,t}X_{1,0})' W_1 + (W_1)'\mu_{2,t}W_1\right) \mid \mathcal{F}_0\right].
\end{aligned}$$

Expression (A.5) then implies that

$$\begin{aligned}
E[L_{1,t}W_1 \mid \mathcal{F}_0] &= \widetilde{E}[W_1 \mid \mathcal{F}_0] \\
&= \left(I_k - 2\mu_{2,t}\right)^{-1}\left(\mu_{0,t} + \mu_{1t}X_{1,0}\right).
\end{aligned}$$

The variance of W_1 under the $\widetilde{\cdot}$ measure satisfies

$$\widetilde{\Sigma}_t = \left(I_k - 2\operatorname{sym}\left[\operatorname{mat}_{k,k}\left(\mathbf{\Psi}_2 + \overline{\Gamma}_2\Lambda_{22} + \overline{\Gamma}_3(\Lambda_{10}\otimes\Lambda_{10})\right)\right]\right)^{-1}.$$

B.2 Approximation of the Shock Elasticity Function

In Section 5.3.1, we constructed the approximation of the shock elasticity function $\varepsilon(x,t)$. The first-order approximation is constructed by differentiating the elasticity function under the perturbed dynamics

$$\varepsilon_1(X_{1,0},t) = \frac{d}{d\mathsf{q}}\nu(X_0(\mathsf{q}))\cdot\left.\frac{E[M_t(\mathsf{q})W_1 \mid X_0 = x]}{E[M_t(\mathsf{q}) \mid X_0 = x]}\right|_{\mathsf{q}=0} = \nu(\bar{x})\cdot E[Y_{1,t}W_1 \mid X_0 = x].$$

The first-derivative process $Y_{1,t}$ can be expressed in terms of its increments, and we obtain a state-independent function

$$\varepsilon_1(t) = \nu(\bar{x}) \cdot E\left[\sum_{j=1}^{t-1} \kappa_x (\psi_x)^{j-1} \psi_w + \kappa_w\right]'$$

where κ_x, ψ_x, κ_w, ψ_w are derivative matrices evaluated at the steady state $(\bar{x}, 0)$.

Continuing with the second derivative, we have

$$\begin{aligned}
\varepsilon_2(X_{1,0}, X_{2,0}, t) \; &= \frac{d^2}{d\mathsf{q}^2} \nu(X_0(\mathsf{q})) \cdot \frac{E[M_t(\mathsf{q}) W_1 \mid X_0 = x]}{E[M_t(\mathsf{q}) \mid X_0 = x]}\bigg|_{\mathsf{q}=0} \\
&= \nu(\bar{x}) \cdot \left\{ E\left[(Y_{1,t})^2 W_1 + Y_{2,t} W_1 \mid \mathcal{F}_0\right] - 2E[Y_{1,t} W_1 \mid \mathcal{F}_0] E[Y_{1,t} \mid \mathcal{F}_0] \right\} \\
&\quad + 2\left[\frac{\partial \nu}{\partial x'}(\bar{x})\right] X_{1,0} \cdot E[Y_{1,t} W_1 \mid \mathcal{F}_0].
\end{aligned}$$

However, notice that

$$\begin{aligned}
E\left[(Y_{1,t})^2 W_1 \mid \mathcal{F}_0\right] &= 2\left(\sum_{j=0}^{t-1} \kappa_x (\psi_x)^j X_{1,0}\right) \left(\sum_{j=1}^{t-1} \kappa_x (\psi_x)^{j-1} \psi_w + \kappa_w\right)' \\
E[Y_{1,t} W_1 \mid \mathcal{F}_0] &= \left(\sum_{j=1}^{t-1} \kappa_x (\psi_x)^{j-1} \psi_w + \kappa_w\right)' \\
E[Y_{1,t} \mid \mathcal{F}_0] &= \sum_{j=0}^{t-1} \kappa_x (\psi_x)^j X_{1,0}
\end{aligned}$$

and thus

$$E\left[(Y_{1,t})^2 W_1 \mid \mathcal{F}_0\right] - 2E[Y_{1,t} W_1 \mid \mathcal{F}_0] E[Y_{1,t} \mid \mathcal{F}_0] = 0.$$

The second-order term in the approximation of the shock elasticity function thus simplifies to

$$\varepsilon_2(X_{1,0}, X_{2,0}, t) = \nu(\bar{x}) \cdot E[Y_{2,t} W_1 \mid \mathcal{F}_0] + 2\left[\frac{\partial \nu}{\partial x'}(\bar{x})\right] X_{1,0} \cdot E[Y_{1,t} W_1 \mid \mathcal{F}_0].$$

The expression for the first term on the right-hand side is

$$\begin{aligned}
E[Y_{2,t} W_1 \mid \mathcal{F}_0] \; &= E\left[\sum_{j=0}^{t-1} \left(Y_{2,j+1} - Y_{2,j}\right) W_1 \mid \mathcal{F}_0\right] = 2\mathrm{mat}_{k,n}(\kappa_{xw}) X_{1,0} \\
&\quad + 2\sum_{j=1}^{t-1} \left[\psi_w' \left(\psi_x'\right)^{j-1} \mathrm{mat}_{n,n}(\kappa_{xx}) (\psi_x)^j + \mathrm{mat}_{k,n}\left[\kappa_x (\psi_x)^{j-1} \psi_{xw}\right]\right] X_{1,0} \\
&\quad + 2\sum_{j=1}^{t-1} \sum_{k=1}^{j-1} \left[\psi_w' \left(\psi_x'\right)^{k-1} \mathrm{mat}_{n,n}\left[\kappa_x (\psi_x)^{j-k-1} \psi_{xx}\right] (\psi_x)^k\right] X_{1,0}.
\end{aligned}$$

To obtain this result, notice that repeated substitution for $Y_{1,j+1} - Y_{1,j}$ into the above formula yields a variety of terms but only those containing $X_{1,0} \otimes W_1$ have a nonzero conditional expectation when interacted with W_1.

ACKNOWLEDGMENTS

We would like to thank John Heaton and Vadim Linetsky for helpful comments. Spencer Lyon and Victor Zhorin provided excellent research assistance.

REFERENCES

Adrian, T., Boyarchenko, N., 2012. Intermediary leverage cycles and financial stability. Federal Reserve Bank of New York Staff Report No. 567.

Ai, H., Croce, M.M., Diercks, A., Li, K., 2013. Production-based term structure of equity returns.

Alvarez, F., Jermann, U.J., 2004. Using asset prices to measure the cost of business cycles. J. Polit. Econ. 112 (6), 1223–1256.

Anderson, E.W., Hansen, L.P., Sargent, T.J., 2003. A quartet of semigroups for model specification, robustness, prices of risk, and model detection. J. Eur. Econ. Assoc. 1 (1), 68–123.

Andreasen, M.M., Fernández-Villaverde, J., Rubio-Ramírez, J.F., 2010. The pruned state space system for non-linear DSGE models: Asset pricing applications to GMM and SMM. Unpublished manuscript.

Belo, F., Collin-Dufresne, P., Goldstein, R.S., 2015. Dividend dynamics and the term structure of dividend strips. J. Financ. 70 (3), 1115–1160.

Benigno, G., Benigno, P., Nisticò, S., 2010. Second-order approximation of dynamic models with time-varying risk. NBER Working Paper W16633.

Bernanke, B.S., Gertler, M., 1989. Agency costs, net worth, and business fluctuations. Am. Econ. Rev. 79 (1), 14–31.

Bernanke, B.S., Gertler, M., Gilchrist, S., 1999. The financial accelerator in a quantitative business cycle framework. In: Handbook of Macroeconomics, vol. 1, Chapter 21. Elsevier B.V., Amsterdam, Netherlands, pp. 1341–1393.

Beveridge, S., Nelson, C.R., 1981. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'. J. Monet. Econ. 7, 151–174.

Bhandari, A., Borovička, J., Ho, P., 2016. Identifying ambiguity shocks in business cycle models using survey data.

Bianchi, F., 2012. Regime switches, agents' beliefs, and post-World War II U.S. macroeconomic dynamics. Rev. Econ. Stud. 67 (2), 380–405.

Bianchi, F., 2015. Rare events, financial crises, and the cross-section of asset returns.

Bianchi, J., 2016. Efficient bailouts? Am. Econ. Rev. Forthcoming.

Bianchi, F., Ilut, C., 2015. Monetary/fiscal policy mix and agents' beliefs.

Bianchi, F., Melosi, L., 2016. Modeling the evolution of expectations and uncertainty in general equilibrium. Int. Econ. Rev. 57 (2), 717–756.

Bianchi, F., Ilut, C., Schneider, M., 2013. Uncertainty shocks, asset supply and pricing over the business cycle.

Blanchard, O.J., Kahn, C.M., 1980. The solution of linear difference models under rational expectations. Econometrica 48 (5), 1305–1312.

Blanchard, O.J., Quah, D., 1989. The dynamic effects of aggregate demand and supply disturbances. Am. Econ. Rev. 79 (4), 655–673.

Bocola, L., 2016. The pass-through of sovereign risk. J. Polit. Econ. Forthcoming.

Borodin, A.N., Salminen, P., 2002. Handbook of Brownian Motion: Facts and Formulae, second edition. Birkhäuser, Basel, Boston, Berlin.

Borovička, J., Hansen, L.P., 2013. Robust preference expansions.

Borovička, J., Hansen, L.P., 2014. Examining macroeconomic models through the lens of asset pricing. J. Econom. 183 (1), 67–90.

Borovička, J., Hansen, L.P., Hendricks, M., Scheinkman, J.A., 2011. Risk-price dynamics. J. Financ. Econom. 9 (1), 3–65.

Borovička, J., Hansen, L.P., Scheinkman, J.A., 2014. Shock elasticities and impulse responses. Math. Finan. Econ. 8 (4), 333–354.

Borovička, J., Hansen, L.P., Scheinkman, J.A., 2015. Misspecified recovery. J. Financ. Forthcoming.

Brunnermeier, M.K., Sannikov, Y., 2014. A macroeconomic model with a financial sector. Am. Econ. Rev. 104 (2), 379–421.

Campbell, J.Y., Shiller, R.J., 1988. Stock prices, earnings, and expected dividends. J. Financ. 43 (3), 661–676. http://dx.doi.org/10.1111/j.1540-6261.1988.tb04598.x.

Chen, H., 2010. Macroeconomic conditions and the puzzles of credit spreads and capital structure. J. Financ. 65 (6), 2171–2212.

Chung, H., Davig, T., Leeper, E.M., 2007. Monetary and fiscal policy switching. J. Money Credit Bank. 39 (4), 809–842.

David, A., 2008. Heterogeneous beliefs, speculation, and the equity premium. J. Financ. 63 (1), 41–83.

Davig, T., Leeper, E.M., Walker, T.B., 2010. "Unfunded liabilities" and uncertain fiscal financing. J. Monet. Econ. 57 (5), 600–619.

Davig, T., Leeper, E.M., Walker, T.B., 2011. Inflation and the fiscal limit. Eur. Econ. Rev. 55 (1), 31–47.

Di Nunno, G., Øksendal, B., Proske, F., 2009. Malliavin Calculus for Lévy Processes with Applications to Finance. Springer Verlag, Berlin, Heidelberg.

Di Tella, S., 2015. Uncertainty shocks and balance sheet recessions. J. Polit. Econ. Forthcoming.

Epstein, L.G., Zin, S.E., 1989. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: a theoretical framework. Econometrica 57 (4), 937–969.

Farmer, R.E., Waggoner, D.F., Zha, T., 2011. Minimal state variable solutions to Markov-switching rational expectations models. J. Econ. Dyn. Control. 35 (12), 2150–2166.

Feller, W., 1952. The parabolic differential equations and the associated semi-groups of transformations. Ann. Math. 55 (3), 468–519.

Feller, W., 1957. On boundaries and lateral conditions for the Kolmogorov differential equations. Ann. Math. 65 (3), 527–570.

Foerster, A., Rubio-Ramirez, J., Waggoner, D.F., Zha, T., 2014. Perturbation methods for Markov-switching DSGE models. NBER Working Paper W20390.

Frisch, R., 1933. Propagation problems and impulse problems in dynamic economics. In: Economic Essays in Honour of Gustav Cassel. Allen and Unwin, Oslo, pp. 171–205.

Hansen, L.P., 2012. Dynamic valuation decomposition within stochastic economies. Econometrica 80 (3), 911–967. Fisher-Schultz Lecture at the European Meetings of the Econometric Society.

Hansen, L.P., 2014. Nobel lecture: uncertainty outside and inside economic models. J. Polit. Econ. 122 (5), 945–987. https://ideas.repec.org/a/ucp/jpolec/doi10.1086-678456.html.

Hansen, L.P., Richard, S.F., 1987. The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. Econometrica 50, 587–614.

Hansen, L.P., Scheinkman, J.A., 2009. Long term risk: an operator approach. Econometrica 77 (1), 177–234.

Hansen, L.P., Scheinkman, J.A., 2012. Pricing growth-rate risk. Finance Stochast. 16, 1–15.

Hansen, L.P., Sargent, T.J., Tallarini Jr., T.D., 1999. Robust permanent income and pricing. Rev. Econ. Stud. 66 (4), 873–907.

Hansen, L.P., Heaton, J.C., Li, N., 2008. Consumption strikes back? Measuring long-run risk. J. Polit. Econ. 116, 260–302.

Hasler, M., Marfè, R., 2015. Disaster recovery and the term structure of dividend strips.

He, Z., Krishnamurthy, A., 2013. Intermediary asset pricing. Am. Econ. Rev. 103 (2), 732–770.

Holmes, M.H., 1995. Introduction to Perturbation Methods. Springer Verlag, New York.

Jacobson, D.H., 1973. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. IEEE Trans. Autom. Control AC-18, 1124–1131.

Justiniano, A., Primiceri, G.E., 2008. The time-varying volatility of macroeconomic fluctuations. Am. Econ. Rev. 98 (3), 604–641.

Karlin, S., Taylor, H.M., 1981. A Second Course in Stochastic Processes. Academic Press, London, United Kingdom.

Kim, J., Kim, S., Schaumburg, E., Sims, C.A., 2008. Calculating and using second-order accurate solutions of discrete time dynamic equilibrium models. J. Econ. Dyn. Control 32 (11), 3397–3414.

Klimenko, N., Pfeil, S., Rochet, J.C., De Nicolò, G., 2016. Aggregate bank capital and credit dynamics.

Kreps, D.M., Porteus, E.L., 1978. Temporal resolution of uncertainty and dynamic choice theory. Econometrica 46 (1), 185–200.

Linetsky, V., 2008. Spectral methods in derivatives pricing. In: Handbooks in Operations Research and Management Science, vol. 15, Chapter 6. Elsevier B.V., Amsterdam, Netherlands, pp. 213–289.

Liu, Z., Waggoner, D.F., Zha, T., 2009. Asymmetric expectation effects of regime switches in monetary policy. Rev. Econ. Dyn. 12 (2), 284–303.

Liu, Z., Waggoner, D.F., Zha, T., 2011. Sources of macroeconomic fluctuations: a regime-switching DSGE approach. Quant. Econ. 2 (2), 251–301.

Lombardo, G., Uhlig, H., 2014. A theory of pruning. European Central Bank. https://ideas.repec.org/p/ecb/ecbwps/20141696.html. Working Paper Series 1696.

Lopez, P., Lopez-Salido, D., Vazquez-Grande, F., 2015. Nominal rigidities and the term structures of equity and bond returns.

Maenhout, P.J., 2004. Robust portfolio rules and asset pricing. Rev. Financ. Stud. 17 (4), 951–983. http://dx.doi.org/10.1093/rfs/hhh003. http://rfs.oxfordjournals.org/content/17/4/951.full.pdf+html, http://rfs.oxfordjournals.org/content/17/4/951.abstract.

Malkhozov, A., Shamloo, M., 2011. Asset prices in affine real business cycle models.

Moreira, A., Savov, A., 2016. The macroeconomics of shadow banking. J. Financ. Forthcoming.

Nakamura, E., Sergeyev, D., Steinsson, J., 2016. Growth-rate and uncertainty shocks in consumption: cross-country evidence. Columbia University. http://www.nber.org/papers/w18128.

Nualart, D., 2006. The Malliavin Calculus and Related Topics, second edition. Springer Verlag, Berlin, Heidelberg, New York.

Park, H., 2015. Ross recovery with recurrent and transient processes.

Pryce, J.D., 1993. Numerical Solution of Sturm-Liouville Problems. Oxford University Press, Oxford, United Kingdom.

Qin, L., Linetsky, V., 2014. Long term risk: a martingale approach. Mimeo, Northwestern University.

Qin, L., Linetsky, V., 2014. Positive eigenfunctions of Markovian pricing operators: Hansen-Scheinkman factorization and Ross recovery. Mimeo, Northwestern University.

Qin, L., Linetsky, V., Nie, Y., 2016. Long forward probabilities, recovery and the term structure of bond risk premiums. Mimeo, Northwestern University.

Schmitt-Grohé, S., Uribe, M., 2004. Solving dynamic general equilibrium models using a second-order approximation to the policy function. J. Econ. Dyn. Control 28 (4), 755–775.

Sims, C., 1980. Macroeconomics and reality. Econometrica 48 (1), 1–48.

Sims, C.A., 2002. Solving rational expectations models. Comput. Econ. 20 (1–2), 1–20.

Sims, C.A., Zha, T., 2006. Were there regime switches in U.S. monetary policy. Am. Econ. Rev. 96 (1), 54–81.

Slutsky, E., 1927. The summation of random causes as the source of cyclic processes. Probl. Econ. Cond. 3 (1).

Tallarini, T.D., 2000. Risk-sensitive real business cycles. J. Monet. Econ. 45 (3), 507–532. http://EconPapers.repec.org/RePEc:eee:moneco:v:45:y:2000:i:3:p:507-532.

van Binsbergen, J.H., Koijen, R.S.J., 2016. The term structure of returns: facts and theory. J. Financ. Econ. Forthcoming.

van Binsbergen, J.H., Brandt, M.W., Koijen, R.S.J., 2012. On the timing and pricing of dividends. Am. Econ. Rev. 102 (4), 1596–1618.

van Binsbergen, J., Hueskes, W., Koijen, R.S., Vrugt, E.B., 2013. Equity yields. J. Financ. Econ. 110 (3), 503–519.

Walden, J., 2014. Recovery with unbounded diffusion processes.

Whittle, P., 1990. Risk Sensitive and Optimal Control. John Wiley and Sons, West Suffix, England.

Yule, G.U., 1927. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. Philos. Trans. R. Soc. 226, 267–298.

Zviadadze, I., 2016. Term structure of consumption risk premia in the cross section of currency returns. J. Financ. Forthcoming.

CHAPTER 21

Quantitative Models of Sovereign Debt Crises

M. Aguiar[*], S. Chatterjee[†], H. Cole[‡], Z. Stangebye[§]
[*]Princeton University, Princeton, NJ, United States
[†]Federal Reserve Bank of Philadelphia, Philadelphia, PA, United States
[‡]University of Pennsylvania, Philadelphia, PA, United States
[§]University of Notre Dame, Notre Dame, IN, United States

Contents

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.04.005

Abstract

This chapter is on quantitative models of sovereign debt crises in emerging economies. We interpret debt crises broadly to cover all of the major problems a country can experience while trying to issue new debt, including default, sharp increases in the spread and failed auctions. We examine the spreads on sovereign debt of 20 emerging market economies since 1993 and document the extent to which fluctuations in spreads are driven by country-specific fundamentals, common latent factors and observed global factors. Our findings motivate quantitative models of debt and default with the following features: (i) trend stationary or stochastic growth, (ii) risk averse competitive lenders, (iii) a strategic repayment/borrowing decision, (iv) multiperiod debt, (v) a default penalty that includes both a reputation loss and a physical output loss, and (vi) rollover defaults. For the quantitative evaluation of the model, we focus on Mexico and carefully discuss the successes and weaknesses of various versions of the model. We close with some thoughts on useful directions for future research.

Keywords

Quantitative models, Emerging markets, Stochastic trend, Capital flows, Rollover crises, Debt sustainability, Risk premia, Default risk

JEL Classification Codes:

D52, F34, E13, G15, H63

1. INTRODUCTION

This chapter is about sovereign debt crises, instances in which a government has trouble selling new debt. An important example is when a government is counting on being able to roll over its existing debt in order to service it over time. When we refer to trouble selling its debt, we include being able to sell new debt but only with a large jump in the spread on that debt over comparable risk-free debt, failed auctions, suspension of payments, creditor haircuts and outright default. So our notion of a debt crisis covers all of the major negative events that one associates with sovereign debt issuance.

We focus on debt crises in developing countries because the literature has focused on them and because these countries provide the bulk of our examples of debt crises and defaults. However, the recent debt crises in the European Union remind us that this is certainly not always the case. While the recent crises in the EU are of obvious interest, they come with a much more complicated strategic dimension, given the role played by the European Central Bank and Germany in determining the outcomes for a country like, say, Greece. For this reason we will hold to a somewhat more narrow focus. Despite this, we see our analysis as providing substantial insight into sovereign debt crises in developed countries as well.

This chapter will highlight quantitative models of the sovereign debt market. We will focus on determining where the current literature stands and where we need to go next. Hence, it will not feature an extensive literature survey, though we will of course survey the literature to some extent, including a brief overview at the end of the chapter. Instead, we will lay out a fairly cutting-edge model of sovereign debt issuance and use that model and its various permutations to gauge the successes and failures of the current literature as we see them.

The chapter will begin by considering the empirical evidence on spreads. We will examine the magnitude and volatility of spreads on sovereign debt among developing countries. We will seek to gauge the extent to which this debt features a risk premium in addition to default risk. We will also seek to characterize the extent to which the observed spread is driven by country-specific fundamentals, global financial risk and uncertainty factors, or other common drivers. To do this, we will estimate a statistical model of the spread process in our data, and this statistical model will feature several common factors that we estimate along with the statistical model. The facts that emerge from this analysis will then form the basis on which we will judge the various models that we consider in the quantitative analysis.

The chapter will then develop a quantitative model of sovereign debt that has the following key features: risk-averse competitive lenders, since it will turn out that risk premia are substantial, and a strategic sovereign who chooses how much to borrow and whether or not to repay, much as in the original Eaton and Gersovitz (1981) model. The sovereign will issue debt that has multiperiod maturity. While we will take the maturity of the debt to be parametric, being able to examine the implications of short and long maturity is an important aspect of the analysis. Default by the sovereign will feature two punishments: a period of exclusion from credit markets and a loss in output during the period of exclusion. Pure reputation effects are known to fail (Bulow and Rogoff, 1989) and even coupling them with a loss of saving as well as borrowing does not generate a sufficient incentive to repay the sorts of large debts that we see in the data. Hence, we include the direct output cost as well.

Our model will feature both fundamental defaults, in which default is taking place under the best possible terms (fixing future behavior). The model will also allow for rollover or liquidity defaults, in which default occurs when lending takes place under the worst possible terms (again, fixing future behavior) as in Cole and Kehoe (2000). We include both types of defaults since they seem to be an important component of the data. Doing so, especially with multiperiod debt maturity, will require some careful modeling of the timing of actions within the period and a careful consideration of both debt issuance and debt buybacks. In addition, the possibility of future rollover crises will affect the pricing of debt today and the incentives to default, much as in the original Calvo (1988) model.

We will consider two different growth processes for our borrowing countries. The first will feature stochastic fluctuations around a deterministic trend with constant growth. The second will feature stochastic growth shocks. We include the deterministic trend process because the literature has focused on it. However, the notion that we have roughly the same uncertainty about where the level of output of a developing country will be in 5 years and in 50 years seems sharply counterfactual, as documented by Aguiar and Gopinath (2007). Hence our preferred specification is the stochastic growth case and, so, we discuss this case as well.

There will be three shocks in the model. The first is a standard output shock that will vary depending on which growth process we assume. The second is a shock to lender wealth. The third is a belief-coordination shock that will determine whether a country gets the best or the worst possible equilibrium price schedule in a period. An important question for us will be the extent to which these shocks can generate movements in the spread that are consistent with the patterns we document in our empirical analysis of the data.

The chapter will examine two different forms of the output default cost. The first is a proportional default cost as has been assumed in the early quantitative analyses and in the theoretical literature on sovereign default. The second form is a nonlinear output cost such as was initially pioneered by Arellano (2008). In this second specification, the share of output lost in default depends positively on (predefault) output. Thus, default becomes a more effective mechanism for risk sharing compared to the proportional cost case. As noted in Chatterjee and Eyigungor (2012), adding this feature also helps to increase the volatility of sovereign spreads.

2. MOTIVATING FACTS

2.1 Data for Emerging Markets

We start with a set of facts that will guide us in developing our model of sovereign debt crises. Our sample spans the period 1993Q4–2014Q4 and includes data from 20 emerging markets: Argentina, Brazil, Bulgaria, Chile, Colombia, Hungary, India, Indonesia, Latvia, Lithuania, Malaysia, Mexico, Peru, Philippines, Poland, Romania, Russia, South Africa, Turkey, and Ukraine. For each of these economies, we have data on GDP in US dollars measured in 2005 domestic prices and exchange rates (real GDP), GDP in US dollars measured in current prices and exchange rates (nominal GDP), gross external debt in US dollars (debt), and market spreads on sovereign debt.[a]

Tables 1 and 2 report summary statistics for the sample.[b] Table 1 documents the high and volatile spreads that characterized emerging market sovereign bonds during this period. The standard deviation of the level and quarterly change in spreads 676 and

[a] Data source for GDP and debt is Haver Analytics' Emerge database. The source of the spread data is JP Morgan's Emerging Market Bond Index (EMBI).

[b] Note that Russia defaulted in 1998 and Argentina in 2001, and while secondary market spreads continued to be recorded post default, these do not shed light on the cost of new borrowing as the governments were shut out of international bond markets until they reached a settlement with creditors. Similarly, the face value of debt is carried throughout the default period for these economies.

Table 1 Sovereign spreads: Summary statistics

Country	Mean $r - r^*$	Std dev $r - r^*$	Std dev $\Delta(r - r^*)$	95th pct $\Delta(r - r^*)$	Frequency crisis
Argentina	1525	1759	610	717	0.18
Brazil	560	393	174	204	0.09
Bulgaria	524	486	129	155	0.03
Chile	146	57	34	34	0.00
Colombia	348	206	88	245	0.05
Hungary	182	154	57	88	0.02
India	225	54	47	85	0.00
Indonesia	285	137	98	73	0.02
Latvia	157	34	16	17	0.00
Lithuania	246	92	48	98	0.00
Malaysia	175	122	75	81	0.03
Mexico	345	253	134	127	0.05
Peru	343	196	84	182	0.06
Philippines	343	153	75	136	0.04
Poland	191	138	54	67	0.01
Romania	271	102	49	68	0.00
Russia	710	1096	478	175	0.06
South Africa	226	116	68	99	0.03
Turkey	395	217	95	205	0.05
Ukraine	760	607	350	577	0.11
Pooled	431	676	229	158	

229 basis points, respectively. Table 2 reports an average external debt-to-(annualized) GDP ratio of 0.46. This level is low relative to the public debt levels observed in developed economies. The fact that emerging markets generate high spreads at relatively low levels of debt-to-GDP reflects one aspect of the "debt intolerance" of these economies documented by Reinhart et al. (2003).

The final column concerns "crises," which we define as a change in spreads that lie in the top 5% of the distribution of quarterly changes. This threshold is a 158 basis-point jump in the spread. By construction, 5% of the changes are coded as crises; however, the frequency of crises is not uniform across countries. Nearly 20% of Argentina's quarter-to-quarter changes in spreads lie above the threshold, while many countries have no such changes.

While many of the countries in our sample have very high spreads, only two—Russia in 1998 and Argentina in 2001—ended up defaulting on their external debt, while a third, Ukraine, defaulted on its internal debt (in 1998). This highlights the fact that periods of high spreads are more frequent events than defaults. Nevertheless, it is noteworthy that the countries with the highest mean spreads are the ones that ended up defaulting during this period. This suggests that default risk and the spread are connected.

Table 2 Sovereign spreads: Summary statistics

Country	Mean $\frac{B}{4{*}Y}$	Corr $(\Delta(r - r^{*}),\Delta y)$	Corr $(r - r^{*},\%\Delta B)$	Corr $(\Delta(r - r^{*}),\%\Delta B)$
Argentina	0.38	−0.35	−0.22	0.08
Brazil	0.25	−0.11	−0.18	−0.01
Bulgaria	0.77	0.09	−0.20	0.06
Chile	0.41	−0.16	−0.18	−0.11
Columbia	0.27	−0.29	−0.40	−0.07
Hungary	0.77	−0.24	−0.56	−0.05
India	0.82	−0.32	0.04	−0.65
Indonesia	0.18	−0.43	−0.03	0.07
Latvia	0.49	−0.18	−0.12	−0.16
Lithuania	1.06	−0.25	−0.17	−0.31
Malaysia	0.54	−0.56	−0.33	0.24
Mexico	0.16	−0.4	0.23	−0.13
Peru	0.48	−0.01	−0.39	−0.05
Philippines	0.47	−0.16	0.06	0.09
Poland	0.57	−0.09	−0.35	−0.38
Romania	0.61	0.5	0.42	−0.33
Russia	NA	−0.45	−0.30	0.02
South	0.26	−0.14	−0.38	−0.24
Turkey	0.38	−0.34	−0.20	0.08
Ukraine	0.64	−0.49	−0.60	−0.07
Pooled	0.46	−0.27	−0.19	0.01

2.2 Statistical Spread Model

To further evaluate the empirical behavior of emerging market government bond spreads, we fit a statistical model to our data. In this model a country's spread is allowed to depend on country-specific fundamentals as well as several mutually orthogonal common factors (common across emerging markets) that we will implicitly determine as part of the estimation. To do this, we use EMBI data at a quarterly frequency. We have data for $I=20$ countries from 1993:Q4–2015:Q2 (so $T=87$), with sporadic missing values. If we index a country by i and a quarter by t, then we observe spreads, debt-to-GDP ratios, and real GDP growth: $\{s_{it}, b_{it}, g_{it}\}_{i=1,t=1}^{I,T}$. We also suppose that there are a set of J common factors that impact all the countries (though perhaps not symmetrically): $\{\alpha_t^j\}_{j=1}^{J}$.

We specify our statistical model as follows:

$$s_{it} = \beta_i b_{it} + \gamma_i g_{it} + \sum_{j=1}^{J} \delta_i^j \alpha_t^j + \kappa_i + \epsilon_{it}, \tag{1}$$

where ϵ_{it} is a mean-zero, normally distributed shock with variance σ_i^2. Notice that we allow for the average spread and innovation volatility to vary across countries. In the estimation we impose the constraint that $\delta_i^j \geq 0$ for all i, so we are seeking common factors that cause all spreads to rise and fall together.

These common factors are permitted to evolve as follows. Let α_t be the J-dimensional vector of common factors at time t. Then

$$\alpha_t = \Gamma \alpha_{t-1} + \eta_t \tag{2}$$

where η_t is a J-dimensional vector of normally distributed i.i.d. innovations orthogonal to each other. Because we estimate separate impact coefficients for each common factor, we normalized the innovation volatilities to 0.01. We restrict Γ to be a diagonal matrix, ie, our common factors are assumed to be orthogonal and to follow AR(1) processes.

To estimate this model, we transform it into state-space form and apply MLE. We apply the (unsmoothed) Kalman Filter to compute the likelihood for a given parameterization. When the model encounters missing values, we will exclude those values from the computation of the likelihood and the updating of the Kalman Filter. Thus, missing values will count neither for nor against a given parameterization.

Table 3 reports the explanatory power of the country-specific fundamentals as well as the two global factors. Specifically, we construct a variance decomposition following the algorithm of Lindeman et al. (1980) as outlined by Gromping (2007). This procedure constructs the average marginal R^2 in the case of correlated regressors by assuming a uniform distribution over all possible permutations of the regression coefficients. We can see

Table 3 Country-specific variance decomposition average marginal R^2

Country (i)	b_{it}	g_{it}	α_t^1	α_t^2	R^2	Obs.
Argentina	0.16	0.01	0.20	0.02	0.39	39
Brazil	0.28	0.01	0.52	0.05	0.87	81
Bulgaria	0.18	0.01	0.44	0.27	0.90	59
Chile	0.05	0.13	0.38	0.21	0.77	63
Colombia	0.20	0.05	0.55	0.16	0.95	55
Hungary	0.28	0.19	0.05	0.12	0.64	63
India	0.10	0.26	0.32	0.32	1.00	8
Indonesia	0.09	0.07	0.38	0.45	0.99	43
Latvia	0.03	0.03	0.86	0.08	1.00	9
Lithuania	0.06	0.01	0.67	0.25	0.99	20
Malaysia	0.23	0.11	0.46	0.16	0.96	24
Mexico	0.01	0.23	0.59	0.17	0.99	51
Peru	0.34	0.04	0.52	0.07	0.97	71
Philippines	0.26	0.05	0.50	0.01	0.83	84
Poland	0.06	0.10	0.23	0.32	0.71	42
Romania	0.15	0.03	0.47	0.23	0.87	12
Russia	0.12	0.05	0.21	0.51	0.90	62
South Africa	0.03	0.32	0.25	0.36	0.96	48
Turkey	0.05	0.09	0.77	0.04	0.94	74
Ukraine	0.02	0.26	0.20	0.41	0.89	44

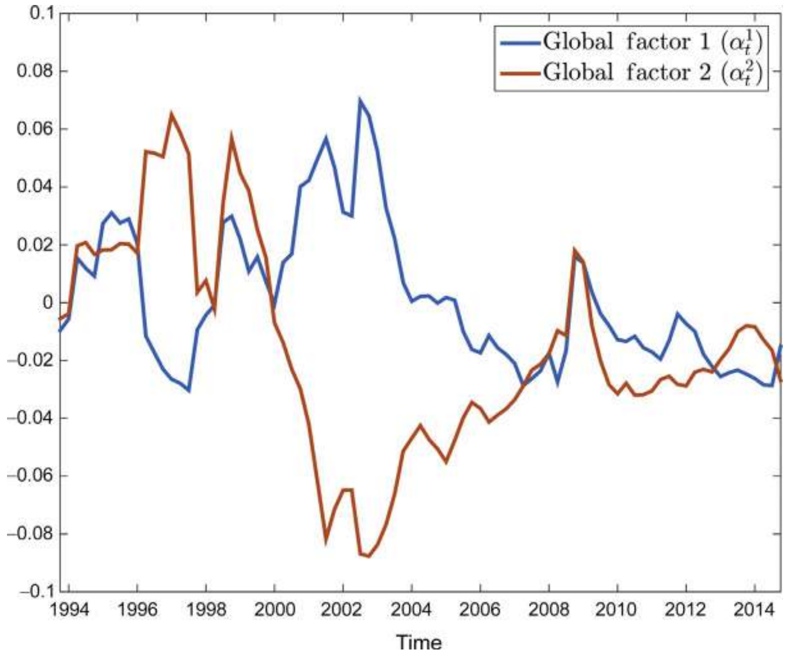


Fig. 1 Estimated common factors.

from this table first that our regressors explain much of the variation for many of the countries (as high as 99.88% for India). We can also see that country-specific fundamentals, here in the form of the debt-to-GDP ratio and the growth rate of output, explain only a modest amount of the variation in the spreads; typically less than 20%. This means that much of the movement in the spreads is explained by our two orthogonal factors.

Fig. 1 plots our two common factors.[c] Given the importance our estimation ascribes to them, we sought to uncover what is really driving their movements. To do this, we use a regression to try to construct our estimated common factors from the CBOE VIX, S&P 500 Diluted Earnings P/E ratio, and the LIBOR.[d] These regressors are standard measures of foreign financial-market uncertainty, price of risk and borrowing costs, respectively. These results are reported in table 4. The top panel reports the results from regressing the level of the factors on the level of foreign financial variables and the bottom reports the comparable regressions in first differences. We find that the foreign financial variables explain a modest amount of the variation in the level of the common factors: Each has an R^2 less than 0.3. To the extent that these objects do explain the common factors, however, it seems as if common factor 1 is driven primarily by measures of investor

[c] See Longstaff et al. (2011) for a related construction of a global risk factor.

[d] The LIBOR is almost perfectly correlated with the fed funds rate, so for precision of estimates we exclude the latter.

Table 4 Common factor regressions: Levels

Index		VIX	P/E ratio	LIBOR	R^2
Levels					
α_t^1	Coefficient	$8.32e-4$ (3.36e−4)	$2.00e-3$ (6.31e−4)	$9.75e-4$ (1.1e−3)	
	Var decomp	0.10	0.17	0.02	0.29
α_t^2	Coefficient	$6.1383e-4$ (5.0460e−4)	−0.0017 (9.4742e−4)	0.0088 (0.0017)	
	Var decomp	$-4.0795e-5$	−0.0058	0.2722	0.27
First differences					
α_t^1	Coefficient	0.001 (0.002)	−0.001 (0.001)	−0.001 (0.002)	
	Var decomp	0.30	0.06	0.00	0.35
α_t^2	Coefficient	0.001 (<0.001)	0.001 (0.001)	0.002 (0.003)	
	Var decomp	0.05	<0.01	0.01	0.06

uncertainty and the price of risk, while common factor 2 is driven primarily by world interest rates. In first differences, the foreign factors explain a third of the variation in the first factor but very little of the second factor.

There is an additional surprising finding about how risk pricing impacts our spreads. The coefficient on the P/E ratio for the level specification is *positive* in common factor 1, where it has a substantial impact. Since an increase in the price of risk will drive down the P/E ratio, this means that our spreads are rising when the market price of risk is falling. This is the opposite of what our intuition might suggest. This coefficient reverses sign in the first-difference specification, reflecting that the medium run and longer correlation between the P/E ratio and our first factor has the opposite sign of the quarter-to-quarter correlation. The first-difference specification is what has been studied in the literature (Longstaff et al., 2011; Borri and Verdelhan, 2011). These results show that the foreign risk premium may influence spreads differentially on impact vs in the longer run.

2.3 Excess Returns

We turn next to the relationship between spreads and defaults. One of the striking facts here is that spreads "over-predict" future defaults in that ex post returns exceed the return on risk-free assets. Hence, risk premia play an important role.

The fact that spreads are compensating lenders for more than the risk-neutral probability of default is suggested by the statistics reported in Table 1. The average spread is relatively high, and there are significant periods in which spreads are several hundred basis points. However, the sample contains only two defaults: Russia in 1998 and Argentina in 2001.

To explore this more systematically, we compute the realized returns on the EMBI+ index, which represents a value-weighted portfolio of emerging country debt constructed by JP Morgan. In Table 5, we report the return on this portfolio for the full

Table 5 Realized bond returns

Period	EMBI+	2-Year treasury	5-Year treasury
1993Q1–2014Q4	9.7	3.7	4.7
1993Q1–2003Q4	11.1	5.4	6.3
2004Q1–2014Q4	8.2	2.0	3.1

sample period the index is available, as well as two subperiods. The table also reports the returns to the portfolio of US Treasury securities of 2 years and 5 years maturity. We offer two risk-free references, as the EMBI+ does not have a fixed maturity structure and probably ranges between 2 and 5 years.

The EMBI+ index paid a return in excess of the risk-free portfolio of 5 to 6%. This excess return is roughly stable across the two subperiods as well. Whether the realized return reflects the ex ante expected return depends on whether our sample accurately reflects the population distribution of default and repayment. The assumption is that by pooling a portfolio of bonds, the EMBI+ followed over a 20 year period provides a fair indication of the expected return on a typical emerging market bond. Of course, we cannot rule out the possibility that this sample is not representative. Nevertheless, the observed returns are consistent with a fairly substantial risk premium charged to sovereign borrowers.

2.4 Deleveraging

The data from emerging markets can also shed light on debt dynamics during a crisis. Table 2 documents that periods of above-average spreads are associated with reductions in the face value of gross external debt. The pooled correlation of spreads at time t and the percentage change in debt between $t-1$ and t is -0.19. The correlation of the *change* in spread and debt is roughly zero. However, a large change in the spread (that is, a crisis period) is associated with a subsequent decline in debt. In particular, regressing the percent change in debt between t and $t+1$ on the indicator for a crisis in period t generates a coefficient of -1.6 and a t-stat of nearly 3. This relationship is robust to the inclusion of country fixed effects. This implies that a sharp spike in spreads is associated with a subsequent decline in the face value of debt.

2.5 Taking Stock

Our empirical analysis has led us to a set of criteria that we would like our model to satisfy. Specifically:

1. Crises, and particularly defaults, are low probability events;
2. Crises are not tightly connected to poor fundamentals;

3. Spreads are highly volatile;
4. Rising spreads are associated with deleveraging by the sovereign; and
5. Risk premia are an important component of sovereign spreads.

In considering which features of real-world economies are important in generating these patterns, the first thing to recognize is that sovereign debt lacks a direct enforcement mechanism: most countries default despite having the physical capacity to repay. Yet, countries seem perfectly willing to service significant amounts of debt most of the time (rescheduling of debts and outright default are relatively rare events). Without any deadweight costs of default, the level of debt that a sovereign would be willing to repay is constrained by the worst punishment lenders can inflict on the sovereign, namely, permanent exclusion from all forms of future credit. It is well known that this punishment is generally too weak, quantitatively speaking, to sustain much debt (this is spelled out in a numerical example in Aguiar and Gopinath, 2006). Thus, we need to posit substantial deadweight costs of default.

Second, defaults actually occurring in equilibrium reflect the fact that debt contracts are not fully state-contingent, and default provides an implicit form of insurance. However, with rational risk-neutral lenders who break even, on average, for every loan they make to sovereigns, the deadweight cost of default (which does not accrue to lenders) makes default an actuarially unfair form of insurance against bad states of the world for the sovereign. And, with risk-averse lenders, this insurance-through-default becomes even more actuarially unfair. Given fairly substantial deadweight costs of default and substantial risk aversion on the part of lenders, the insurance offered by the possibility of default appears to be quite costly in practice. The fact that countries carry large external debt positions despite the costs suggests that sovereigns are fairly impatient.

However, while myopia can explain in part why sovereigns borrow, it does not necessarily explain why they default. As noted already, default is a very costly form of insurance against bad states of the world. This fact—via equilibrium prices—can be expected to encourage the sovereign to stay away from debt levels for which the probability of default is significant. This has two implications. First, when crises/defaults do materialize, they come as a surprise, which is consistent with these events being low probability. Unfortunately, the other side of this coin is that getting the mean and volatility of spreads right is a challenge for quantitative models. Getting high and variable spreads means getting periods of high default risk as well as substantial variation in expected future default risk. This will be difficult to achieve when the borrower has a strong incentive to adjust his debt-to-output level to the point where the probability of future default is (uniformly) low.

3. ENVIRONMENT

Our environment is a simplified version of the one introduced in ACCS (2016). The analysis focuses on a sovereign government that makes consumption and savings/

borrowing decisions on behalf of the denizens of a small open economy facing a fluctuating endowment stream. The economy is small relative to the rest of the world in the sense that the sovereign's decisions do not affect any world prices, including the world risk-free interest rate. However, the sovereign faces a segmented credit market in that it can only borrow from a set of risk-averse potential lenders with limited wealth. In this section, we proceed by describing the economy of which the sovereign is in charge, the sovereign's decision problem and the lenders' decision problem. We then give the definition of an equilibrium and discuss issues related to equilibrium selection. We conclude the section by briefly describing how we compute the model.

3.1 The Economy

3.1.1 Endowments

Time is discrete and indexed by $t = 0, 1, 2, \ldots$. The economy receives a stochastic endowment $Y_t > 0$ each period. We assume that

$$\ln Y_t = \sum_{s=1}^{t} g_s + z_t, \tag{3}$$

where g_t and z_t follow first-order Markov processes. This specification follows Aguiar and Gopinath (2006, 2007) and nests the endowment processes that have figured in quantitative studies. In particular, setting $g_t = g$ generates a deterministic linear trend. More generally, g_t can be random, which corresponds to the case of a stochastic trend. In either case, z_t is transitory (but potentially persistent) fluctuations around trend growth. In this chapter we will study both specifications in some detail.

3.1.2 Preferences

The economy is run by an infinitely-lived sovereign government. The utility obtained by the sovereign from a sequence of aggregate consumption $\{C_t\}_{t=0}^{\infty}$ is given by:

$$\sum_{t=0}^{\infty} \beta^t u(C_t),\ 0 < \beta < 1 \tag{4}$$

and

$$u(C) = \begin{cases} C^{1-\sigma}/(1-\sigma) & \text{for } \ \sigma \geq 0 \text{ and } \sigma \neq 1 \\ \ln(C) & \text{for } \ \sigma = 1 \end{cases} \tag{5}$$

It is customary to assume that the sovereign has enough instruments to implement any feasible consumption sequence as a competitive equilibrium and, thus, abstract from the problem of individual residents of the economy. This does not mean that the government necessarily shares the preferences of its constituents, but rather that it is the relevant decision maker vis-a-vis international financial markets.[e]

3.1.3 Financial Markets and the Option to Default

The sovereign issues noncontingent bonds to a competitive pool of lenders. Bonds pay a coupon every period up to and including the period of maturity, which, without loss of generality, we normalize to r^* per unit of face value, where r^* is the (constant) international risk-free rate. With this normalization, a risk-free bond will have an equilibrium price of one. For tractability, bonds are assumed to mature randomly as in Leland (1994).[f] Specifically, the probability that a bond matures next period is a constant $\lambda \in [0,1]$. The constant hazard of maturity implies that all bonds are symmetric before the realization of maturity at the start of the period, regardless of when they were issued. The expected maturity of a bond is $1/\lambda$ periods and so $\lambda = 0$ is a consol and $\lambda = 1$ is a one-period bond. When each unit of a bond is infinitesimally small and any given unit matures independently of all other units, a fraction λ of any nondegenerate portfolio of bonds will mature with probability 1 in any period. With this setup, a portfolio of sovereign bonds of measure B gives out a payment (absent default) of $(r^* + \lambda)B$ and has a continuation face value of $(1 - \lambda)B$.

We will explore the quantitative implications of different maturities, but in any given economy, bonds with only one specific λ are traded. The stock of bonds at the start of any period—inclusive of bonds that will mature in that period—is denoted B. We do not restrict the sign of B, so the sovereign could be a creditor ($B < 0$) or a debtor ($B > 0$). If $B < 0$, the sovereign's (foreign) assets are assumed to be in risk-free bonds that mature with probability λ and pay interest (coupon) of r^* until maturity. The net issuance of bonds in any period is $B' - (1 - \lambda)B$, where B' is the stock of bonds at the end of the period. If the net issuance is negative, the government is either purchasing its outstanding debt or accumulating foreign assets; if it is positive, it is either issuing new debt or deaccumulating foreign assets.

If the sovereign is a debtor at the start of a period, it is contractually obligated to pay λB in principal and r^*B in interest (coupon) payments. The sovereign has the option to default on this obligation. The act of default immediately triggers exclusion from international financial markets (ie, no saving or borrowing is permitted) starting in the next period. Following the period of mandatory exclusion, exclusion continues with constant probability $(1 - \xi) \in (0,1)$ per period. Starting with the period of mandatory exclusion and continuing for as long as exclusion lasts, the sovereign loses a proportion $\phi(g, z)$ of

[e] In particular, one interpretation of the environment is that C_t represents public spending and Y_t the available revenue that is allocated by the government.

[f] See also Hatchondo and Martinez (2009), Chatterjee and Eyigungor (2012), and Arellano and Ramanarayanan (2012).

(nondefault state) output Y. When exclusion ends, the sovereign's debts are forgiven and it is allowed to access financial markets again.

3.1.4 Timing of Events

The timing of events within a period is depicted in Fig. 2. A sovereign in good standing observes S, the vector of current-period realizations of all exogenous shocks, and decides to auction $B' - (1-\lambda)B$ units of debt, where B' denotes the face value of debt at the start of the next period. If the sovereign does not default at settlement, it consumes the value of its endowment plus the value of its net issuance (which could be positive or negative) and proceeds to the next period in good standing.

If the sovereign defaults at settlement, it does not receive the auction proceeds and it is excluded from international credit markets. Thus it consumes its endowment and proceeds to the next period in which it is also excluded from borrowing and lending. We assume that the amount raised via auction, if any, is disbursed to all existing bondholders in proportion to the face value of their bond positions, ie, each unit of outstanding bonds is treated equally and receives $q(S,B,B')(B' - (1-\lambda)B)/B'$. The implication is that as long as $B > 0$ purchasers of newly issued bonds suffer an immediate loss following default. If the sovereign defaults at settlement after purchasing bonds (ie, after a buyback of existing debt), we assume that it defaults on its new payment obligations along with any remaining outstanding debt (this is a simplification relative Aguiar et al., 2016). Thus the sovereign consumes its endowments in this case as well (and moves on to the next period in a state of financial exclusion).

Our timing regarding default deviates from that of Eaton and Gersovitz (1981), which has become the standard in the quantitative literature. In the Eaton–Gersovitz timing, the bond auction occurs after that period's default decision is made. That is, the government is the Stackelberg leader in its default decision in a period. Thus newly auctioned bonds do not face any within-period default risk and, so, the price of bonds depend only on the exogenous states S and the amount of bonds the sovereign exits a period with, B'. Our timing expands the set of equilibria relative to the Eaton–Gersovitz timing, and in

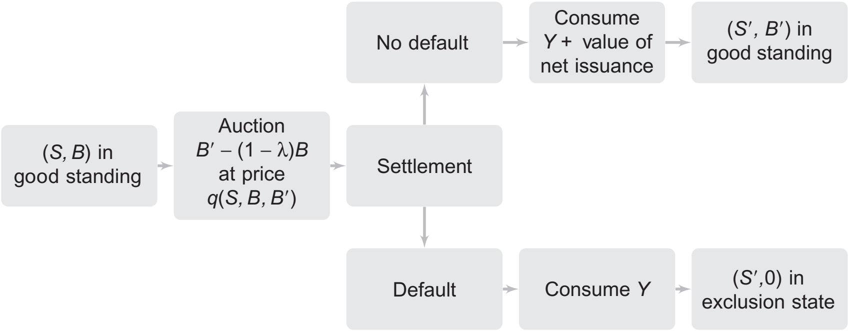

No default
Consume Y + value of net issuance
(S', B') in good standing
(S, B) in good standing
Auction $B' - (1-\lambda)B$ at price $q(S, B, B')$
Settlement
Default
Consume Y
(S',0) in exclusion state

Fig. 2 Timing within a Period.

particular allows a tractable way of introducing self-fulfilling debt crises, as explained in Section 3.5.[g] It is also worth pointing out that implicit in the timing in Fig. 2 is the assumption that there is only one auction per period. While this assumption is standard, it does allow the sovereign to commit to the amount auctioned within a period.[h]

3.2 The Sovereign's Decision Problem

We will state the sovereign's decision problem in recursive form. To begin, the vector $S \in \mathscr{S}$ of exogenous state variables consists of the current endowment Y and current period realizations of the endowment shocks g and z; it also contains W, the current period wealth of the representative lender, as this will affect the supply of foreign credit; and it contains $x \in [0,1]$, a variable that indexes investor beliefs regarding the likelihood of a rollover crisis (explained more in Section 3.5). Both W and x are stochastic and assumed to follow first-order Markov processes. We assume that all conditional expectations of the form $\mathbb{E}_S f(S', \cdot)$ encountered below are well defined.

Let $V(S, B)$ denote the sovereign's optimal value conditional on S and B. Working backwards through a period, at the time of settlement the government has issued $B' - (1-\lambda)B$ units of new debt at price $q(S,B,B')$ and owes $(r^* + \lambda)B$. If the government honors its obligations at settlement, its payoff is:

$$V^R(S,B,B') = \begin{cases} u(C) + \beta \mathbb{E}_S V(S',B') & \text{if } C \geq 0 \\ -\infty & \text{otherwise} \end{cases}. \tag{6}$$

where

$$C = Y + q(S,B,B')[B' - (1-\lambda)B] - (r^* + \lambda)B. \tag{7}$$

If the sovereign defaults at settlement, its payoff is:

$$V^D(S) = u(Y) + \beta \mathbb{E}_S V^E(S') \tag{8}$$

where

$$V^E(S) = u(Y(1-\phi(g,z))) + \beta \mathbb{E}_S \left[\xi V(S',0) + (1-\xi) V^E(S')\right] \tag{9}$$

[g] The timing in Fig. 2 is adapted from Aguiar and Amador (2014b), which in turn is a modification of Cole and Kehoe (2000). The same timing is implicit in Chatterjee and Eyigungor's ((2012)) modeling of a Cole–Kehoe type rollover crisis. In both setups, the difference relative to Cole and Kehoe is that the sovereign is not allowed to consume the proceeds of an auction if it defaults. This simplifies the off-equilibrium analysis without materially changing the results. See Auclert and Rognlie (2014) for a discussion of how the Eaton–Gersovitz timing in some standard environments has a unique Markov equilibrium, thus ruling out self-fulfilling crises.

[h] For an exploration of an environment in which the government cannot commit to a single auction, see Lorenzoni and Werning (2014) and Hatchondo and Martinez (undated).

is the sovereign's value when it is excluded from financial markets and incurs the output costs of default. Recall that ξ is the probability of exiting the exclusion state and, when this exit occurs, the sovereign reenters financial markets with no debt. Note also that the amount of new debt implied by B' is not relevant for the default payoff as the government does not receive the auction proceeds if it defaults at settlement.

Finally, the current period value function solves:

$$V(S,B) = \max\left\langle \max_{B' \leq \theta Y} V^R(S,B,B'),\, V^D(S)\right\rangle, \forall\ S \text{ and } B. \tag{10}$$

The upper bound θY on the choice of B' rules out Ponzi schemes.

Let $\delta(S,B,B')$ denote the policy function for default at settlement conditional on B'. For technical reasons, we allow the sovereign to randomize over default and repayment when it is indifferent, that is, when $V^R(S,B,B') = V^D(S)$. Therefore, $\delta(S,B,B') : \mathscr{S} \times \mathbb{R} \times (-\infty,\ \theta Y] \rightarrow [0,1]$ is the probability the sovereign defaults at settlement, conditional on (S,B,B'). Let $A(S,B) : \mathscr{S} \times \mathbb{R} \rightarrow (-\infty,\ \theta Y]$ denote the policy function that solves the inner maximization problem in (10) when there is at least one B' for which C is strictly positive. The policy function of consumption is implied by those for debt and default.

3.3 Lenders

We assume financial markets are segmented and only a subset of foreign investors participates in the sovereign debt market. This assumption allows us to introduce a risk premium on sovereign bonds as well as to explore how shocks to foreign lenders' wealth influence equilibrium outcomes in the economy, all the while treating the world risk-free rate as given. For simplicity, all period t lenders participate in the sovereign bond market for one period and are replaced by a new set of lenders.

We assume there is a unit measure of identical lenders each period. Let W_i be the wealth of an individual lender in the current period (W is the *aggregate* wealth of investors and is included in the state vector S in this capacity). Each lender allocates his wealth across two assets: the risky sovereign bond and an asset that yields the world risk-free rate r^*. Lenders must hold nonnegative amounts of the sovereign bond but can have any position, positive or negative, in the risk-free asset. The lender's utility of next period (terminal) wealth, $\widetilde{W}_i$, is given by

$$k(\widetilde{W}_i) = \begin{cases} \widetilde{W}_i^{1-\gamma}/(1-\gamma) & \text{for } \gamma \geq 0 \text{ and } \gamma \neq 1 \\ \ln(\widetilde{W}_i) & \text{for } \gamma = 1 \end{cases}.$$

Note that $\widetilde{W}_i$ is distinct from the W' that appears in S' (next period's exogenous state vector) as the latter refers to the aggregate wealth of next period's new cohort of lenders.

The one-period return on sovereign bonds depends on the sovereign's default decision within the current period as well as on next period's default decision. Let $\widetilde{D}$ and $\widetilde{D}'$

denote the sovereign's realized default decisions, either 0 (no default) or 1 (default), at settlement during the current and next period, respectively. A lender who invests a fraction (or multiple) μ of his current wealth W_i has random terminal wealth $\widetilde{W}_i$ given by

$$(1-\mu)W_i(1+r^*)+\mu W_i/q(S,B,B')\,[(1-\widetilde{D})(1-\widetilde{D}')]\,[r^*+\lambda+(1-\lambda)q(S',B',B'')],\tag{11}$$

where,

$$\begin{aligned}&\widetilde{D}=1 \text{ with probability } \delta(S,B,B')\\&\widetilde{D}'=1 \text{ with probability } \delta(S',B',A(S',B'))\\&B''=A(S',B').\end{aligned}\tag{12}$$

The wealth evolution equation omits terms that are only relevant off equilibrium; namely, it omits any payments from the settlement fund after a default. These will always be zero in equilibrium.

The representative lender's decision problem is how much sovereign debt to purchase at auction. Specifically:

$$L(W_i,S,B,B')=\max_{\mu\geq 0}\mathbb{E}_S\left[k\left(\widetilde{W}_i\right)\middle|B,B'\right],$$

subject to (11) and the expressions in (12). The solution to the lender's problem implies an optimal $\mu(W_i,S,B,B')$.

The market-clearing condition for sovereign bonds is then

$$\mu(W,S,B,B')\cdot W=q(S,B,B')\cdot B' \text{ for all feasible } B'>0,\tag{13}$$

where W is the aggregate wealth of the (symmetric) lenders. The condition requires that the bond price schedule be consistent with market clearing for any potential $B'>0$ that raises positive revenue. This is a "perfection" requirement that ensures that when the sovereign chooses its policy function $A(S,B)$, its beliefs about the prices it will face for different choices of B' are consistent with the "best response" of lenders. There are no market-clearing conditions for $B'\leq 0$; the sovereign is a small player in the world capital markets and, thus, can save any amount at the world risk-free rate.

Differentiation of the objective function of the lender with respect to μ gives an FOC that implies

$$q(S,B,B')=\frac{\mathbb{E}_S[\widetilde{W}^{-\gamma}(1-\widetilde{D})(1-\widetilde{D}')(r^*+\lambda+(1-\lambda)q(S',A(S',B')))]}{(1+r^*)\mathbb{E}_S[\widetilde{W}^{-\gamma}]}\tag{14}$$

where $\widetilde{W}$ is evaluated at $\mu(W,S,B,B')$.

Eq. (14) encompasses cases that are encountered in existing quantitative studies. As noted already, in the Eaton–Gersovitz timing of events there is no possibility of default at settlement. This means $\delta(S,B,B')=0$ and the pricing of bonds at the end of the current period reflects the possibility of default in future periods only. This means $\delta(S',B',B''(S',B'))$ does not depend on B'', only on (S',B'). Thus, q depends on (S,B') only. If lenders are risk neutral and debt is short term ($\gamma=0$ and $\lambda=1$), $q(S,B,B')$ is simply the probability of repayment on the debt next period; if lenders are risk neutral but debt is long term ($\gamma=0$ and $\lambda>0$)

$$q(S,B,B')=\frac{\mathbb{E}_S(1-D(S',B'))(r^*+\lambda+(1-\lambda)q(S',A(S',B')))]}{(1+r^*)}. \tag{15}$$

3.4 Equilibrium

Definition 1 (Equilibrium) Given a first-order Markov process for S, an *equilibrium* consists of a price schedule $q:\mathscr{S}\times\mathbb{R}\times(-\infty,\theta Y]\rightarrow[0,1]$; sovereign policy functions $A:\mathscr{S}\times\mathbb{R}\rightarrow(-\infty,\theta Y]$ and $\delta:\mathscr{S}\times\mathbb{R}\times(-\infty,\theta Y]\rightarrow[0,1]$; and lender policy function $\mu:\mathbb{R}^+\times\mathscr{S}\times\mathbb{R}\times(-\infty,\theta Y]\rightarrow\mathbb{R}$; such that: (i) $A(S,B)$ and $\delta(S,B,B')$ solve the sovereign's problem from Section 3.2, conditional on $q(S,B,B')$; (ii) $\mu(W,S,B,B')$ solves the representative lender's problem from Section 3.3 conditional on $q(S,B,B')$ and the sovereign's policy functions; and (iii) market clearing: Eq. (13) holds.

3.5 Equilibrium Selection

Because the default decision is made at the time of settlement, the equilibrium of the model features defaults that occur due to lenders' refusal to roll over maturing debt. To see how this can occur, consider the decision problem of a lender who anticipates that the sovereign will default at settlement on new debt issued in the current period, ie, the lender believes $\delta(S,B,B')=1$ for all (feasible) $B'>(1-\lambda)B$. Then, the lender's optimal μ is 0 and the market-clearing condition (13) implies that $q(S,B,B')=0$ for $B'>(1-\lambda)B$. In this situation, the most debt the sovereign could exit the auction with is $(1-\lambda)B$ and consistency with lender beliefs requires that $V^D(S)\geq V^R(S,B,(1-\lambda)B)$.[i] On the other hand, for a given stock of debt and endowment, there may be a positive price schedule that can also be supported in equilibrium. That is, if $q(s,B,\tilde{B})>0$ for some $\tilde{B}>(1-\lambda)B$ (which necessarily implies that lenders do not anticipate default at settlement for $B'=\tilde{B}$) and $V^D(S)<V^R(S,B,\tilde{B})$, the sovereign would prefer issuing new bonds to help pay off maturing debt and thus find it optimal to repay at settlement. Defaults caused by lenders offering the adverse equilibrium price schedule when a more generous price schedule that induces repayment is also an equilibrium price schedule are called *rollover crises*. A default that occurs because there is no price schedule that can induce

[i] If this condition is violated, the sovereign would strictly prefer to honor its obligation even after having acquired some small amount of new debt, contrary to lender beliefs

repayment (because endowments are too low and/or debt is too high) is called a *fundamental default*.

We incorporate rollover crises via the belief shock variable x. We assume that x is uniformly distributed on the unit interval, and we denote values of $x \in [0,\pi)$ as being in the crisis zone and values of $x \in [\pi, 1]$ as being in the noncrisis zone. In the crisis zone, a rollover crisis occurs *if* one can be supported in equilibrium. That is, a crisis occurs with $q(S, B, B') = 0$ for all $B' > (1-\lambda)B)$ if $V^R(S,B,(1-\lambda)B) < V^D(S)$*and* $x(S) \in [0,\pi)$. On the other hand, if a positive price of the debt can be supported in equilibrium, conditional on the sovereign being able to roll over its debt, then this outcome is selected if $x(S) \in [\pi,1]$. If S is such that $V^R(S,B,(1-\lambda)B) \geq V^D(S)$, then no rollover crisis occurs even if $x(S) \in [0,\pi)$. We let π index the likelihood a rollover crisis, if one can be supported in equilibrium.

We end this section with a comment on the incentive to buy back debt in the event of a failed auction, defined as a situation where lenders believe that $\delta(S,B',B) = 1$ for all $B' > (1-\lambda)B$ (either because of a rollover crisis or because of a solvency default). With a failed auction and long-term debt, the government has an incentive to buy back its debt on the secondary market if the price is low enough and then avoid default at settlement. For instance, this incentive will be strong if $q(S,B,B') = 0$ for $B' < (1-\lambda)B$. In this case, the sovereign could purchase its outstanding debt at zero cost and if

$$u(Y + (r^* + \lambda)B) + \beta\mathbb{E}_S V^R(S', B, 0) > u(Y) + \beta\mathbb{E}_S V^E(S'),$$

the sovereign's incentive to default at settlement will be gone. But, then, a lender would be willing to pay the risk-free price for the last piece of debt and outbid the sovereign for it.

To square the sovereign's buyback incentives with equilibrium, we follow Aguiar and Amador (2014b) and assume that in the case of a failed auction, the price of the debt $q(S, B, B')$ for $B' \leq (1-\lambda)B$, is high enough to make the sovereign just indifferent between defaulting on the one hand and, on the other, paying off its maturing debt and buying back $(1-\lambda)B - B'$ of its outstanding debt. Given this indifference, we further assume that the sovereign randomizes between repayment and default following a buyback, with a mixing probability that is set so that current period lenders are willing to hold on to the last unit of debt in the secondary market in the event of a buyback (more details on the construction of the equilibrium price schedule are provided in the computation section).

3.6 Normalization

Since the endowment Y has a trend, the state vector S is unbounded. To make the model stationary for computation we normalize the nonstationary elements of the state vector S by the trend component of Y_t,

$$G_t = \exp\left(\sum_1^t g_s\right). \tag{16}$$

The elements of the normalized state vector s are (g, z, w, x), where w is W/G. Since Y/G is a function of z only and z already appears in S, s contains one less element than S. It will be convenient to use the same notation defined above for functions of S for functions of the normalized state vector s. Normalizing both sides of the budget constraint (7) by G and denoting C/G by c, B/G by b and B'/G by b' yields the normalized budget constraint

$$c = \exp(z) + q(s, b, b')[b' - (1-\lambda)b] - (r^* + \lambda)b. \tag{17}$$

Here we are imposing the restriction that the pricing function is homogeneous of degree 0 in the trend endowment G and, so, denote it by $q(s,b,b')$.[j]

Next, since $u(C) = G^{1-\sigma}u(c)$, we guess $V^R(S,B,B') = G^{1-\sigma}V^R(s,b,b')$ and $V(s,b) = G^{1-\sigma}V(S,B)$. This gives

$$V^R(s, b, b') = u(c) + \beta \mathbb{E}_s g'^{1-\sigma} V(s', b'/g'). \tag{18}$$

Analogous guesses for the value functions under default and exclusion yield

$$V^D(s) = u(\exp(z)) + \beta \mathbb{E}_s g'^{1-\sigma} V^E(s') \tag{19}$$

and

$$V^E(s) = u(\exp(z)(1 - \phi(g, z))) + \beta \mathbb{E}_s g'^{1-\sigma}\left[\xi V(s', 0) + (1-\xi) V^E(s')\right]. \tag{20}$$

So,

$$V(s, b) = \max\left\langle \max_{b' \le \theta \exp z} V^R(s, b, b'),\, V^D(s)\right\rangle, \forall\ s \text{ and } b. \tag{21}$$

We denote the sovereign's default decision rule from the stationarized model by $\delta(s,b,b')$ and we denote by $a(s,b)$ the solution to $\max_{b' \le \theta \exp z} V^R(s, b, b'))$, provided repayment is feasible at (s,b).

Turning to the lender's problem, observe that given constant relative risk aversion, the optimal μ (the fraction devoted to the risky bond) is independent of the investor's wealth. Let $\mu(1,s,b,b')$ be the optimal μ of a lender with unit wealth. The FOC associated with the optimal choice of μ implies a normalized version of (14), namely,

[j] In particular, we are assuming that prices are functions of the ratios of debt and lenders' wealth to trend endowment but not of the level of trend endowment G itself. One could conceivably construct equilibria where this is not the case by allowing lender beliefs to vary with the level of trend endowment, conditional on these ratios. We are ruling out these sorts of equilibria.

$$q(s,b,b') = \frac{\mathbb{E}_s[\widetilde{w}^{-\gamma}(1-D)(1-D')(r^* + \lambda + (1-\lambda)q(s',b',a(s',b')))]}{(1+r^*)\mathbb{E}_s[\widetilde{w}^{-\gamma}]}, \tag{22}$$

where $\widetilde{w}$ is the terminal wealth of the lender with unit wealth evaluated at $\mu(1,s,b,b')$ and the expectation is evaluated using the sovereign's (normalized) decision rules.

The normalized version of the key market-clearing condition is then

$$\mu(1,s,b,b') \cdot w = q(s,b,b') \cdot b' \text{ for all feasible } b' > 0. \tag{23}$$

For a given pricing function $0 \leq q(s,b,b') \leq 1$, standard Contraction Mapping arguments can be invoked to establish the existence of all value functions. For this, it is sufficient to bound b' from below by some $\underline{b} < 0$, ie, impose an upper limit on the sovereign's holdings of foreign assets (in addition to the upper limit on its issuance of debt to rule out Ponzi schemes), and assume that $\beta \mathbb{E} g'^{1-\sigma}|g < 1$ for all $g \in \mathcal{G}$.

3.7 Computation

Computing an equilibrium of this model means finding a price function $q(s,b,b')$ and associated optimal stationary decision rules $\delta(s,b,b')$, $a(s,b)$ and $\mu(1,s,b,b')$ that satisfy the stationary market-clearing condition (23). That is, it means finding a collection of functions that satisfy

$$\mu(1,s,b,b') \cdot w = \left[\frac{\mathbb{E}_s[\widetilde{w}^{\,-\gamma}(1-\widetilde{D})(1-\widetilde{D}')(r + \lambda + (1-\lambda)q(s',b',a(s',b')))]}{(1+r^*)\mathbb{E}_s[\widetilde{w}^{\,-\gamma}]}\right] b' \;\; \forall\, s,\ b \text{ and } b'. \tag{24}$$

If such a collection can be found, an equilibrium in the sense of Definition 1 will exist in which all the nonstationary decision rules are scaled versions of the stationary decision rules, ie, $A(S,B) = a(s,b)G$, $\delta(S,B,B') = \delta(s,b,b')$ and $\mu(W,S,B,B') = \mu(1,s,b,b')wG$.

On the face of it, this computational task seems daunting given the large state and control space. It turns out, however, that (24) can be solved by constructing the solution out of the solution of a computationally simpler model. This simpler model adheres to the Eaton–Gersovitz timing, so $\delta(s,b,b') = 0$, and thus q is a function of s and b' only. But, unlike the standard Eaton–Gersovitz model, it is modified to have rollover crises.[k] The modification is as follows: If s is such that the belief shock variable $x(s)$ is in $(\pi,1]$ (ie, it is not in the crisis zone), the sovereign is offered $q(s,b')$ where b' can be any feasible choice of debt (think of this as the price schedule in "normal times"). But if $x(s)$ is in $[0,\pi]$, the sovereign is offered a truncated *crisis* price schedule in which $q(s,b') = 0$ for all $b' > (1-\lambda)b$ provided default strictly dominates repayment under the crisis price schedule;

[k] This model is described in section E of Chatterjee and Eyigungor (2012).

if the proviso is not satisfied, the sovereign is offered the normal (nontruncated) price schedule.

To see how this construction works, let $q(s,b')$ be the equilibrium price function of this rollover-modified EG model. That is, $q(s,b')$ satisfies

$$\mu(1,s,b')\cdot w = \left[\frac{\mathbb{E}_s[\widetilde{w}^{-\gamma}(1-D(s',b'))(r+\lambda+(1-\lambda)q(s',a(s',b')))]}{(1+r)\mathbb{E}_s[\widetilde{w}^{-\gamma}]}\right]b' \tag{25}$$

where $D(s,b)$ and $a(s,b)$ are the associated equilibrium policy functions. And let $V(s,b)$ and $V^D(s)$ be the associated value functions. Next, let $G(Q;s,b,b')$ be defined as the utility gap between repayment and default at settlement when the auction price is Q:

$$u[\exp(z(s))-(r^*+\lambda)b+Q(b'-(1-\lambda)b)]+\beta\mathbb{E}_s g'^{1-\sigma}V(s',b'/g')-V^D(s).$$

G encapsulates the incentive to default or repay at settlement in a model in which default at settlement is not permitted. The logic underlying the construction of the price schedule for the model in which default at settlement *is* permitted is this: If $G(s,b,b')$ evaluated at $Q = q(s,b')$ is nonnegative, $q(s,b,b')$ is set equal to $q(s,b')$, as there is no incentive to default at settlement; if $G(s,b,b')$ evaluated at $Q = q(s,b')$ is negative, $q(s,b,b')$ is set to 0 if the incentive to default is maintained at an auction price of zero, or it is set to some positive value between 0 and $q(s,b')$ for which the sovereign is indifferent between default and repayment.

1. For $b' \geq (1-\lambda)b$

$$q(s,b,b') = \begin{cases} 0 \text{ if } G(q(s,b');\, s,b,b') < 0 \\ q(s,b') \text{ if } G(q(s,b');\, s,b,b') \geq 0. \end{cases}$$

The top branch deals with the case where the sovereign's incentive to default at settlement is strictly positive after having issued debt at price $q(s,b')$. Since G is (weakly) increasing in Q in this case, the incentive to default at settlement is maintained at $Q = 0$ and, so, we set $q(s,b,b') = 0$. The bottom branch deals with the case where the sovereign (weakly) prefers repayment over default. In this case, the price is unchanged at $q(s,b')$.

2. For $b' < (1-\lambda)b$:

$$q(s,b,b') = \begin{cases} 0 \text{ if } G(0;\, s,b,b') < 0 \\ Q(s,b,b') \text{ if } Q \in [0,q(s,b')) \\ q(s,b') \text{ if } G(q(s,b');\, s,b,b') \geq 0. \end{cases}$$

The bottom branch offers $q(s,b')$ if $G(q(y,b');s,b,b') \geq 0$. If $G(q(y,b');s,b,b') < 0$, then two cases arise. Since G is weakly decreasing in Q, it is possible that there is a $Q\in [0,q(s,b'))$ for which the $G(Q;s,b,b')=0$. In this case, we set $q(s,b,b')=Q$. If there is no such Q, then $G(0;s,b,b') < 0$ and we set $q(s,b,b') = 0$.

Next, we verify that given $V(s,b)$ and $V^D(s)$ (the value functions under $q(s,b)$), the optimal action under $q(s,b)$ is also an optimal action under $q(s,b,b')$. First, consider (s,b) for which the optimal action is to choose $a(s,b)$. This implies that $G(q(s,b);s,b,a(b,s)) \geq 0$. Then, by construction, $q(s,b,b') = q(s,b)$ and the payoff from choosing $a(s,b)$ is the same as under $q(s,b)$ and this payoff will (weakly) dominate the payoff from choosing any other b' for which $q(s,b,b') = q(s,b')$ (by optimality). Furthermore, the payoff from any b' for which $q(s,b,b') \neq q(s,b)$ is never better than default. It follows that $a(s,b)$ (coupled with $\delta(s,b,a(s,b))$ $= 0$) is an optimal choice under $q(s,b,b')$. Next, consider (s,b) for which it is optimal to default under $q(s,b)$. This implies $G(q(s,b);s,b,b') < 0$ for all feasible b'. Then, by construction, default at settlement is the best option, or one of the best for all b' under $q(s,b,b')$.

Finally, we have to verify that $q(s,b,b')$ is consistent with market clearing. For (s,b,b') such that $q(s,b,b') = q(s,b)$, market clearing is ensured because the market clears (by assumption) under $q(s,b)$. For (s,b,b') such that $q(s,b,b') = 0$, market clearing is ensured trivially. For (s,b,b') such that $q(s,b,b') \in (0,q(s,b))$, market clearing can be ensured by selecting $\delta(s,b,b')$ appropriately. For instance, if lenders are risk-neutral, $\delta(s,b,b')$ is set to satisfy $q(s,b,b') = [1 - \delta(s,b,b')]q(s,b')$. Then, with probability $\delta(s,b,b')$ the sovereign defaults and the bonds are worthless, and with probability $1 - \delta(s,b,b')$, the sovereign repays and the bonds are worth $q(s,b')$. With risk-averse lenders, $\delta(s,b,b')$ can be similarly set to make lenders willing to lend b' at $q(s,b,b')$.[1]

We conclude the description of the construction of $q(s,b,b')$ by noting how it modifies the rollover price schedule under $q(s,b')$. Under $q(s,b')$, a rollover crisis is a price schedule with (a) $x(s) \in [0,\pi]$, (b) for $b \geq ((1 - \lambda)b$, $q(s,b') = 0$, and (c) $D(s,b) = 1$. Under $q(s,b,b')$, a rollover has (a) $x(s) \in [0,\pi]$, (b) for $b' \geq (1 - \lambda)b$, $q(s,b,b') = 0$ (which, in this case, is also $q(s,b')$) and (c) for $b' < (1 - \lambda)b$, $q(s,b,b')$ is given by the construction under (ii). Thus, the only modification to the crisis price schedule is to raise the prices associated with buybacks (as discussed earlier in Section 3.5).

In the rest of this section, we describe the iterative process by which the (stationary) equilibrium of the rollover-modified EG model is computed. First, the space of feasible b' is discretized. Second, the space of x (the belief shock variable) is also discretized with "crisis" equal to a value of 1, taken with probability π, and "normal" equal to a value of 0, taken with probability $(1 - \pi)$. Suppose that $\{q^k(s,b')\}$ is the price schedule at the start of iteration k. Let $a(s,b;q^k),D(s,b;q^k)\}$ be the sovereign's decision rules conditional

[1] If $\delta(s,b,b') = 0$ lenders would be just willing to lend b' at the price $q(s,b')$ (because they are willing to do so under $q(s,b')$). If the probability of default at settlement is kept at zero and the price of the bond is lowered to $q(s,b,b')$, there will be an excess demand for bonds. This excess demand can be choked off by lowering $\delta(s,b,b')$ sufficiently.

on $q^k(s,b')$. Then, for every feasible $b' > 0$ for which $q^k(s,b')b' > 0$, the price implied by the lender's optimal choice of μ and market clearing is

$$J^k(s,b')) = \frac{\mathbb{E}_s[\widetilde{w}^{-\gamma}(1-D(s',b';q^k))(r+\lambda+(1-\lambda)q^k(s',a(s',b';q^k)))]}{(1+r^*)\mathbb{E}_s[\widetilde{w}^{-\gamma}]}, \tag{26}$$

where, using (23), the $\mu(1,s,b';q^k)$ that appears in $\widetilde{w}$ is replaced by $[q^k(s,b)\cdot b']/w(s)$. If $|\max J^k(s,b') - q^k(s,b')|$ is less than some chosen tolerance $\epsilon > 0$, the iteration is stopped and the collection $\{q^k(s,b'),a(s,b;q^k),D(s,b;q^k),\mu(1,s,b';q^k)\}$ is accepted as an approximation of the equilibrium. If not, the price schedule is updated to

$$q^{k+1}(s,b') = \xi q^k(s,b') + (1-\xi)J^k(s,b'), \tag{27}$$

where $\xi \in (0,1)$ is a damping parameter (generally close to 1).

In a purely discrete model in which all shocks and all choices belong to discrete sets, the iterative procedure described above typically fails to converge for a wide choice of parameter values. The reason is that the equilibrium we are seeking is, in effect, a Nash equilibrium of a game between the sovereign and its lenders and we should not expect the existence of an equilibrium in *pure* strategies, necessarily. To remedy the lack of convergence, it is necessary to let the sovereign randomize appropriately between two actions that give virtually the same payoff. The purpose of the continuous i.i.d. shocks (z in the SG model and m in the DG model) is to provide this mixing. We refer the reader to Chatterjee and Eyigungor (2012) for a discussion of how continuous i.i.d. shocks allow robust computation of default models.

4. BENCHMARK MODELS

We calibrate two versions of the basic model under the assumption that rollover crises never happen. In one version, labeled DG, the endowment process of the sovereign and the wealth process of investors are modeled as independent stationary fluctuations around a common deterministic growth path. In the second version, labeled SG, the growth rates of endowments and investor wealth follow independent stationary processes with a common mean growth.

To calibrate the endowment process we use quarterly real GDP data for Mexico for the period 1980Q1–2015Q2. For the DG model, $G_t = (1+g)^t$ and log income is a stationary process plus a linear trend. The stationary component, z_t, is assumed to be composed of two parts: a persistent part e_t that follows an AR1 process and a purely transitory part m_t:

$$z_t = e_t + m_t, \quad m_t \sim N(0,\sigma_m^2) \text{ and } e_t = \rho_e e_{t-1} + v_t \quad v_t \sim N(0,\sigma_v^2) \tag{28}$$

As explained at the end of the previous section, the transitory shock m_t is required for robust computation of the equilibrium bond price function. We set $\sigma_m^2 = 0.000025$

and estimate (28) using standard state-space methods. The estimation gives $\rho_e = 0.85$ (0.045) and $\sigma_v^2 = 0.000139$ ($1.08e - 05$) (standard errors in parenthesis). The slope of the trend line implies a long-run quarterly growth rate of 0.56% (or annual growth rate of 2.42%).

For the SG model, the growth rate g_t is stochastic. Now, $\ln(Y_t) = \sum_0^t g_t + z_t$ and the growth rate of the period t endowment, $\ln(Y_t) - \ln(Y_{t-1}) \equiv \Delta y = g_t + z_t - z_{t-1}$. We assume

$$g_t = \alpha + \rho_g g_{t-1} + v_t, \quad v_t \sim N(0, \sigma_v^2) \text{ and } z_t \sim N(0, \sigma_z^2) \tag{29}$$

and use the observed growth rate of real GDP to estimate (29) using state-space methods. The estimation yields $\alpha = 0.0034$ (0.0012), $\rho_g = 0.45$ (0.12), $\sigma_v^2 = 0.000119$ (0.0000281) and $\sigma_z^2 = 0.000011$ ($8.12e - 06$). The estimates of α and ρ_g imply an average growth rate of 2.45% at an annual rate. These estimates are summarized in Table 6

Regarding $\phi(g,z)$, which determines the level of output under exclusion from credit markets, we assume

$$\text{for DG} : \phi(g, z) = d_0 \exp(z)^{d_1} \text{ and for SG} : \phi(g, z) = d_0 \exp(g)^{d_1}. \tag{30}$$

In either model, setting $d_1 = 0$ leads to default costs that are proportional to output. If $d_1 > 0$, then default costs rise more than proportionately with z in the DG model, and more than proportionately with g in the SG model.

We assume that g takes values in a finite set $\mathcal{G}$. In the deterministic growth case $\mathcal{G}$ is a singleton. The specification of z depends on what is being assumed for g. When g is stochastic, z is drawn from a distribution H with compact support $[-\bar{h}, \bar{h}]$ and continuous CDF. When g is deterministic, $z = e + m$, where e follows a first-order Markov process with values in a finite set $\mathcal{E}$ and m is drawn from H. In either case, z is first-order Markov in its own right (in the stochastic g case, trivially so) but it is *not* finite-state.

Aside from the parameters of the endowment process, there are 12 parameters that need to be selected. The model has 3 preference parameters, namely, β (the sovereign's discount factor), σ (the curvature parameter of the sovereign's utility function) and γ (the curvature parameter of the investors utility function). It has 2 parameters with respect to

Table 6 Parameters of endowment processes

Parameter	Description	DG	SG
–	Average annual growth rate of endowments	2.42	2.45
ρ_e	Autocorrelation of y	0.85	–
σ_v	Standard deviation of innovations to e or g	0.012	0.011
σ_m	Standard deviation of m	0.005	–
ρ_g	Autocorrelation of g	–	0.45
σ_z	Standard deviation of z	–	0.003

the bond market, namely, λ (the probability with which a bond matures), and r_f (the risk-free rate of return available to investors). It has 3 parameters with respect to the default state, namely, d_0 and d_1, the parameters of the $\phi(g,z)$ function, and ξ, the probability of reentry into credit markets from the exclusion state. Finally, there are 3 parameters governing the stochastic evolution of investor wealth w_t. For the DG version, w_t is defined as $\ln\left(W_t/\omega(1+g)^t\right)$ and for the SG version as $\ln(W_t/\omega Y_t)$, where ω controls the average wealth of investors relative to the sovereign. In either case w_t follows an AR1 process with persistence parameter ρ_w and unconditional variance σ_w^2.

Turning first to preference parameters, σ is set to 2, which is a standard value in the literature. The curvature parameter of the investor's utility function, γ, affects the compensation required by investors for default risk (risk premium). However, for any γ, the risk premium also depends on ω, as this determines the fraction of investor wealth that must reside in sovereign bonds in equilibrium. Thus, we can fix γ and vary ω to control the risk premium. With this in mind, γ was also set equal to 2.

With regard to the bond market parameters, we set the (quarterly) risk-free rate to 0.01. This value is roughly the average yield on a 3-month US Treasury bill over the period 1983–2015.[m] . The probability of a bond maturing, λ, is set to $1/8 = 0.125$ which implies that bonds mature in 2 years, on average. This is roughly consistent with the data reported in Broner et al. (2013) which show that the average maturity of bonds issued by Mexico during the Brady bonds era prior to the Tequila crisis (1993–95) was 2.5 years (postcrisis, the average maturity lengthened substantially).

The exclusion state parameters, d_0, d_1 and ξ, affect the value of the default option. The value of ξ was set to 0.125, which implies an average exclusion period of 2 years, on average. Settlements following default have generally been quick in the Brady era, so a relatively short period of exclusion seems appropriate.

Finally, we use the US P/E ratio as a proxy for investor wealth. We set the autocorrelation of the investor wealth process to 0.91, which is the autocorrelation of the P/E ratio at a quarterly frequency for the period 1993Q1–2015Q2. We assume that w takes values in a finite set $\mathscr{W}$ and its (first-order) Markov process has an unconditional mean $\omega > 0$, where ω determines the relative wealth of investors via-a-vis the sovereign.

These parameter choices are summarized in Table 7.

The remaining five parameters (β, d_0, d_1, ω, σ_w^2) are jointly determined to match moments in the data. The moments chosen are the average debt-to-GDP ratio for Mexico, the average EMBI spreads on Mexican sovereign debt, the standard deviation of the spread, the fraction of variation in Mexican spreads accounted for by the variation in investor wealth proxied by the variation in the US P/E ratio, and an annualized default frequency of 2%.[n]

[m] We use constant maturity yield computed by the Treasury and this data series begins in 1983Q3.

[n] If we date the beginning of private capital flows into emerging markets in the postwar era as the mid-1960s, Mexico has defaulted once in 50 years.

Table 7 Other parameters selected independently

Parameter	Description	Value
σ	Risk aversion of sovereign	2.000
γ	Risk aversion of investors	2.000
r_f	Risk-free rate	0.010
λ	Reciprocal of average maturity	0.125
ξ	Probability of exiting exclusion	0.125
ρ_w	Autocorrelation of wealth process	0.910

Table 8 Targets and model moments with proportional default costs

Description	Target	DG	SG
Debt-to-annual GDP	0.66	0.66	0.66
Average default freq	0.02	0.003	0.02
Average EMBI spread	0.03	0.001	0.03
R^2 of spreads on P/E	0.22	0.20	0.27

We do the moment matching exercise in two steps. First, we set the curvature parameter for default costs, d_1, to 0 so that default costs are simply proportional to output and we drop the standard deviation of spreads as a target. The results are shown in Table 8. The finding is that the SG model can be calibrated to the data quite well but the DG model could not. The DG model could get the debt-to-GDP ratio and the R^2 of the spreads on P/E regression, but the average spread and the average default frequency are an order of magnitude below their targets. These results echo those in Aguiar and Gopinath (2006).

Given the poor quantitative performance of the DG model with proportional costs, the rest of this chapter focuses on models with asymmetric default costs. We return to the proportional default cost and discuss its shortcomings in the next section after presenting our benchmark results.

5. BENCHMARK RESULTS WITH NONLINEAR DEFAULT COSTS

Table 9 reports the results of the moment matching exercise when all five parameters are chosen to match the four targets above and the standard deviation of spreads. As is evident, the performance of the DG model improves substantially and it can now deliver the target level of average spreads and default frequency.

A surprising finding is that neither model can match the observed spread volatility, which is an order of magnitude larger in the data than in the models. The finding is surprising because asymmetric default cost models have been successful in matching the volatility of spreads on Argentine sovereign bonds (the case that is most studied in the

Table 9 Targets and model moments with asymmetric default costs

Description	Target	DG	SG
Debt-to-annual GDP	0.66	0.66	0.66
Average default freq	0.02	0.02	0.02
Average EMBI spread	0.03	0.03	0.03
R^2 of spreads on P/E	0.22	0.23	0.26
SD of EMBI spread	0.03	0.005	0.002

Table 10 Parameters selected jointly

Parameter	Description	DG	SG
β	Sovereign's discount factor	0.892	0.842
d_0	Level parameter for default costs	0.075	0.068
d_1	Curvature parameter for default costs	10.0	10.0
ω	Wealth of investors relative to mean endowment	2.528	2.728
σ_w	SD of innovations to wealth	2.75	0.275

quantitative default literature). As explained later in the paper, the reason for the models' inability to match spread volatility is that neither z nor g is sufficiently volatile for Mexico (compared to Argentina) for the asymmetry in default costs to matter. Given this, the curvature parameter for default costs cannot be pinned down and we simply set it to a relatively large value and chose the remaining four parameters to match the other four targets.

The parameter values implied by this moment matching is reported in Table 10.

5.1 Equilibrium Price and Policy Functions

In this section we characterize the equilibrium bond price schedules and policy functions for debt issuance. We discuss the benchmark stochastic-growth (SG) and deterministic-growth (DG) versions of the model.

The price schedules and policy functions for our two growth cases are depicted in Fig. 3. As one can see from the first panel of the figure, the price schedules for the two different growth processes are quite similar. In both cases the price schedules are highly nonlinear, reflecting the positive feedback between the value of market access and q: the option to default lowers q for any B'/Y, which, in turn, lowers the value of market access and further increases the set of states in which default is optimal. Careful inspection will show that the DG schedule responds slightly less to an increase in debt right at the bend point.

The government's policy functions for debt issuance are depicted in the second panel of Fig. 3. These two functions exhibit an important difference. The striking fact about the

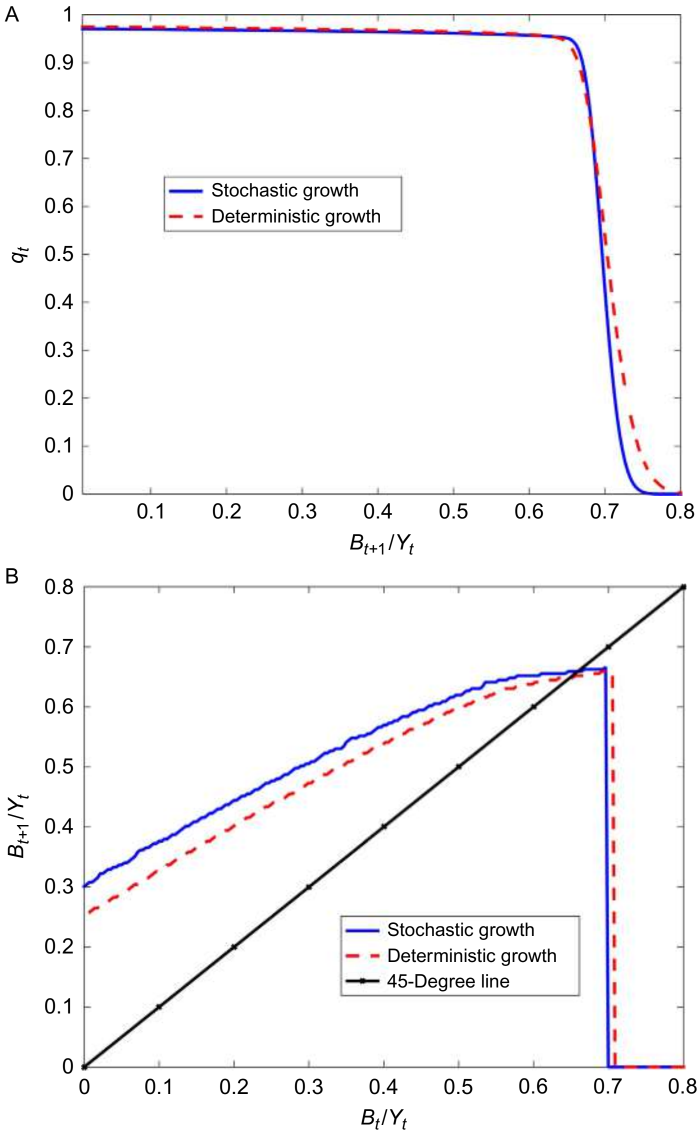


Fig. 3 Pricing schedules and policy functions. (A) Pricing schedules. (B) Policy functions.

SG debt policy functions is that it is quite flat around the 45-degree line: This implies that the optimal policy features sharp leveraging and deleveraging that offsets the impact of good and bad growth shocks, respectively, and returns B'/Y to the neighborhood of the crossing point quite rapidly. Notice also that the crossing point is not very far from the levels of debt for which default is triggered. This “distance to default,” and therefore the equilibrium spreads, are essentially determined by the output costs of default.

In contrast, the policy function for debt issuance for the DG economy depicts a significantly more modest leveraging and deleveraging response to deviations in the debt-to-output ratio around the 45-degree line. As we will see below, this will lead to sharp differences in the predicted outcomes of the two versions of our model.

We turn next to trying to understand how our model will respond to shocks. To do that we examine how our bond demand schedule responds to output and wealth shocks. These are plotted in Figs. 4 and 5, respectively. With respect to output shocks, we see a

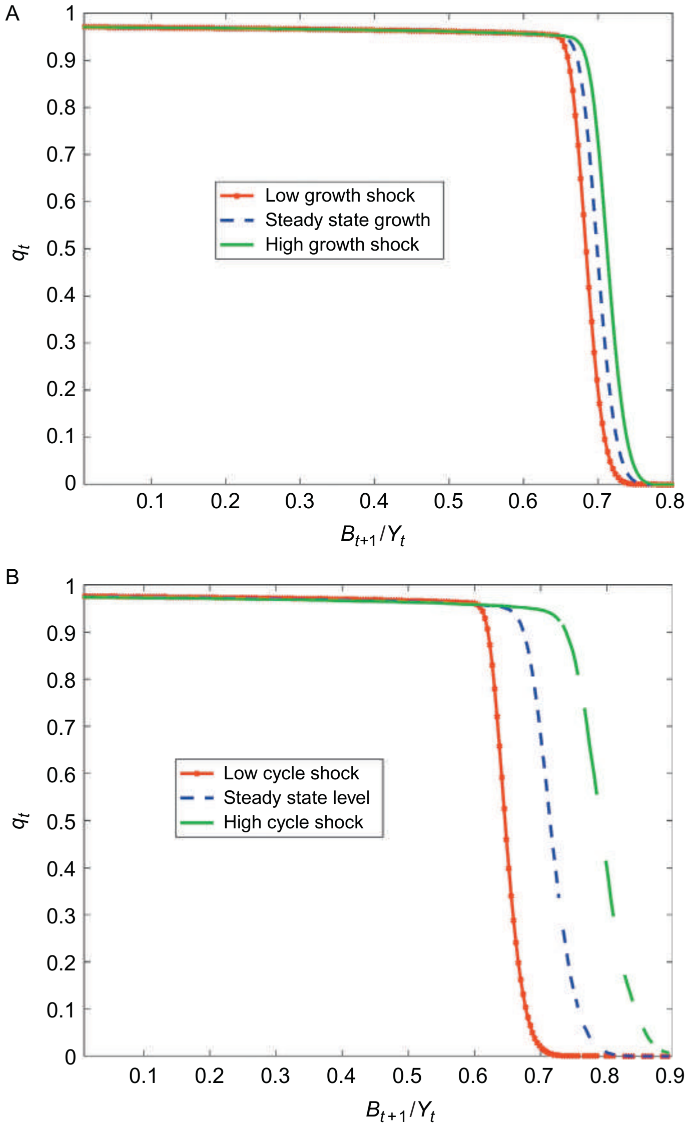


Fig. 4 Pricing schedules and output shocks. (A) Stochastic Schedule. (B) Deterministic Schedule.

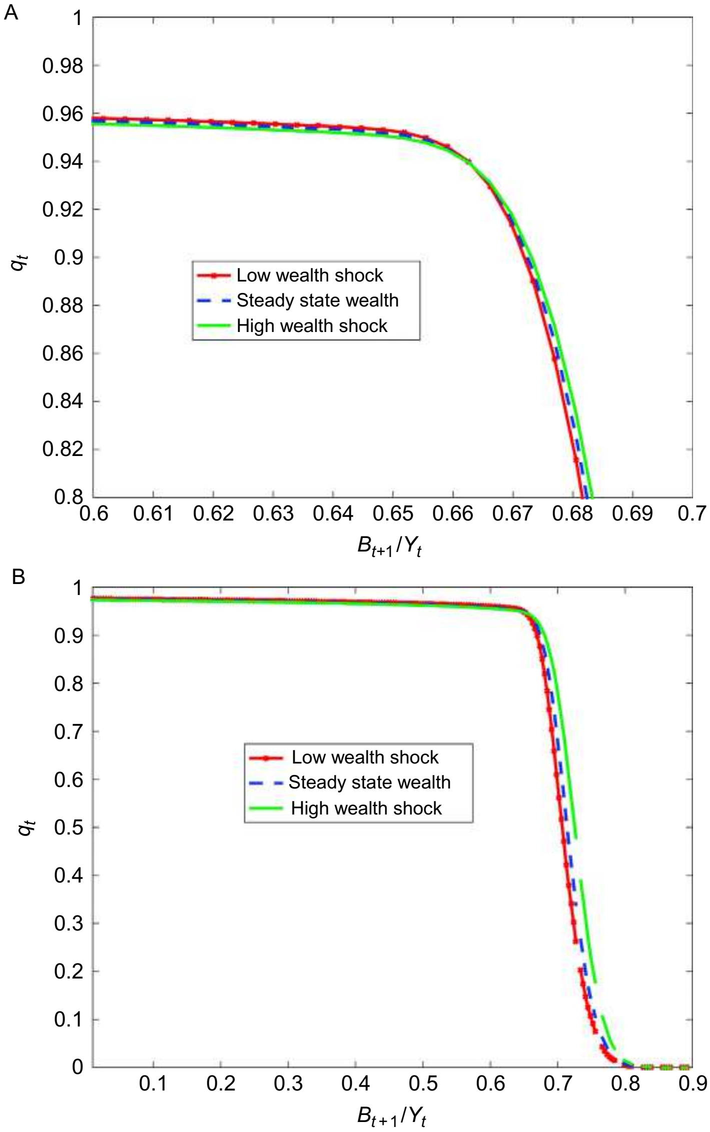


Fig. 5 Pricing schedules and wealth shocks. (A) Stochastic schedule. (B) Deterministic schedule.

fairly stark difference between our two models. Growth shocks have very little impact on the bond demand schedule in the SG model. But shocks that move output away from its deterministic trend have a fairly large effect in the DG version. This suggests that the stochastic growth version of our model will be much less responsive to output shocks than the deterministic growth version.

The reason for the difference in the response to output shocks between our two models stems from the interaction of two factors. First, when output is substantially below trend in the DG model, the agents in the economy anticipate that a recovery to trend is highly likely, making the future level of output look positive relative to the present. At the same time, our assumption of asymmetric default costs means that defaulting when output is below trend is less costly than defaulting when output has recovered to trend. Overall this creates a stronger incentive to default in the near term for given levels of B/Y and B', and this shifts in (out) the pricing schedule in response to a negative (positive) output shock. The shift in the price schedule offsets the country's desire for smoothing, but, at the same time, generates movement in the spread. Below we compare this to proportional default cost case and show that the shifts result mostly from the asymmetric default cost.

In contrast, negative growth shocks in the stochastic growth model make the expectation of future growth lower because these growth shocks are positively autocorrelated. Thus nonlinear output costs makes delaying default more attractive. In addition, the negative trajectory of output encourages the country to save, not borrow. The first effect dampens the shift in the price schedule, while the second effect dampens the incentive to borrow. Together this means that there is little or no increase in the spread today. As we will see, these differences will lead to differences in equilibrium outcomes such as the dispersion in debt-to-output levels and spreads.

Both models are quite unresponsive to wealth shocks. Interestingly, a wealth shock tends to twist the price schedule. For example, a positive wealth shock pushes out the price for high borrowing levels and but pulls it down for low borrowing levels. This last part arises from the increased incentive to dilute the current bonds in the future since the "price" of such dilution is not as high. We graphed the SG schedule on a magnified scale in order to make this twisting more apparent. This mechanism is explored in detail in Aguiar et al. (2016).

In the deterministic case, we see relatively large movements in the pricing schedule with shocks. In Fig. 6 we plot the pricing schedules for the proportional default cost case. In the DG model the price schedule does not respond to the output shock. This is because the expected positive trajectory of output makes the current debt-to-output ratio less onerous, while the proportionate default costs do not generate as strong an incentive to default today relative to the nonlinear case. Hence, the incentive to default is fairly stable and the price schedule does not shift in. At the same time, the feedback effect in the DG model with proportionate costs is so strong that the price schedule completely collapses past a certain B/Y ratio. This leads the country to stay sufficiently far inside of the collapse point that the probability of default tomorrow is virtually zero. In particular, it is very hard to generate a modest default probability and spread premium given this extreme pricing schedule. This is why this model is so hard to calibrate and why we get no volatility in the spread.

5.2 Boom-and-Bust Response

The sharp difference between our models comes from their responses to output shocks. To further understand the response of our models to growth rate shocks, we consider

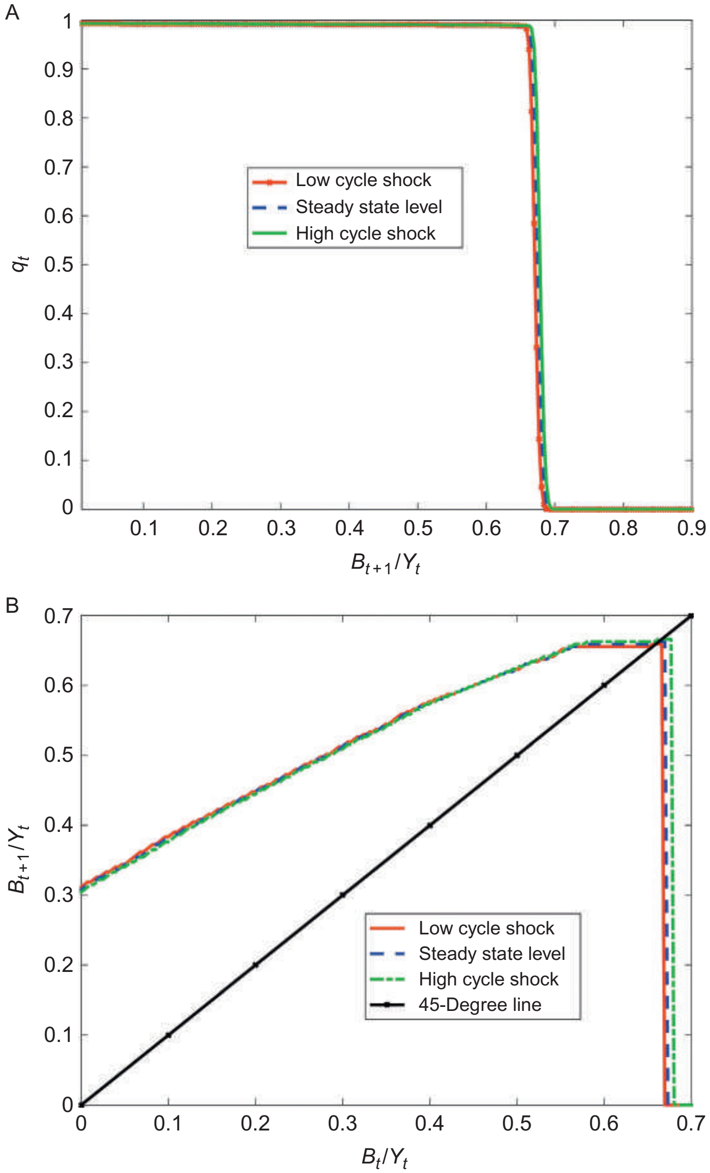


Fig. 6 DG model pricing schedule and policy function with proportionate costs. (A) Price schedule. (B) Policy function.

what happens after a sequence of positive shocks terminates in an negative shock. We refer to this as a boom-and-bust cycle.

In Fig. 7 we show the policy response to a series of positive output shocks of varying length, followed by a bad output shock. We also show the impact on the equilibrium spread. In both cases, the fairly high degree of persistence in our output shocks leads

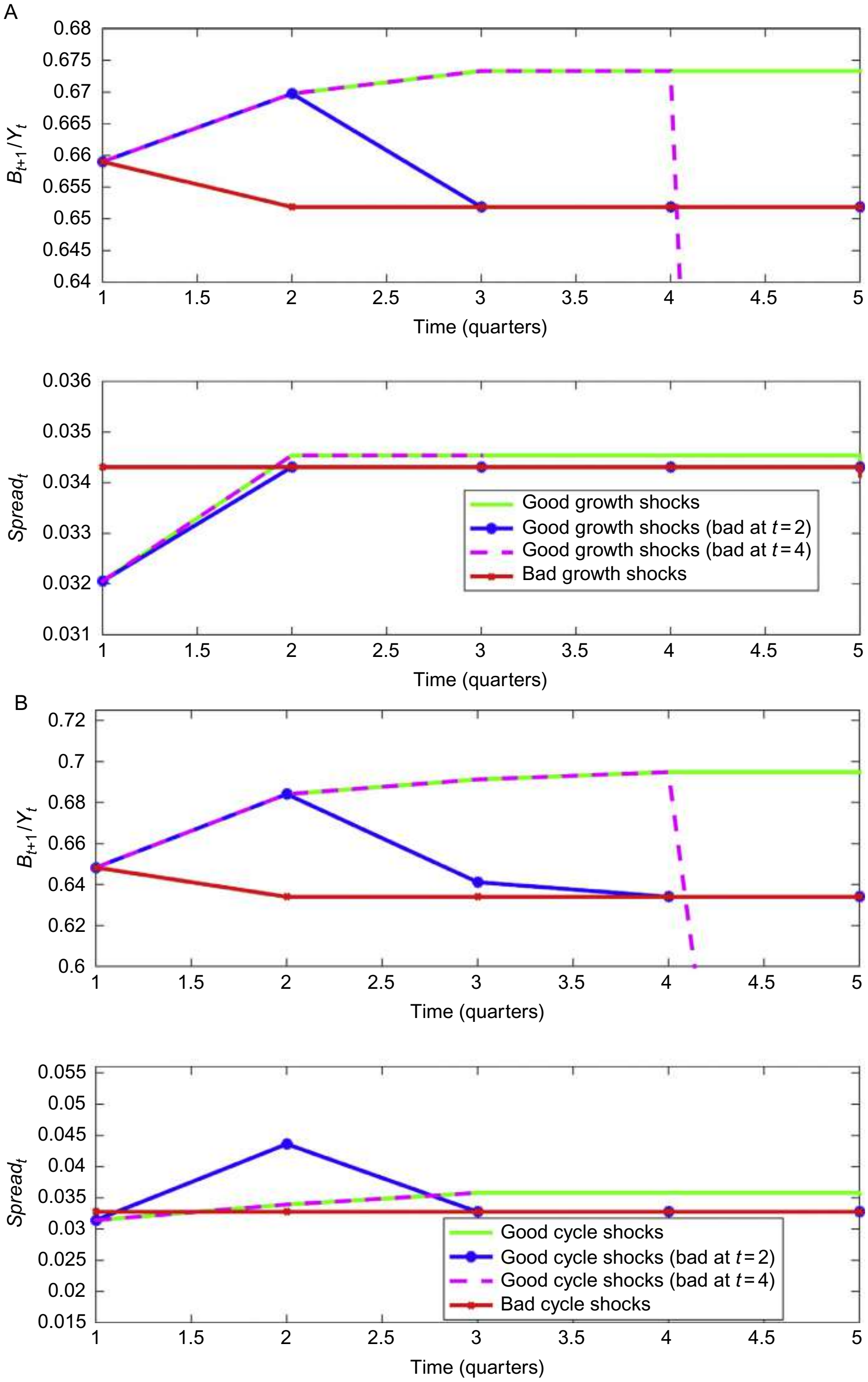


Fig. 7 Boom-and-bust cycle. (A) Stochastic. (B) Deterministic.

the government to borrow into a boom, raising the debt-to-output ratio. In the SG model, the government chooses to immediately delever in response to the negative output shock if it comes early enough in the boom; if it comes late, it defaults. The government in the DG model behaves similarly, except that it chooses to delever slightly more slowly in the case of a boom of intermediate length.

The spread behaves somewhat differently across the two versions of our model. In the SG version, the spread initially falls in respond to a positive output shock, but then it bounces back to essentially the same level as before in response to continued positive growth rate shocks because of the government's decision to lever up. More important, even in the period in which a negative growth rate shock first occurs, the government's decision to sharply delever means that the spread does not change in response to the negative shock. While the policy response of the government in the DG model is very similar to that of the SG model, the slightly slower deleveraging in response to a negative output shock leads to a sharp temporary rise in the spread.

5.3 Equilibrium Outcomes

In this section we lay out the results for both versions of our model with nonlinear output loses. Our first set of results are presented in Table 11. The first three statistics, which were targeted, match the long-run data for Mexico and are in the ball park for other emerging economies. The sixth statistic we report is the R^2 of a regression of the spread on the investor wealth shock w. This too is targeted to match the results of the regression of the spread on the US price-earnings ratio and is roughly in line with the data.

There are two nontargeted moments in Table 11. The first is the correlation of the average excess return and the growth rate of output. For the stochastic growth economy, the sign of this correlation is positive, which is surprising, since one would expect positive growth rate shocks to lower the spread. However, the magnitude of this correlation is in the ball park in that the correlation is quite weak as it is in the data. In the DG model this correlation is both of the wrong sign and also substantially higher. This reflects that

Table 11 Basic statistics: Stochastic and deterministic growth models

	Stochastic benchmark	Deterministic benchmark	Deterministic Argentina
Debt-to-GDP	0.66	0.66	0.28
Average default freq.	0.02	0.02	0.04
Average spread	0.03	0.03	0.06
SD of spreads	0.002	0.004	0.07
Corr of spreads with Δy or z	0.15	0.46	−0.76
R^2 of spreads on w	0.26	0.17	0.01

economy's greater responsiveness to output shocks, which we discussed earlier in reference to Fig. 4. Below we more closely examine the evidence on spreads and shocks using regression analysis to compare model and data results.

The other nontargeted moment is the standard deviation of the spread. This moment is too low, since it should be roughly equal to the average level of the spread. The fact that the spread's relative variation was still so low even with nonlinear default costs is surprising given that the literature has found that such costs can generate relatively realistic variation levels. However, the papers that have found this result have been calibrated to Argentina, which has a much more volatile output series.

To examine whether this might be at the root of our failure, we examined the implications of the DG model when we calibrate output to Argentina. When we calibrate our output process to Argentina, the autocorrelation coefficient for our output deviation from trend, z_t, rises from 0.853 to 0.930, thereby becoming more persistent. In addition, the standard deviation of z rises from 0.023 to 0.074, so the output deviations from trend are more volatile overall. All of the other model parameters are left unchanged. We report the results from this experiment in the last column of Table 11.

When we switch to the Argentine growth process for the deterministic model, the average debt-to-output level falls sharply, to 0.28, which is somewhat inconsistent with the fact that Argentina has a much higher value of this ratio than Mexico. In addition, the average spread rises sharply, to 0.06, and the volatility of the spread increases to 0.07. Both of these changes are consistent with the data in that Argentina has a much higher average spread and a much more volatile spread. This last finding indicates that the key to the literature's positive finding on spread volatility is the combination of nonlinear default costs and quite high output volatility. However, this story cannot explain the spread volatility in a country like Mexico with lower output volatility.

One other stark difference between the results with the Mexico and the Argentina output calibrations concerns the correlation of the spread and the percent deviation of output from trend. This has now become very negative. In Table 2 the average correlation in our sample was −0.27, and the highest value was only −0.56 for Malaysia. The correlation in Argentina was −0.35 and in Mexico it was −0.4. So a value of −0.76 with the Argentine calibration for output looks too high. Below in the regression analysis, we examine more closely the extent to which this success comes at the price of making spreads too dependent on output fluctuations.

The ergodic distributions of the debt-to-income ratio and the spread is depicted in Fig. 8. For the stochastic growth case, both the debt-to-income and the spread distributions are very tight and symmetric around their mean. The distribution of the debt-to-income ratio for the DG case is also symmetrical, but it is substantially more dispersed. For the spread distribution, the deterministic growth distribution is not completely symmetric and is again substantially more dispersed than the stochastic case. The greater dispersion in the debt-to-GDP ratio and the spread in the deterministic growth model is

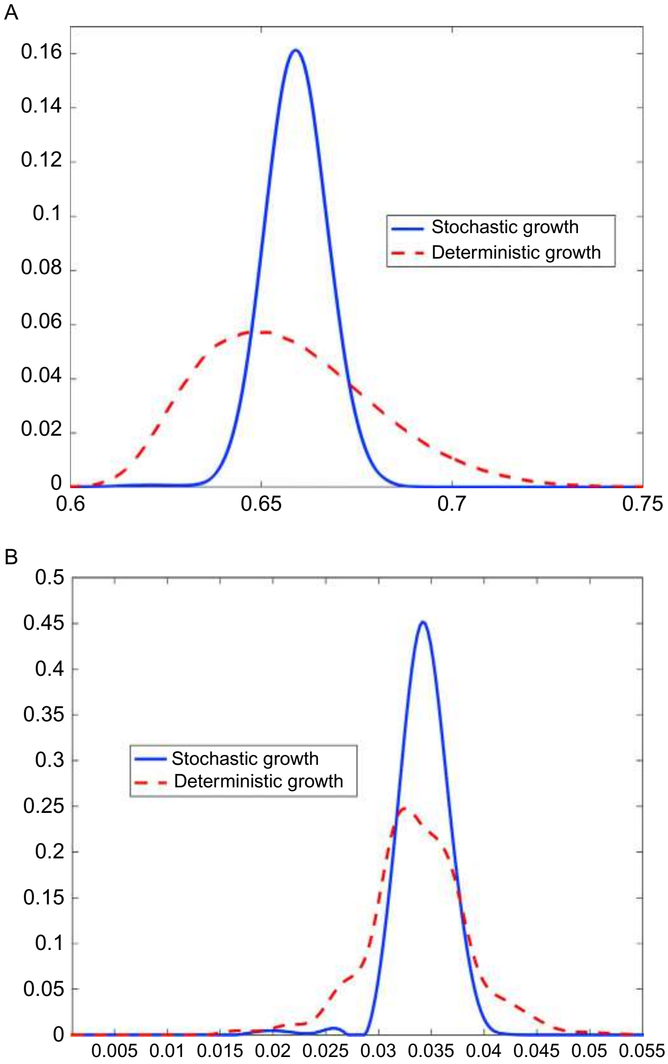


Fig. 8 Ergodic distributions. (A) Debt-to-income. (B) Spreads.

consistent with our earlier observation that the deterministic economy was more responsive to output shocks.

This spread can be decomposed into a default premium and a risk premium. Specifically, the risk premium is the standard difference between the expected implied yield on sovereign bonds and the risk-free interest rate. The default premium is the promised yield that would equate the expected return on sovereign bonds (inclusive of default) to a risk-free bond; that is, the yield that would leave a risk-neutral lender indifferent. The top panel of Fig. 9 depicts the risk premium and the bottom panel depicts the default

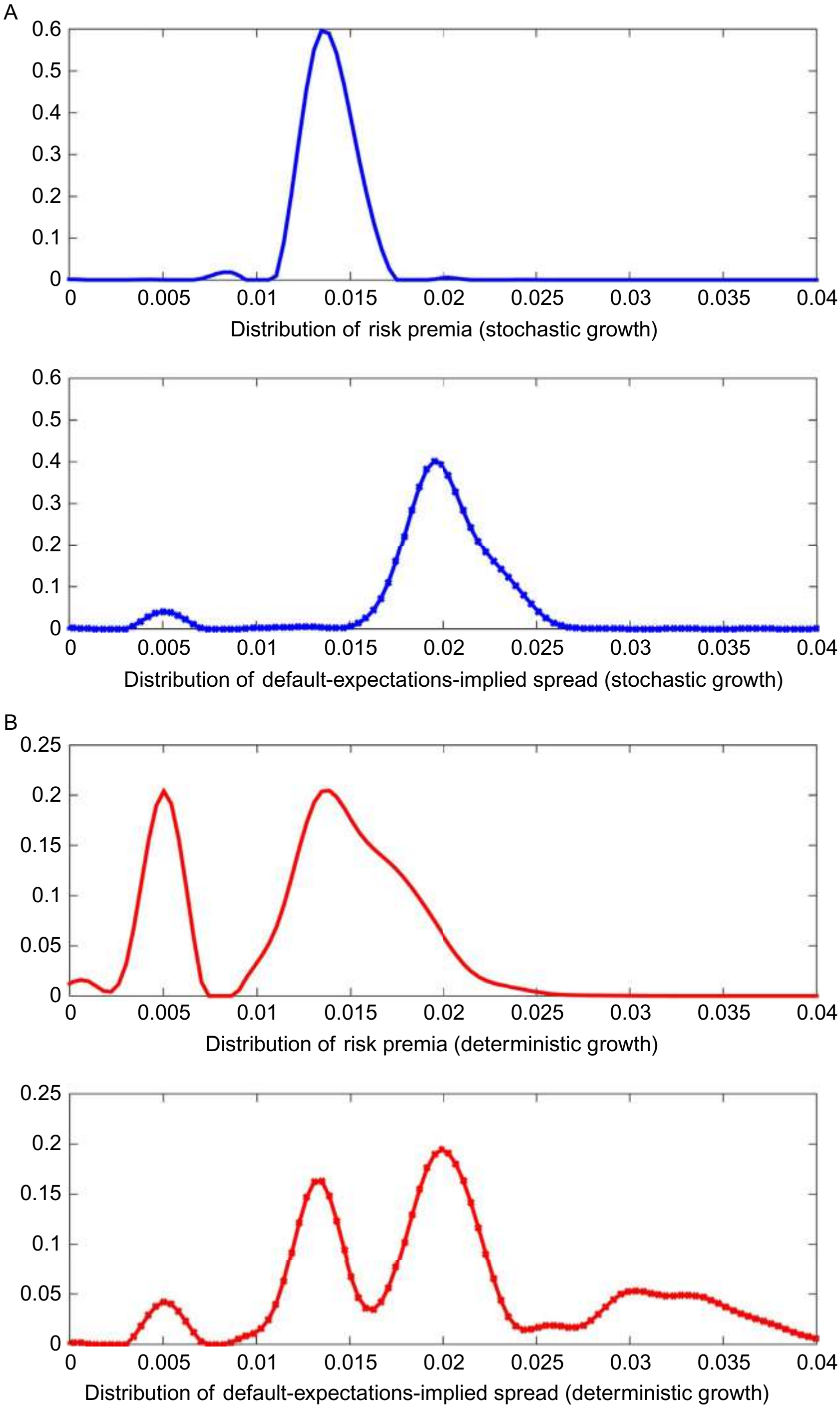


Fig. 9 Decomposition of the spread. (A) Stochastic growth. (B) Deterministic growth.

Table 12 Default and crisis statistics for the nonlinear default cost economies

	Def. share with output collapse	Def. Share with *w* collapse	Crisis share with output collapse	Crisis share with *w* collapse
Stochastic	0.80	0.02	0.31	0.01
Deterministic	0.60	0.06	0.66	0.03

premium. In both cases the risk and default spreads quite similar to each other, suggesting that the two are moving closely in parallel. On average, roughly 60% is the default premium and the rest is risk premium. This reflects our calibration target of 3% average spread and 2% default probability.

To understand the circumstances in which we are getting defaults and crises in our models, we examine the share of defaults and crises with large negative output changes and large negative investor wealth shocks. These negative changes are 1.5 standard deviations relative to the unconditional distribution. We use negative growth rate realizations for output so we are using the same metric for both models. The results are reported in Table 12. The results imply that, in the SG model, defaults are almost always associated with negative growth rate shocks and almost never with negative wealth shocks. In the deterministic growth model, the dependence of defaults on negative output shocks is a bit weaker and investor wealth shocks play essentially no role. When we turn to spread crises, we see much less dependence on growth shocks in the SG model and again essentially no dependance on wealth shocks. This is because a very negative growth shock leads to either an immediate default or rapid deleveraging. In contrast, in the DG model, the dependence of spread crises on growth shocks is even higher than it is for defaults.

5.4 Simulation Regressions

To compare the model to the data more closely, we take our model-simulated data and regress the spread on a constant and our three shocks. Besides the benchmark versions of SG and DG, we also included the results when we calibrate the output process to Argentina in the DG case. The results are in Table 13. We have already reported the results of estimating our statistical model in Table 3. However, those regressions included our two common factors. To make a closer comparison with the model regressions, we examine regressions for several of our countries with just the financial controls we considered in decomposing the common factors. We believe that including these financial controls as important in making this comparison. In our model data the output and wealth shocks are orthogonal by construction. In the actual data, an important concern is the feedback from interest rate or risk premium shocks to growth (as emphasized by Neumeyer and Perri, 2005).

Table 13 Spread regressions with wealth (simulated data)

	B_t/Y_t	g_t or z_t	w_t	R^2
SG benchmark calibration				
Coefficient	0.0286	0.0191	0.0070	
Var decomp	0.3850	0.0154	0.1660	0.5663
DG benchmark calibration				
Coefficient	0.0412	−0.0707	9.2928e−4	
Var decomp	0.3016	0.1145	0.1323	0.5484
DG argentina calibration				
Coefficient	0.307443	−0.77599	−0.00067	
Var decomp	0.030814	0.532024	0.000354	0.563191

While our two benchmark models were calibrated to Mexico, which we view as representive of countries subject to sovereign debt crises, the data series are fairly short to evaluate these somewhat rare events. Hence, it is useful to compare our model regression results to a range of countries in the data. To aid in this comparison, we also consider the DG version of our model with a growth process calibrated to Argentina.

In the SG model the output shock is the growth rate, or g_t, while in the DG model it is the deviation from trend, or z_t. To make a consistent comparison to the data-based regressions, we did them both with the growth rate of output as the shock and with the deviation of log output from a linear trend. These results are reported in Tables 14 and 15.

When we examine the results for the SG benchmark model with nonlinear default costs, one sees that the debt-to-output ratio has a positive coefficient and is explaining 38% of the movements in the spread as measured by the marginal R^2. This finding is consistent with the data regressions where this variable always has a positive coefficient and explains almost half of the spread in three of our countries and virtually nothing in two of them. The marginal R^2 for the growth rate shock is 0.01, which is very consistent with our growth rate regressions in Table 14 and the sign of that coefficient is positive. The wealth shock explains 17% of the variation according to the marginal R^2. This too is consistent with the data, since in some countries the financial variables explain very little, and in several others, particularly Mexico, they explain a great deal.

There are two major surprises in the SG model regression. First, the sign of the output shock is positive in the SG model, indicating that positive growth rate shocks raise the spread. This is contrary to the sign of this term in the data regressions. However, this result seems consistent with the results we showed for a boom-bust cycle in Fig. 7. There, only the initial response to a good output shock was negative while a sequence of good output shocks led the government to raise its debt-to-output ratio and thereby induce an

Table 14 Spread regressions (data): output shock = growth rate

Country	B_t/Y_t	g_t	**VIX**	**P/E ratio**	**LIBOR**	R^2
Argentina:						
Coefficients	0.0067 (9.9307*e*−4)	−1.0480 (0.6770)	7.8592*e* − 4 (0.0013)	0.0034 (0.0046)	−0.0372 (0.0072)	
Var decomp	0.4962	0.0120	0.0059	0.0085	0.0880	0.6105
Brazil:						
Coefficients	0.0026 (3.1092*e*−4)	−0.3134 (0.2297)	0.0013 (4.4568*e*−4)	1.8695*e* − 4 (8.1841*e*−4)	0.0023 (0.0014)	
Var decomp	0.4943	0.0150	0.0537	0.0093	0.0482	0.6204
Colombia:						
Coefficients	0.0018 (1.5892*e*−4)	−0.1535 (0.1102)	0.0011 (1.2462*e*−4)	7.5692*e* − 4 (3.0964*e*−4)	5.2909*e* − 4 (5.3118−4)	
Var decomp	0.4900	0.0236	0.2594	0.1017	0.0062	0.8809
Mexico:						
Coefficients	7.2889 − 4 (2.2988−4)	−0.1595 (0.0467)	6.2858*e* − 4 (6.0880*e*−5)	6.4423*e* − 4 (1.2801*e*−4)	1.1697*e* − 4 (3.5179*e*−4)	
Var decomp	−0.0226	0.1350	0.6598	0.1212	−0.0087	0.8847
Russia:						
Coefficients	2.5708*e* − 4 (8.0133−4)	−0.7400 (0.7191)	0.0025 (0.0015)	0.0117 (0.0031)	0.0058 (0.0075)	
Var decomp	0.0210	0.0109	0.0540	0.3217	0.0696	0.4771
Turkey:						
Coefficients	0.0012 (1.7406−4)	−0.2489 (0.0660)	7.3488*e* − 4 (1.9433*e*−4)	0.0028 (3.4963*e*−4)	4.8343*e* − 4 (7.1599*e*−4)	
Var decomp	0.1520	0.0911	0.1413	0.3847	0.0068	0.7759

Table 15 Spread regressions (data): Output shock = deviation from trend

Country	B_t/Y_t	z_t	**VIX**	**P/E Ratio**	**LIBOR**	R^2
Argentina						
Coefficients	0.0058 (0.0016)	−22.2293 (38.4213)	0.0014 (0.0013)	0.0011 (0.0062)	−0.0384 (0.0080)	
Var decomp	0.2463	0.1060	0.0138	0.0098	0.2080	0.5839
Brazil						
Coefficients	0.0027 (0.0004)	11.6322 (12.8857)	0.0015 (0.0004)	0.0005 (0.0009)	0.0021 (0.0014)	
Var decomp	0.3778	0.0638	0.0512	0.0657	0.0564	0.6150
Colombia						
Coefficients	0.0015 (0.0002)	−19.6663 (7.9572)	0.0011 (0.0001)	0.0009 (0.0003)	0.0013 (0.0006)	
Var decomp	0.3178	0.1130	0.2353	0.2000	0.0245	0.8903
Mexico						
Coefficients	0.0007 (0.0002)	−4.8005 (3.3338)	0.0007 (6.0951e−5)	0.0006 (0.0001)	0.0006 (0.0005)	
Var decomp	0.0371	0.1085	0.5613	0.1058	0.0473	0.8599
Russia						
Coefficients	$-7.0e-4$ (0.0006)	−96.9253 (17.1416)	0.0027 (0.0013)	0.0024 (0.0030)	0.0185 (0.0051)	
Var decomp	0.0705	0.2624	0.0494	0.1642	0.1072	0.6536
Turkey						
Coefficients	0.0009 (0.0002)	−18.3784 (4.1594)	0.0008 (0.0002)	0.0013 (0.0005)	0.0027 (0.0008)	
Var decomp	0.0956	0.2719	0.1433	0.2271	0.0519	0.7898

increase in the spread. Note that this response is not present in the DG model. Instead, because the government was slower to delever, a sequence of positive shocks followed by a negative one led to a temporary jump upwards in the spread.

Second, the sign of the wealth factor is positive, indicating that an increase in investor wealth, which should lower risk pricing holding everything else fixed, actually raises the spread. This result is consistent, however, with our earlier surprise finding that the sign of the P/E ratio in the data regressions is positive, indicating that a fall in the risk premium in the data also raises the spread. We will seek to better understand this finding in our quantitative exercises below.

In the simulated data regressions from the DG benchmark and DG Argentine models we also see that the debt-to-output ratio explains 30% of the variation in output and that the sign of this term is positive. However, if we compare this explanatory power to the regressions in Table 15, this is high relative to what we find when we take the output shock to be a deviation from trend. The sign on the deviation is negative, as one would expect and as we see in the data. In the DG benchmark the explanatory power of the output shock is only 11% which is consistent with the regression results. However, the explanatory power of this variable under the Argentine growth process is over 50%, which is much higher than anything we see in the data regressions. Thus it does seem like the ability of the nonlinear output cost element to increase the spread volatility when the variability of output is sufficiently high comes at the expense of tying the spread much too closely to output fluctuations. In addition, the sign of the wealth term changes when we move from the benchmark to the Argentinian output calibration. However, the positive sign in the benchmark case is consistent with the positive sign of the P/E ratio in the data regressions.

5.5 Comparative Experiments

We want to examine how the equilibrium predictions of our two benchmark models respond to changes in several key parameters. This will help us understand exactly what is driving our outcomes. In these experiments we change only the parameter in question, and we explicitly do not recalibrate the other parameters. The results are given in Table 16.

The first set of results in column 2 examines the impact of shortening the average maturity from 2 years to 1 quarter. In both the SG and the DG versions, this shortening of the maturity sharply reduces the default rate and the average spread almost to zero. This occurs because with debt that matures in a single period, future debt issuance has no effect on the value of bonds currently being issued. With longer maturity bond this is not the case and future issuances dilutes the value of current debt. Since capital loss on outstanding bonds from new issuance of debt is not borne by the sovereign, long maturity bonds induce over-borrowing and higher default risk. Put differently, with short maturity debt,

Table 16 Comparative statistics: Stochastic and deterministic growth models

	Stochastic growth				
	Benchmark	**Short maturity**	**High risk aversion**	**i.i.d. *w***	**i.i.d. *g***
Debt-to-GDP	0.66	0.68	0.66	0.66	0.78
Average default freq.	0.02	0.007	0.001	0.02	0.006
Average spreads	0.03	0.002	0.03	0.03	0.01
SD of spreads	0.002	0.001	0.002	0.002	0.002
Corr of spreads with Δy	0.15	0.15	0.14	0.17	−0.23
R^2 of spreads on w	0.26	0.008	0.43	0.003	0.29
	Deterministic growth				
	Benchmark	**Short maturity**	**High risk aversion**	**i.i.d. *w***	**Low auto. *z***
Debt-to-GDP	0.66	0.67	0.65	0.66	0.87
Average default freq.	0.02	0.002	0.01	0.02	0.003
Average spreads	0.03	0.003	0.03	0.03	0.007
SD spreads	0.004	0.001	0.005	0.004	0.001
Corr of spreads with z	0.46	0.09	0.39	0.51	−0.21
R^2 of spreads on w	0.17	0.01	0.36	0.001	0.23

the government is forced to internalize the full cost of a rise in default risk and therefore chooses to constrains its borrowing.

The second set of results concerns the impact of risk aversion on our equilibrium outcomes. In both the SG and DG cases, the frequency of default falls sharply. However, the increase in the price of risk just offsets this drop, so the average spread stays roughly unchanged. This indicates the greater discipline imposed on sovereign's borrowing behavior from a higher risk aversion on part of lenders. The greater discipline comes from the fact that the spread required per unit of default risk is higher with greater risk aversion, making default risk much more expensive for the sovereign. As a result, the sovereign optimally chooses to lower its expected future default risk. This result can also sheds light on why an increase in w raised the spread rather than lowering it. Future risk pricing can discipline future behavior. How strong that is will determine the extent to which it shows up as an increase or a decrease in the spread today. But it will increase the frequency of defaults.

The third set of results concerns the impact of making wealth shocks i.i.d. In this case, the disciplining affect of having a high future price of risk because of a low value of w today is removed. In the benchmark cases, this future discipline led to a twisting of the price schedules. When the debt-to-output is low, the future disciplinary effect dominates the static risk pricing effect and, as a result, a high w shocks lowers the price of debt. When the debt-to-output ratio is high, the static pricing effect dominates and rise in w increases the price of debt (see Fig. 5). With i.i.d. w, this twisting effect is gone and an increase in w strictly increases q where it is below the risk-free rate. In both the SG and DG models this leads to a sharp fall in the impact of wealth shocks on the spread as measured by the R^2. Consistent with this, the correlation of the wealth shock and the spread goes from 0.15 in the benchmark to 0.002 with i.i.d. w in the SG model and from 0.40 to 0.03 in the DG model.

The final set of results concerns the impact of autocorrelation of output shocks. For the SG model we reduce the correlation in the output growth rate g from 0.45 to 0, and in the DG model we reduce the correlation in the deviations from trend from 0.85 to 0.45. In both models the debt-to-output ratio goes up as the hedging motive goes up. In both models the default frequency goes down as the likelihood of a sequence of bad shocks driving a country into default goes down. In addition, in both models the incentive to borrow into a boom goes down as the likelihood of the good times continuing is reduced. As a result, the correlation of the spread and the growth rate of output is now negative in both models. At the same time spreads and default frequencies fall in both models.

5.6 Taking Stock

Our models of sovereign borrowing, default and the spread can match a number of key facts in the data. They can match the overall borrowing level, but this comes at the expense of assuming that default costs are large so that we can get the sovereign to repay, and that the sovereign is fairly myopic since borrowing and occasionally defaulting is, as we noted earlier a poor way of getting insurance.

Risk aversion on part of lenders leads to the average spread being greater than the average frequency of default, hence lenders earn a positive risk premium of about 1%.

The sovereign tends to borrow into booms, which is consistent with the boom-bust cycle we observe in many emerging economies. Also, the end of the boom is associated with a sudden shift in the price schedule for debt, which resembles the lending cutoff (sudden stops) observed in the data. This borrowing into booms depends on future optimism, which here comes through the autocorrelation in output shocks. If we make growth rates i.i.d. in the SG model or reduce the persistence of deviations from trend in the DG model, borrowing-into-booms effect largely goes away. This in turn leads to a sharp fall in the frequency of default and therefore the spread.

When we compare the spread regressions in the model simulated data with those in the data, the overall behavior is broadly consistent with that observed in the data. For both the SG and DG benchmark models, the importance of the debt-to-output ratio and the output shock is consistent with the regression results. However, the positive impact of a growth shock on the spread in the SG model is not consistent with the negative sign of the coefficient on this variable in the regression. This indicates that the reliance on a boom-bust cycle as opposed to the smoothing of consumption is excessive in this version of the model.

Global risk pricing shocks, which we model as shocks to the wealth of investors, have a surprisingly limited impact in our model. Interestingly, an increase in lenders' risk aversion that stems from a decrease in their wealth leads to a *fall* in the spread. A similar impact occurs when we increase investors' risk aversion in our comparative statics exercises. This result comes through the higher price of debt issuance, which lowers the extent to which current lenders need to worry about the dilution of the value of their claims in the future. The threat of future dilution goes away with short maturity debt. This is why we see a sharp fall in default rates and spreads when we switch to one-period debt.

The impact of persistent wealth shocks stemming from changes in borrowing discipline in the future leads to one of the surprising empirical successes of our models. In our spread regressions, a decrease in the price of risk increases the P/E ratio, but increases in the P/E ratio are associated with increases, not decreases, in the spread on emerging market sovereign bonds. This inverse relationship between the price of risk and spreads is predicted by both models. In our comparative statics exercise we saw that this correlation essentially goes away when wealth shocks become i.i.d., confirming that the inverse relationship is driven by anticipation of changes in future borrowing behavior.

The major failure of our benchmark models is with respect to the volatility of the spread. It is much too low in the model relative to the data. This indicates that the levering/delevering response to output shocks is too strong, resulting in a spread that is too smooth. This was particularly true in the initial version of our model with proportionate output costs, but is still true when we switch to nonlinear output costs of default (which improves the insurance offered by defaulting).

Increasing the variance of the output process in the DG model can substantially increase the variance of the spread, bringing it in line with the data for most countries. However, this positive result comes at a cost. First, it implies that the model cannot account for counties in our sample, such as Mexico, which have less volatile output processes. Additionally, relative to the data, higher volatility leads to too strong a dependence of spreads on output shocks.

These results suggest that what is needed is:

1. An additional shock to the pricing of debt that is not tied to country fundamentals or global risk pricing factors. This is indicated by the importance of the two common factors in the spread regressions and their lack of dependance on global asset pricing factors.
2. A reduction in the levering/delevering incentive or at least a drawing out of debt crises, which leads to high levels of the spread in response to these crises.

6. ROLLOVER CRISES

Our model was constructed to allow for rollover crises along the lines of Cole and Kehoe (2000). Here we conduct a preliminary investigation of the potential for rollover crises to add to the volatility of the spread in our models without tying this volatility too tightly to country fundamentals.

Rollover crises emerge from investors' failure to coordinate their beliefs on the good equilibrium outcome in which the government is offered a generous price schedule and therefore chooses to not default. Instead, investors adopt pessimistic beliefs about government's behavior, which leads them to offer an adverse price schedule—specifically, a zero price for new issuance of bonds—and this, in turn, induces the government to default. The government's default then validates the investors' pessimistic beliefs. What is empirically attractive about this mechanism is that while requiring that the country's fundamentals be bad enough to generate a default under the adverse price schedule, it allows relatively wide latitude in the timing of a sovereign debt crises.

In constructing a quantitative model of rollover crises, the first question is: what is a plausible process for beliefs? Beliefs, unlike, say, output, cannot be directly observed, and hence its impact and its stochastic evolution must be inferred. Aguiar et al. (2016) estimate shifts in beliefs from spreads. Another alternative is to adopt a state-space approach in which the belief process and it's realizations are estimated jointly along with other parameters of the model as in Bocola and Dovis (2015). A related alternative would be to construct belief processes that replicated the impact and time series properties of the common factors estimated in the spread regressions reported earlier. However, undertakings such as these are beyond the scope of a handbook chapter. So, instead, we follow Cole and Kehoe (2000) and its quantitative implementation in Chatterjee and Eyigungor (2012) and assume that there is a constant probability of a crisis. This limits the empirical scope of self-fulfilling rollover crises, but does allow us to partially gauge their potential impact. Also, we do not recalibrate the models so this too is a quantitative comparative statics exercise.

The results are presented in Table 17 along with our baseline results (for the nonlinear output loss from default). Here, we assume that if a country is in the crisis zone (ie, a rollover crisis can be supported in equilibrium) then a rollover crisis transpires with a 20% probability. Several results stand out. First, the possibility of rollover crises reduces

Table 17 Stochastic and deterministic growth models: Benchmark vs rollover crises

	Stochastic benchmark	**Stochastic w. RC**	**Deterministic benchmark**	**Deterministic w. RC**
Debt-to-GDP	0.66	0.63	0.66	0.65
Average default freq.	0.02	0.02	0.02	0.02
Average spread	0.03	0.04	0.03	0.04
SD of spreads	0.002	0.002	0.004	0.004
Corr of spreads with Δy or z	0.15	0.06	0.46	0.11
R^2 of spreads on w	0.26	0.18	0.17	0.09
Share of rollover defaults	0	0.70	0	0.30

the average debt-to-output ratio. This makes sense because rollover defaults are generally more costly than fundamental defaults (they can occur even when output is relatively high) and this makes the sovereign wary about borrowing too much. In contrast, the average default frequency does not change much with the addition of rollover crisis and, as a result, the impact on the average spread is fairly small (in the SG model it stays the same, while in the DG model it rises slightly). However, there is significant change in the nature of defaults since many of them are now being induced by rollover crises. This is particularly pronounced in the case of the SG model, where 70% are now rollover-induced defaults. Along with this change in the nature of the defaults comes a change in the relationship between the spread and our fundamental shocks. In both models the correlation of the spread and the output shocks falls. This is particularly pronounced in the DG model, where it falls from 0.46 to 0.11.[o] In a similar fashion, the R^2 of the regressions of our spread on our wealth shock w falls in both models. In the SG model it falls by one-third, while in the DG model it falls by one-half. At the same time, the standard deviation of the spread hardly changes with belief shocks.

The lack of increase in the spread's volatility is surprising. To understand a bit better what is going on, we plot default indifference curves for both the benchmark SG model and the SG model with rollover crises in response to belief shocks in Fig. 10. We start first with the benchmark model. The indifference condition between defaulting and not defaulting traces out combinations of the debt-to-output ratio and the current growth rate. Since growth is positively autocorrelated in this model, high growth today is good news about future output and hence reduces the incentive to default. Of course, a high debt burden encourages default. This gives us the trade-off we see in the first panel of the figure. We have also plotted the stationary debt levels (ie, the debt level where $b = a(s,b)$) as a function of the current growth rate of output. These debt levels are important because the government finds it optimal to lever/delever back to this point in response to a shock. The fact that the stationary points are positively sloped reflects the tendency to borrow into a boom discussed earlier. Defaults occur in equilibrium largely because a sufficiently low growth rate shock from a debt position close to the stationary points last period generated a current debt-to-output level that is on the wrong side of the indifference curve. In which case, the government optimally chooses to default. The fact that the gap between the indifference curve and the stationary point is increasing in g illustrates why default is closely associated with low output shocks.

In the second panel of Fig. 10, we see a similar graph for the SG model with rollover crises. Only now there are two indifference curves: one for fundamental defaults as in the

[o] Another feature of rollover defaults is that they can occur for fundamentals that are, on average, better than in the case of fundamental defaults. Thus, the correlation between defaults and fundamentals is also weakened, consistent with evidence reported in Tomz and Wright (2007). See also Yeyati and Panizza (2011) for an empirical evaluation of the timing of output losses surrounding default episodes.

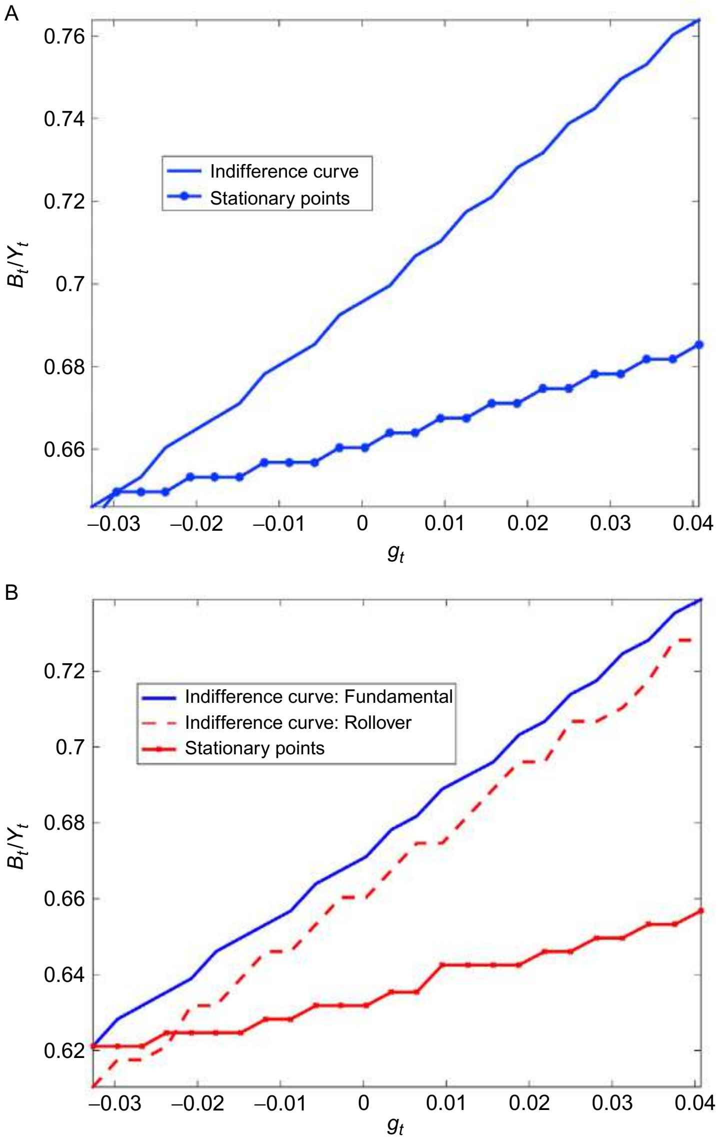


Fig. 10 Default indifference curves and stationary policy choices. (A) Benchmark. (B) w. Rollover crises.

benchmark model and one for rollover crises defaults. Since the lending terms are worse, the rollover indifference curve lies below the fundamental curve, indicating that a rollover crisis is possible for a given growth rate g_t at a strictly lower level of b_t. Note that the fundamental indifference curve is lower than in the benchmark model. This is because the future prospect of rollover crises lowers the payoff even when these crises do not occur today and this has shifted down the solvency indifference curve. As a result, defaults will occur at lower debt levels fixing g than in the benchmark model. Next, note that the stationary debt level curve has also been shifted down. This is because the increased likelihood of a default and its adverse consequences means that the optimal level of borrowing has decreased. The fact that 70% of the defaults occur under the crisis pricing schedule means that the likelihood of drawing a sufficiently bad output shock to force the government over the fundamental indifference curve has gone down substantially. In this sense the gap between the solvency indifference curve and the stationary debt levels has widened.

There is a sense in which virtually all of the defaults in the model with rollover crises are driven by beliefs. This is because, if we asked whether the states in which realized defaults are in equilibrium, very few of them are on the wrong side of the benchmark indifference schedule. It is also worth noting that if we suddenly switch from a situation in which the benchmark pricing schedule, policy function and beliefs applied, to one in which the rollover ones did, then the government would have to sharply delever in the face of a worse price schedule, even if a crisis did not formally occur in the current period. This sort of transition might be a way to generate more volatility in the spread, especially if the government could be induced to slow down the rate at which it delevered.

7. EXTENSIONS AND LITERATURE REVIEW

Beginning with Aguiar and Gopinath (2006) and Arellano (2008), there is now a substantial body of work drawing on the Eaton–Gersovitz framework. Aguiar and Amador (2014a) discuss the theoretical and conceptual issues in this area. This section provides a brief guide to the evolving quantitative literature (the reader is encouraged to consult the studies mentioned here for additional related work).

Existence and Uniqueness of Equilibrium: The existence of an equilibrium when both endowments and assets are continuous is an open question.[p] Aguiar and Amador (2014a) discuss that the operator whose fixed point characterizes the equilibrium (with permanent autarky as punishment) is monotone and note how this can be useful to compute an equilibrium. When both b and the non i.i.d. component of endowments are discrete, Chatterjee and Eyigungor (2012) establish the existence of an equilibrium for debt with arbitrary maturity and temporary or permanent autarky following default.

[p] Eaton and Gersovitz (1981) pointed out that *if* the probability of default $\mathbb{E}_s(D(s', b')$ is differentiable in b', the solution to the bond pricing equation amounts to the solution of a first-order nonlinear differential equation. However, differentiability of $\mathbb{E}_s D(s', b')$ requires everywhere differentiability of the value function, which is not true in a model with default.

The issue of uniqueness of equilibrium is more subtle. For the case where default is punished with permanent autarky, Auclert and Rognlie (2014) prove uniqueness for the Eaton–Gersovitz model with one-period debt. Passadore and Xandri (2015) study the multiplicity that arises when the state space for debt is restricted to be nonnegative (that is, no saving). Stangebye (2015a) and Aguiar and Amador (2016) discuss how multiplicity in the Eaton–Gersovitz model arises in the absence of one-period debt due to the vulnerability to dilution. More generally, one can often construct multiple equilibria with variations on the standard set up. Cole and Kehoe (2000) alter the Stackelberg nature of the government's default decision in order to generate self-fulfilling rollover crises. Chatterjee and Eyigungor (2012) exploits a similar variation to generate (investor) belief-driven rollover crises in a model that otherwise resembles the Eaton–Gersovitz setup.

The Strategic Structure of the Debt Market: In the Eaton–Gersovitz setup, the sovereign accesses the debt market at most once within a period. If the sovereign may access the market as many times within a period as it wishes, lenders at any given round of borrowing must anticipate the sovereign's future within-period borrowing decisions (Bizer and DeMarzo, 1992). As shown in Hatchondo and Martinez (undated) equilibrium implications of this is that investors will offer the sovereign a state-dependent pair of bond price and debt limit, $\{\bar{q}(y,b),\bar{x}(y,b)\}$, with the sovereign free to borrow any $b' \leq \bar{x}(y,b)$ at the price $\bar{q}(y,b)$. Interestingly, the bond *price* depends on inherited debt b (while in the standard setup the bond price *schedule* $q(y,b')$ is independent of b) and, so, borrowing history matters for the terms of credit. Lorenzoni and Werning (2014) and Ayres et al. (2015) discuss this issue in detail.

Contract Choice: In the standard setup, the structure of a unit bond is fixed and described by the pair (z,λ). At the cost of enlarging the state space, more flexible contractual structures are possible. Bai et al. (2014) define a unit bond by (T,δ), where the bond pays $(1+\delta)^{-\tau}$, $0 \leq \tau \leq T$ periods from maturity. Sanchez et al. (2015) consider the case where $\delta = 0$. Both relax the fixity of the contractual structure by letting the sovereign replace old debt each period with new debt with a different contractual structure.

Maturity Choice: Cole and Kehoe (1996) discuss the role of maturity in the presence of self-fulfilling debt crises. In the standard setup, market incompleteness is extreme in that only one type of debt contract can be issued at any time. Arellano and Ramanarayanan (2012) consider the case where the sovereign can simultaneously buy and sell bonds of different maturities and show that the average maturity shortens as fundamentals weaken. Aguiar and Amador (2014b) show that when default probabilities are high, the sovereign has an incentive to reduce its stock of one-period debt. Shorter maturity provides the sovereign the correct incentives to minimize the inefficiencies represented by default. Bocola and Dovis (2015) discuss the role of maturity choice in the presence of both fundamental and rollover crises and analyze their separate roles in the recent Eurozone debt crisis.

Exchange Rates, Default Risk, and Currency Denomination: Sovereign defaults are generally preceded by a depreciation of the country's currency, with a further sharp depreciation occurring soon after default. Asonuma (2014) documents these facts and develops a two-country model with traded and nontraded goods in which one country is the borrower and the other the lender. Negative shocks to productivity in the borrowing country can trigger a real exchange rate depreciation which, in turn, can raise the likelihood of a default on sovereign debt. Gumus (2013) examines the currency denomination of debt in a similar model with two types of debt: In one, the payoff is linked to the domestic price index (a proxy for local currency debt) and in the other to the price of the tradeable good (a proxy for foreign currency debt). Although the default risk on "local currency debt" is not uniformly lower than the default risk on "foreign currency debt," the former is found to be the better (higher welfare) arrangement.

Explicit Treatment of the Government: For some purposes, it is important to model the sovereign separately from private-sector agents. Cuadra and Sapriza (2008) analyze borrowing and default behavior when redistributive conflict and the risk of political turnover impart myopia (present-bias) *a lá* Alesina and Tabellini (1990). In their model, the sovereign discounts the future more than citizens do, which helps to partially rationalize the low discount factors often used in quantitative models. Hatchondo et al. (2009) consider two types of governments that differ in their discount factors with the goal of analyzing how political risk affects default probabilities and the volatility of spreads. Cuadra et al. (2010) model the government sector in order to give an account of the strongly procyclical nature of fiscal policy in emerging economies.[q]

Settlement Following Default: Sovereign defaults end with a settlement on the defaulted debt, wherein creditors accept a haircut and the sovereign regains (unencumbered) access to credit markets. Generally, settlement occurs after a significant amount of time has elapsed since default. In the context of one-period debt and equal treatment of all creditors in default (the so-called *pari passu* clause), Yue (2010) models settlement as the outcome of a one-shot Nash bargain between the sovereign and the representative creditor in the period of default. Following agreement, the sovereign is assumed to repay the renegotiated debt over time, with no possibility of default or access to new borrowing. This produces a theory of haircuts but not of delays. Bi (2008) assumes that defaulted debts must be settled in cash but employs the stochastic alternating-offers game developed in Merlo and Wilson (1995) to produce a theory of both haircuts and delays. Benjamin and Wright (2009) observe that settlement is typically done with new debt (rather than just cash) and allow for this possibility within the context of the stochastic

[q] Amador (2012) shows that once the equilibrium of the political game between different groups comprising the government is taken into account, it becomes possible to sustain positive levels of debt even when punishment for default is limited to exclusion from future credit (*contra*; Bulow and Rogoff, 1989)

alternating-offers game. In both models, delays arise because it is optimal for both parties to defer settlement until the sovereign's endowment is sufficiently high.[r]

Restructuring Without Default: Default and debt restructuring is a form of ex-post state contingency. Logically, and in practice, ex-post state contingency need not involve default. Hatchondo et al. (2014) point to voluntary debt exchanges as debt write-offs that occur when a sequence of bad endowment shocks places the sovereign on the wrong side of the revenue Laffer curve. Relatedly, Asonuma and Trebesch (2015) document that about a third of all restructurings in the last several decades occurred in the absence of default, termed *preemptive* restructuring. They extend the Eaton–Gersovitz model to allow for such restructurings and show that they occur when the likelihood of a future default is high. Salomao (2014) has analyzed how the presence of a credit default swap (CDS) market impacts debt renegotiation, when the outcome of the negotiation determines whether a "credit event" is triggered.

Partial Default: Default is typically modeled as a binary event on a single type of debt. In reality, sovereigns have a range of external obligations outstanding at any point in time, including trade credit, bank loans, bonds, bilateral (government-to-government) loans, loans from multilateral agencies (IMF, World Bank and other agencies) and they may choose to default on some types of loans but not on others. Thus, in the aggregate, default tends to be partial. Based on this observation, Arellano et al. (2013) develop a one-period debt model in which the sovereign can partially default on existing debt. Unpaid debts accumulate arrears and there is an output loss that is increasing in the ratio of unpaid to total debts. In their model, moderately bad output shocks trigger partial default that gets "cured" as output recovers.

Reputation: Quantitative sovereign debt models generally do not give any role to reputation in sustaining debt, although the idea that reputation matters is invoked in Eaton and Gersovitz, and, more comprehensively, in Tomz (2007). D'Erasmo (2012) extends the Eaton–Gerovitz model to the case where investors are uncertain about the sovereign's discount factor (degree of patience), which is taken to be stochastic. Investors' perception of the likelihood that the sovereign is the patient type now appears as an additional state variable in the sovereign's dynamic program. The patient type's desire to separate itself from the impatient type encourages more disciplined borrowing behavior on its part. In equilibrium, the patient type can sustain a higher level of debt on average. Generally speaking, the impatient type defaults and the patient type reaches settlement on the defaulted debt.

[r] Bai and Zhang (2012) explore the role of asymmetric information in creating delays in reaching settlement in a stylized environment. The uninformed party (the sovereign) screens creditors (who privately know their reservation value) by making successively attractive offers over time. They show that delay is shorter when the defaulted debt is traded in the secondary market because the price partially reveals the creditors' reservation value.

Sudden Stops: There is a large literature on "sudden stops" that focuses on the macroeconomic implications of a halt of capital inflows into emerging markets. This literature does not base the "sudden stop" on a rollover problem and abstracts from the possibility of sovereign default induced by the sudden stop (see, for instance, Mendoza (2010) and the references cited therein). Bianchi et al. (2014) make the connection to sovereign default by extending the Eaton–Gersovitz model to allow for an exogenous stop in capital inflows and study the role of international reserves (which cannot be grabbed by foreign investors in the event of default) as a hedge against such stops.[s]

Fiscal Rules and Default: There is a literature aimed at understanding the equilibrium implications of fiscal policy rules. Ghosh et al. (2011) analyze a model where the government adheres to some given fiscal rule as long as the deficit implied by the rule can be financed at a finite interest rate. In terms of our notation, this is a setup in which there is some function $c(y,b)$ (the fiscal policy rule) and b' is chosen each period to satisfy $q(y,b') \cdot [b' - (1 - \lambda)b] = y - (r^* + \lambda)b - c(y,b)$. Because the revenue curve $q(y,b')b'$ is an inverted U, there may be no b' that satisfies this equation in which case the sovereign defaults. Furthermore, if there is one b' that satisfies the budget constraint, there will always exist another b' on the "wrong side" of the revenue Laffer curve that will also satisfy this equation. Ghosh et al. assume that the sovereign and investors avoid the wrong side of the Laffer curve and compute the highest debt level $\bar{b}$ beyond which default is certain. Lorenzoni and Werning (2014) and Stangebye (2015b) study a similar setup but the focus is on the rise in interest rates if investors temporarily coordinate on the low price (and therefore high debt) equilibrium path. These authors focus on the recent Eurozone experience.

Debt Dilution and Alternative Trading Arrangements: In quantitative models with long-term debt, "debt-dilution" is an important force leading to excessive borrowing and default. This leads to consideration of alternative trading arrangements that mitigate the adverse effects of debt dilution. Chatterjee and Eyigungor (2015) analyze how respecting seniority during (postdefault) debt renegotiations can improve incentives and the welfare of the sovereign. Hatchondo et al. (2015) analyze how adherence to a fiscal policy rule that binds future sovereigns' borrowing decisions can improve the welfare of the current sovereign.

Decentralized Borrowing and Centralized Default: A growing portion of a country's external debt is debts incurred by private borrowers. Kim and Zhang (2012) analyze an Eaton–Gersovitz model in which private agents choose how much to borrow but the sovereign chooses whether to default. Because private borrowers act as price-takers,

[s] The accumulation of foreign reserves to mitigate rollover risk has been examined from an optimal contracting perspective in Hur and Kondo (2014). They point to the drop-off in the frequency of sudden stops following reserve accumulation by emerging markets as evidence that reserves affect the likelihood of a rollover crisis.

the equilibrium resembles one in which the sovereign can access the credit market unboundedly many times within a period.

Contagion and Correlated Defaults: Lizarazo (2009) studies how the terms of credit offered to sovereigns are affected if sovereigns share a common risk-averse lender. Correlated defaults may occur because a default by one sovereign lowers the wealth of the lender and reduces the supply of credit to all sovereigns. The reduction in supply could push another sovereign into default. Arellano and Bai (2014) study a similar environment but include renegotiation on the defaulted debt and show that bargaining protocols (independent vs coordinated bargaining with sovereigns following default) differentially affect the likelihood of correlated defaults.

Inflation and Default: The bulk of the quantitative-theoretic literature on debt and default models real economies. Two exceptions are Nuno and Thomas (2015) and Du and Schreger (2015). The former compares (in a continuous-time setting) outcomes where sovereign debt is denominated in real terms (with the possibility of outright default) to one where it is nominal and the sovereign chooses monetary and fiscal policy under discretion. The latter studies default risk on sovereign debt denominated in local currency, when private borrowers issue debt denominated in foreign currency. The existence of foreign currency private debt makes inflating away local currency sovereign debt expensive and, thus, keeps default risk on local currency sovereign debt positive (as observed in the data).

News Shocks: Sovereign defaults do not occur only when fundamentals are weak. One possible explanation of this fact could be that they occur when the sovereign and investors receive bad news about the future. Durdu et al. (2013) extend the standard Eaton–Gersovitz set up to include news shocks about future TFP. In addition to default triggered by bad news, the precision of news about future TFP is shown to have quantitatively significant effects the bond pricing schedule.

Default Costs: Quantitative-theoretic models of debt and default typically take the structure of the output costs of default as given. Two exceptions to this practice are Mendoza and Yue (2012) and Perez (2015). In the former, the default costs are grounded in producers' inability to import foreign intermediate inputs when the country is in default. The key implication of this setup is asymmetric default costs: the output costs of default are proportionally higher when TFP is high because that is when the loss of foreign intermediate inputs is proportionately more costly. In the latter, the output costs of default are grounded on the loss of net worth of financial intermediaries (who hold sovereign debt) that occurs with default and the consequent fall in the level and efficiency of financial intermediation, which then depresses output.

Investment and Default: The quantitative debt and default literature has uniformly examined endowment economies. An exception is Gordon and Guerron-Quintana (2016) who extend—both substantively and computationally—the long-term debt model of Chatterjee and Eyigungor (2012) to include capital accumulation (with costly adjustment) and labor-leisure choice. Their goals are a more complete understanding of emerging market business cycles and of the impact of phyiscal capital on debt sustainability.

8. CONCLUSION: WHERE WE'VE BEEN AND WHERE WE NEED TO GO?

This chapter has documented a number of important facts about sovereign default crises, including:

1. Average spreads, spread volatility and the frequency of spread crises vary quite a bit across developing countries.
2. Fundamentals explain only a limited share of spread movements.
3. Spreads have some common factors driving them. However, these factors do not seem tightly connected to standard measures of risk pricing, uncertainty or the risk-free rate.

We have also examined alternative versions of the standard model of sovereign borrowing and defaults. Some of these versions explain many of the main facts, such as the average spread, the default frequency and average debt-to-GDP ratios. However, all of these models struggle to simultaneously explain the volatility of spreads and its apparent lack of connection to country fundamentals. Specifically:

1. In our model countries engage in very limited borrowing and saving to smooth consumption. While this leveraging and deleveraging behavior is found in the data, it seems much less pronounced. As a result, the variation in the debt-to-output ratio is smaller in the model than in the data. This leads to much less variation in the models' implied spread.
2. Nonlinear default costs can increase the volatility of the spread in the DG model when the volatility of output is high. But this increase in the spread comes at the expense of tightly tying movements in the spread to country fundamentals.
3. The SG model is much less sensitive to including nonlinear default costs in part because the low current output realizations do not stimulate much borrowing as growth rates are modestly positively persistent and because the volatility in growth rates is small relative to volatility in the deviation from trend.

Both increases in the risk aversion of our lenders and negative shocks to their wealth did not lead to sharp increases in the spread as simple intuition might suggest. Instead the disciplinary effect of the increase in the price of default risk reduces the future incentive of the government to issue debt into the range that will generate a positive probability of default. This increase in future discipline lower creditors' anticipation of future dilution of their claims by the government and can actually reduces spreads. This negative relationship between the pricing of default risk and the equilibrium spread also appears to be an important factor in the data, thus, validating this surprising implication of our models.

The failure of our models to explain the volatility of spreads stems from the fact that the debt-to-output ratio is largely pinned down by a couple of key features. First, because the government is quite myopic, smoothing plays a limited role in it's optimal policy choice; instead, borrowing is driven by impatience that is ultimately held in check by lack of commitment. Second, because of the strong feedback effect of default risk and risk premia on the government's incentive to default, the debt price schedule is highly nonlinear in the relevant region. As a result, the location of the kink in the price schedule interacts with the sovereign's myopia to almost completely determine its borrowing behavior. In the end, this leads to sharp leveraging/deleveraging in response to positive and negative output shocks and very little variation in the spread. These forces are somewhat ameliorated in cases where the output shock is sufficiently volatile (so the nonlinearity in the default cost can play a role), but even in those cases the sovereign's behavior responds sharply to the contemporaneous shock realization and does not display the history dependance that expenditure-smoothing would have implied. As a result, only the current output shock matters for spreads and this ends up overloading its importance relative to the data.

Rollover crises are a promising way of generating debt crises, particularly since they don't imply an overly tight connection to country fundamentals. However, the sort of stationary rollover risk that we have considered here is not sufficient to produce the kind of variability in the spread that we see in the data. Instead, they seem to simply crowd out standard fundamental crises. What is needed is a more dynamic version of time-varying risks. At the same time, we need to rationalize a reduction in the speed with which the government chooses to undo the impact of negative shocks on the spread by borrowing less and yet not default on the debt.

ACKNOWLEDGMENTS

We thank the editors Harald Uhlig and John Taylor, our discussant Manuel Amador, and participants at the March 2015 Handbook conference in Chicago for thoughtful comments. We also thank St. Martin's "Conference on the Sand." The views expressed here are those of the authors and do not necessarily represent the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

REFERENCES

Aguiar, M., Amador, M., 2014a. Sovereign debt. In: Gopinath, G., Helpman, E., Rogoff, K. (Eds.), Handbook of International Economics, vol. 4. North Holland, pp. 647–687.

Aguiar, M., Amador, M., 2014b. Take the short route: how to repay and restructure sovereign debt with multiple maturities. Mimeo.

Aguiar, M., Amador, M., 2016. Maturity and multiplicity in sovereign debt models. Working Paper.

Aguiar, M., Chatterjee, S., Cole, H.L., Stangebye, Z.R., 2016. Self-fulfilling Debt Crisis, Revisited: The Art of the Desperate Deal. Mimeo.

Aguiar, M., Gopinath, G., 2006. Defaultable debt, interest rate and the current account. J. Int. Econ. 69, 64–83.

Aguiar, M., Gopinath, G., 2007. Emerging market business cycles: the cycle is the trend. J. Polit. Econ. 115, 69–102.

Alesina, A., Tabellini, G., 1990. A positive theory of fiscal deficits and government debt in a democracy. Rev. Econ. Stud. 57, 403–414.

Amador, M., 2012. Sovereign debt and the tragedy of the commons. Mimeo.

Arellano, C., 2008. Default risk and income fluctuations in emerging markets. Am. Econ. Rev. 98 (3), 690–712.

Arellano, C., Bai, Y., 2014. Linkages across sovereign debt markets. Research Department Staff Report 491, Federal Reserve Bank of Minneapolis.

Arellano, C., Mateos-Planos, X., Rios-Rull, V., 2013. Partial default. Federal Reserve Bank of Minneapolis, Mimeo, Working Paper.

Arellano, C., Ramanarayanan, A., 2012. Default and maturity structure in sovereign bonds. J. Polit. Econ. 120, 187–232.

Asonuma, T., 2014. Sovereign defaults, external debt and real exchange rate dynamics. International Monetary Fund, Mimeo.

Asonuma, T., Trebesch, C., 2015. Sovereign debt restructurings: preemptive or post-default. Discussion Paper 10950, Center for Economic Policy Research.

Auclert, A., Rognlie, M., 2014. Unique equilibrium in the Eaton-Gersovitz model of sovereign debt. Mimeo.

Ayres, J., Navarro, G., Nicolini, J.P., Teles, P., 2015. Sovereign default: the role of expectations. Federal Reserve Bank of Minneapolis Working Paper 723.

Bai, Y., Zhang, J., 2012. Duration of sovereign debt renegotiation. J. Int. Econ. 86 (2), 252–268.

Bai, Y., Kim, S.T., Mihalache, G., 2014. The maturity and payment schedule of sovereign debt. Mimeo.

Benjamin, D., Wright, M.L., 2009. Recovery before redemption: a theory of sovereign debt renegotiation. Mimeo.

Bi, R., 2008. "Beneficial delays" delays in restructuring negotiations. Working Paper WP/08/38, International Monetary Fund.

Bianchi, J., Hatchondo, J.C., Martinez, L., 2014. International reserves and rollover risk. Mimeo.

Bizer, D.S., DeMarzo, P.M., 1992. Sequential banking. J. Polit. Econ. 100, 41–61.

Bocola, L., Dovis, A., 2015. Self-fulfilling debt crisis: a quantitative analysis. Mimeo.

Borri, N., Verdelhan, A., 2011. Sovereign risk premia. Working Paper.

Broner, F., Lorenzoni, G., Schmukler, S.L., 2013. Why do emerging economies borrow short term? J. Eur. Econ. Assoc. 11 (S1), 67–100.

Bulow, J., Rogoff, K.S., 1989. Sovereign debt: is to forgive to forget? Am. Econ. Rev. 79 (1), 43–50.

Calvo, G.A., 1988. Servicing the public debt: the role of expectations. Am. Econ. Rev. 78 (4), 647–661.

Chatterjee, S., Eyigungor, B., 2012. Maturity, indebtedness and default risk. Am. Econ. Rev. 102 (6), 2674–2699.

Chatterjee, S., Eyigungor, B., 2015. A seniority arrangement for sovereign debt. Am. Econ. Rev. 105 (12), 3740–3765.

Cole, H.L., Kehoe, T., 1996. A self-fulfilling model of Mexico's 1994-1995 debt crisis. J. Int. Econ. 41, 309–330.

Cole, H.L., Kehoe, T., 2000. Self-fulfilling debt crisis. Rev. Econ. Stud. 67 (1), 91–116.

Cuadra, G., Sanchez, J.M., Sapriza, H., 2010. Fiscal policy and default risk in emerging markets. Rev. Econ. Dyn. 13, 452–469.

Cuadra, G., Sapriza, H., 2008. Sovereign defaults, interest rates and political uncertainty in emerging markets. J. Int. Econ. 76, 77–88.

D'Erasmo, P., 2012. Government reputation and debt repayment in emerging economies. Mimeo.

Du, W., Schreger, J., 2015. Sovereign risk, currency risk and corporate balance sheets. Mimeo.

Durdu, B., Nunes, R., Sapriza, H., 2013. News and default risk in small open economies. J. Int. Econ. 91 (1), 1–17.

Eaton, J., Gersovitz, M., 1981. Debt with potential repudiation: theoretical and empirical analysis. Rev. Econ. Stud. 48 (2), 289–309.

Ghosh, A.R., Kim, J.I., Mendoza, E.G., Ostry, J.D., Qureshi, M.S., 2011. Fiscal fatigue, fiscal space and debt sustainability in advanced economies. Working Paper 16782, National Bureau of Economic Research.

Gordon, G., Guerron-Quintana, P.A., 2016. Dynamics of investment, debt, and default. Mimeo.

Gromping, U., 2007. Estimators of relative importance in linear regression based on variance decomposition. Am. Stat. 61 (2), 139–147.

Gumus, I., 2013. Debt denomination and default risk in emerging markets. Macroecon. Dyn. 17, 1070–1095.

Hatchondo, J.C., Martinez, L., 2009. Long duration bonds and sovereign defaults. J. Int. Econ. 79 (1), 117–125.

Hatchondo, J.C., Martinez, L., undated. Credit risk without commitment. Mimeo.

Hatchondo, J.C., Martinez, L., Sapriza, H., 2009. Heterogeneous borrowers in quantitative models of sovereign default. Int. Econ. Rev. 50 (4), 1129–1151.

Hatchondo, J.C., Martinez, L., Sosa-Padilla, C., 2014. Voluntary debt exchanges. J. Monet. Econ. 61, 32–50.

Hatchondo, J.C., Martinez, L., Roch, F., 2015. Fiscal rules and the sovereign default premium. Mimeo.

Hur, S., Kondo, I.O., 2014. A theory of rollover risk, sudden stops, and foreign reserves. Mimeo.

Kim, Y.J., Zhang, J., 2012. Decentralized borrowing and centralized default. J. Int. Econ. 88, 121–133.

Leland, H., 1994. Bond prices, yield spreads, and optimal capital structure with default risk. IBER Finance Working Paper 240.

Lindeman, R., Merenda, P.F., Gold, R., 1980. Introduction to Bivariate and Multivariate Analysis. Scott Foresman, Glenview, IL.

Lizarazo, S.V., 2009. Contagion of financial crisis in sovereign debt markets. Munich Personal RePec Archive, Discussion paper.

Longstaff, F., Pan, J., Pedersen, L., Singleton, K., 2011. How sovereign is sovereign credit risk. Am. Econ. J.: Macroecon. 3, 75–103.

Lorenzoni, G., Werning, I., 2014. Slow moving debt crises. Mimeo.

Mendoza, E.G., 2010. Sudden stops, financial crises, and leverage. Am. Econ. Rev. 100 (5), 1941–1966.

Mendoza, E.G., Yue, V.Z., 2012. A general equilibrium model of sovereign default and business cycles. Q. J. Econ. 127 (2), 889–946.

Merlo, A., Wilson, C., 1995. A stochastic model of sequential bargaining with complete information. Econometrica 63 (2), 371–399.

Neumeyer, P.A., Perri, F., 2005. Business cycles in emerging economies: the role of interest rates. J. Monet. Econ. 52 (2), 345–380.

Nuno, G., Thomas, C., 2015. Monetary policy and sovereign debt vulnerability. Mimeo.

Passadore, J., Xandri, J.P., 2015. Robust conditional prediction in dynamic games: an application to sovereign debt. Mimeo.

Perez, D.J., 2015. Sovereign debt, domestic banks and the provision of public liquidity. Mimeo.

Reinhart, C.M., Rogoff, K.S., Savastano, M.A., 2003. Debt intolerance. Brook. Pap. Econ. Act. 34, 2003-1, 1–74.

Salomao, J., 2014. Sovereign debt renegotiations and credit default swaps. University of Minnesota, Mimeo.

Sanchez, J., Sapriza, H., Yurdagul, E., 2015. Sovereign default and choice of maturity. FRB St. Louis Working Paper 2014-031B.

Stangebye, Z.R., 2015a. Dynamic panics: theory and application to the eurozone. Working Paper.

Stangebye, Z.R., 2015b. Lifetime-laffer curves and the eurozone. University of Notre Dame, Mimeo.

Tomz, M., 2007. Reputation and International Cooperation. Princeton University Press, Princeton, NJ.

Tomz, M., Wright, M.L.J., 2007. Do countries default in “bad times”? J. Eur. Econ. Assoc. 5 (2-3), 352–360.

Yeyati, E.L., Panizza, U., 2011. The elusive costs of sovereign defaults. J. Dev. Econ. 94, 95–105.

Yue, V.Z., 2010. Sovereign default and debt renegotiation. J. Int. Econ. 80 (2), 176–187.

SECTION 4

Models of Economic Growth and Fluctuations

CHAPTER 22

RBC Methodology and the Development of Aggregate Economic Theory

E.C. Prescott
Arizona State University, Tempe, AZ, United States
Federal Reserve Bank of Minneapolis, Minneapolis, MN, United States

Contents

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.001

Abstract

This essay reviews the development of neoclassical growth theory, a unified theory of aggregate economic phenomena that was first used to study business cycles and aggregate labor supply. Subsequently, the theory has been used to understand asset pricing, growth miracles and disasters, monetary economics, capital accounts, aggregate public finance, economic development, and foreign direct investment.

The focus of this essay is on real business cycle (RBC) methodology. Those who employ the discipline behind the methodology to address various quantitative questions come up with essentially the same answer—evidence that the theory has a life of its own, directing researchers to essentially the same conclusions when they apply its discipline. Deviations from the theory sometimes arise and remain open for a considerable period before they are resolved by better measurement and extensions of the theory. Elements of the discipline include selecting a model economy or sometimes a set of model economies. The model used to address a specific question or issue must have a consistent set of national accounts with all the accounting identities holding. In addition, the model assumptions must be consistent across applications and be consistent with micro as well as aggregate observations. Reality is complex, and any model economy used is necessarily an abstraction and therefore false. This does not mean, however, that model economies are not useful in drawing scientific inference.

The vast number of contributions made by many researchers who have used this methodology precludes reviewing them all in this essay. Instead, the contributions reviewed here are ones that illustrate methodological points or extend the applicability of neoclassical growth theory. Of particular interest will be important developments subsequent to the Cooley and Hansen (1995) volume, *Frontiers of Business Cycle Research*. The interaction between theory and measurement is emphasized because this is the way in which hard quantitative sciences progress.

Keywords

Neoclassical growth theory, Aggregate economic theory, RBC methodology, Aggregation, Business cycle fluctuations, Development, Aggregate financial economics, Prosperities, Depressions

JEL Classification Codes

B4, C10, E00, E13, E32, E60

1. INTRODUCTION

This chapter reviews the development and use of a quantitative, unified theory of aggregate variables both across time and across economies at a point in time. This theory accounts not only for traditional business cycle fluctuations but also for prosperities and depressions, as well as for the vast difference in living standards across countries. This unified quantitative dynamic general equilibrium theory accounts for the large movements in asset values relative to gross national income (GNI), the consequences of alternative monetary policies and tax systems, and the behavior of current accounts as well.

No competing quantitative theory has been developed for the study of aggregate economic behavior. This disciplined theory is unified and has been tested through successful use. The assumptions made when constructing a model economy, or in some cases a set of economies, to address a given question must be consistent with assumptions made in the previous successful applications. Deviations from this theory have arisen, which is evidence that some real theory is involved.[a] Other deviations remain to be discovered. Some of the recognized deviations or puzzles have been resolved via further development of the theory, others by better measurement. This interaction between theory and measurement is the way in which a hard quantitative science progresses.

We call this theory neoclassical growth theory. Key features of this theory are the allocation of productive time between market and household activities and the allocation of output between consumption and investment. Depending on the application, other features of reality must be included, such as sector detail, the nature of the financial system as specified by laws and regulations, and the contracting technology available. Heterogeneity of people in the model economy, with respect to age and idiosyncratic shocks, must be and has been included in models used to address issues such as the consequences of an aging population for various tax policy regimes.

The underlying theoretical framework is the theory of value, in particular the capital theory variant. This means the models used to draw scientific inference will have a recursive structure. This is a crucial feature for the model economies being used to draw scientific inference because the national account statistics can be constructed and compared with actual statistics.

To summarize, aggregate economics is now a hard quantitative science. It has been tested through successful use in all substantive fields of economics.

2. A BRIEF HISTORY OF BUSINESS CYCLES

Fluctuations in the level of business activity have long been a topic of concern. Mitchell (1913, 1927) collected many indicators of the level of economic activity. He viewed the level of economic activity as being cyclical with alternating periods of contractions and expansions. He developed the National Bureau of Economic Research (NBER) definition of recession, which is a period of contraction in the level of economic activity. This definition is still used by the NBER. He categorized his set of indicators into leading indicators, lagging indicators, and contemporaneous indicators. This was the framework he used for forecasting, and it did improve forecasting.

Mitchell called these fluctuations "business cycles." Wicksell (1918) used a rocking horse analogy to think about business cycles. Rocking horses display damped oscillations absent new shocks. This development led the profession to search for an economic structure with these properties. Frisch (1933) viewed business cycle research as the search for shocks or impulses to the economy and a damped oscillatory propagation mechanism.

[a] Trade theory is a disciplined theory. All using the discipline of trade theory come up with essentially the same findings. See Arkolakis et al. (2012).

Samuelson (1939) developed his multiplier–accelerator macroeconomic model that displayed these properties. His model had a consumption function and an investment equation. His model was also a second-order linear equation in real output with parameters that gave rise to damped oscillatory behavior.

The NBER definition of recessions is flawed along three dimensions. First, no corrections are made for trend growth or population size. With the NBER definition, the economy is in expansion 90% of the time and in recession or contraction 10% of the time. With trend-corrected real gross domestic product (GDP) per person 16 years and older, the economy is expanding approximately half of the time and contracting half of the time. Second, the NBER definition of recession is not revised subsequent to revisions in the economic time series. These revisions are sometimes large and are made years later as recent census data become available. If the revised data were used, the timing and magnitude of recessions and expansions would change. Third, the NBER definition of recession is not well defined and has a large subjective element.

The biggest problem in business cycle theory is that these so-called business cycles are not cyclical. This was established by Adelman and Adelman (1959), who found that the Klein–Goldberg model—the first econometric model to be used to forecast business cycles—displays damped nonoscillatory behavior. This finding, however, does not rule out the existence of longer cycles in the level of business activity. Kuznets's (1930) view was that there were 15- to 20-year cycles in output and prices in the United States. He labeled these fluctuations "secondary secular movements." Subsequently, they were called Kuznets cycles. Kondratieff and Stolper (1935) hypothesized even longer business fluctuations with 50- to 60-year cycles.

There are, of course, seasonal cycles, which are cycles in the true sense of the word. But they are of little interest and receive little attention in aggregate analysis. To handle them, the economic data used in aggregate analyses are seasonally adjusted.

2.1 The National Accounts: Defining Macroeconomics

A goal in the early 1930s was to come up with a measure of the performance of the business sector. Kuznets (1930) came up with one that proved to be useful. This measure is gross national product (GNP), the value of all final goods and services produced. Other researchers measured the value of the inputs to the business sector, which are the services of capital stocks. The most important category of these services is the services of different types of human capital. The aggregate value of human capital services is commonly called labor income. The services of tangible capital make up the other major category. The aggregate value of these services is called capital income. Claims against output are by definition income, and given that all businesses have a residual claimant, income equals product.

In the late 1930s, Tinbergen (1952) developed quantitative dynamic time series models and used them for forecasting. Given his background in physics, he thought in terms of empirically determined dynamic systems with instruments and targets.

On the other hand, Lawrence R. Klein, the father of macroeconometric modeling, had a theory underlying the dynamic aggregate models he developed and used for forecasting. The theory is the Hicksian IS-LM theory, later augmented with a Phillips curve. The beauty of Klein's work was that it featured a fully specified dynamic system, which had national accounts. All accounting identities held, which resulted in a consistent set of forecasts for all of the variables. Over time, these macroeconometric models grew in size as the sector detail became richer. Klein's model and other macroeconometric models in his framework came to dominate because their use dramatically improved forecasting. After World War II, for example, most economists thought the United States would experience another Great Depression. Using his model, Klein correctly forecasted that no depression would occur.

The nature of macroeconomics in the 1960s was coming up with a better equation to be included in the basic macroeconomic model. The generally held view was that the neoclassical foundations for the empirically determined aggregate dynamic system would subsequently be developed. The famous Phelps Conference at the University of Pennsylvania in 1969, entitled "Micro Foundations of Wage and Price Determination," tried to bring about the synthesis of macroeconometric models into neoclassical economics.

This neoclassical synthesis, however, was not to be. Lucas (1976a), in his paper entitled "Econometric Policy Evaluation: A Critique," found that the existence of a policy-invariant dynamic system is inconsistent with dynamic economic theory. The implication of this finding was that there was no hope for the neoclassical synthesis. The use of dynamic economic theory to evaluate policy requires that the dynamic system governing the evolution of the national accounts be an endogenous element and not a policy-invariant element, which can be empirically determined.

What happens at a point in time depends on what policy regime will be followed in the future. An implication of this fact is that economic theory cannot predict what will happen as a consequence of a possible current policy action choice. What will happen as the result of a policy action is not a well-posed question in the language of dynamic economic theory. What will happen if some policy rule or regime is followed in the future is a well-posed economic question—a point made by Lucas (1976a).

No one challenged Lucas's conclusions, and those who continued to support the use of macroeconometric models for evaluating policy took the position that a different theoretical framework was needed for the study of business cycle fluctuations. Indeed, many used the theory underlying macroeconometric models of the 1960s to confidently predict that the unemployment rate could be decreased by increasing the inflation rate. In 1969 the unemployment rate and inflation rate were both about 4%. The policy consensus based on the perceived trade-off between inflation and unemployment was that the unemployment rate should be reduced because the social gains from having a lower unemployment rate exceeded the cost of the higher inflation.

This consensus led to an attempt to exploit this trade-off in the 1970s. As Lucas and Sargent (1979) point out, this attempt failed—and failed spectacularly, as predicted by dynamic economic theory.[b] Given this failure of Keynesian macroeconomics, the question was what would replace it.

2.2 Neoclassical Growth Theory: The Theory Used in Aggregate Analysis

The development of aggregate measures of outputs and inputs to the business accounts led to the identification of a set of growth facts. Kaldor's (1957) stylized view of these facts for long-term economic growth in the United States and the United Kingdom are as follows. Roughly constant are capital and labor shares of national income, consumption and investment shares of output, the return on investment, and the capital–output ratio. Growing at the same rate over time are national income and the real wage.

Solow (1956) developed a simple, elegant model that accounted for these facts. The model has an aggregate production function with constant returns to scale, with labor and capital being paid their marginal product. All productivity change is labor augmenting. Investment is a constant share of output, and the time allocated to market production per worker is a constant. Thus, the household makes no decisions. Following Frisch (1970), I therefore refer to the model as being classical.

Around the same time, Swan (1956) developed his growth model that is also consistent with the Kaldor growth facts. The key difference between his model and Solow's model is that Swan did not require neutral technology change. Instead, he assumed a unit elasticity of substitution between the factors of production. In the Swan (1956) paper, he carries out some output accounting. The Swan model is the one that has been used for output accounting.

2.3 The Classical Growth Model and Business Cycle Fluctuations

Lucas (1976b) defined business cycles as being recurrent fluctuations of output and employment about trend and the key facts to be the nature of comovements of aggregate variables about trend. But without a definition of trend, this is not a fully specified definition of business cycle fluctuations. This led Hodrick and Prescott (1980) to develop an operational definition of trend, and they used it to represent time series as the sum of a trend component and a business cycle component. In constructing the trend, a penalty was imposed on the sum of squares of the second differences of the trend. In mathematical terms, a time series y_t is represented as the sum of a trend component g_t and a cyclical component c_t; that is,

$$y_t = g_t + c_t.$$

Given the values of the y_t, the g_t is selected to minimize

[b] Lucas (1972), in what was probably the first dynamic aggregate theory paper, developed a model that displayed an empirical Phillips curve. He predicted that if attempts were made to exploit, they would fail. This prediction was made prior to the attempts to lower the unemployment rate by increasing the inflation rate.

$$\sum_{t=1}^{T} c_t^2 + \lambda \sum_{t=-1}^{T} [(g_t - g_{t-1}) - (g_{t-1} - g_{t-2})]^2.$$

This simple operational procedure has a single smoothing parameter, $\lambda \geq 0$. This parameter is chosen to mimic the smooth curve researchers would draw through the data. The larger its value, the smoother is the trend component. For quarterly data, the first number that Hodrick and I chose and ended up using was 1600. There is no right or wrong number, and it cannot be estimated because it is part of an operational definition. What is desirable is that the same statistics are used across studies of business cycle fluctuations of this type. This uniformity permits comparisons across studies.

A feature of this procedure is that the same linear transformation of the logarithm of all the inputs and outputs to the business sector is made. Consequently, Swan's (1956) output accounting could be used for the operationally defined cyclical component of the time series.

In examining the nature of these fluctuations, researchers documented some business cycle facts for the deviations from trend for the US economy for the 1950.1 to 1979.2 period:

(i) Consumption, investment, market hours, and labor productivity all moved procyclically.

(ii) The standard deviation of fixed investment was 5.1%, and the standard deviation of consumption was only 1.3%.

(iii) Market hours and GDP per hour were roughly orthogonal, with hours having twice the variance.

(iv) The standard deviation of quarterly log output was 1.8%, and the first-order serial correlation was 0.74.

(v) Stocks of capital lagged output, with the lag increasing with the durability of the capital. Inventory stock was almost contemporaneous, producer durables stocks lagged a few quarters, and structures lagged a couple of years.

2.4 The Neoclassical Growth Model

Kydland and Prescott (1982) added an aggregate household to the classical growth model in order to endogenize two key allocation decisions. The first of these allocation decisions is the split of output between investment and consumption. The split varies cyclically. The second of these allocation decisions is how much productive time is allocated to the business sector and how much to the household sector. These allocations are endogenous elements of the neoclassical growth model and, with respect to the aggregate household, depend on both its willingness and its ability to substitute. Thus, this extension of the growth model made it neoclassical in the sense of Frisch (1970).

Kydland and I found that if there were persistent shocks to factors determining the balanced growth path *level* of the neoclassical growth model and if the aggregate household was sufficiently willing to intertemporally substitute market time, the neoclassical growth model displayed fluctuations of the business cycle variety. The aggregate utility function of the stand-in household had a high Frisch labor supply elasticity, much higher than the one labor economists estimated using a representative household construct.

If there are common homothetic convex preferences across households, the aggregated household's labor supply elasticity is the same as that of the individuals being aggregated. Empirically, however, these elasticities are not the same. Kydland and Prescott (1982) found that the aggregate labor supply elasticity must be in excess of 3 for the neoclassical growth model to predict business cycle fluctuations, whereas MaCurdy (1981), using panel data, estimated the labor supply elasticity of prime-age males working continuously to be only 0.15. The aggregate and disaggregate estimates must be consistent, and a reason for this difference is needed.

2.5 Why the Discrepancy Between Micro and Aggregate Elasticity Estimates?

Rogerson (1984) came up with the reason for the discrepancy between micro and aggregate estimates. He observed that the principal margin of adjustment in aggregate labor supply was in the number of people working in a given week and not in the hours worked per worker. Consequently, the micro estimate of the labor supply using a theoretical structure predicting just the opposite has to be dismissed as an estimate of the aggregate labor supply elasticity. The labor economist conclusion that tax rates had little consequence for aggregate labor supply was wrong. This is an important example of the failure of micro theory in drawing *aggregate* scientific inference. Aggregation matters. This was recognized by Marshall in his classic textbook first published in 1890 and by Wicksell around the same time. The aggregate production function, given that there is entry and exit of production units, is very different from the production functions of individual units.

Rogerson (1984) developed a formal theory of the aggregate utility function when there was labor indivisibility. This theory was developed in a static context. Hansen (1985) introduced it into the basic neoclassical growth model and found that the resulting model displayed business cycle fluctuations. This research resolved the puzzling discrepancy between micro and aggregate observations.

2.6 Why Is There Labor Indivisibility?

The puzzle of what could give rise to labor indivisibility was resolved by Hornstein and Prescott (1993), who found that if individuals' outputs of labor services is a function of the capital that each worker uses, the margin of adjustment is the number of people working and not the number of hours worked. The fraction working is the margin used up to the point at which all are working. This model endogenized labor indivisibility in a simple version of the optimal growth model. An important point is that it breaks the clean separation between preferences and technology in determining the aggregate elasticity of labor supply.

An alternative theory of labor indivisibility was subsequently developed by Prescott et al. (2009). The key feature of this theory is that the mapping of time allocated to the market to units of labor services supplied is not linear. The increasing mapping is initially convex. Reasons for this nonlinearity include the time needed to update information on which decisions are made and the time needed to get organized. Then the mapping becomes concave; one reason is that workers become tired and perform tasks less well or at a lower rate.

One implication of this theory is that workweeks of different lengths are different commodities. This was recognized by labor economist Rosen (1978). Hansen and Sargent (1988) have two workweek lengths in their business cycle paper: a standard workweek and an overtime workweek. The micro evidence in support of workweeks of different lengths being different commodities is strong. For example, two half-time workers on average are paid significantly less than one full-time worker with similar human capital. Additional evidence is that the normal workweek length differs across occupations. With this theory, the reason for the differences in workweek lengths across occupations is that the mapping from time allocated to the market to units of labor services produced is different across occupations. When important nonconvexities are present, the micro and aggregate elasticities are different even if all the micro units are identical.

This is true for both the household and the business sectors. At the production unit level, investment is very lumpy, yet at the aggregate level, aggregate investment is smooth. Thomas (2002) established that valuation equilibrium theory predicts that the fraction of units making discrete adjustments to production capacity will be the margin of adjustment used, as it is, and aggregate investment will be smooth.

Time series methods used to model aggregate time series use linear models. This is because there are no obvious nonlinearities in the time series. The one case in which nonlinearity was found to be significant was in the Hansen and Prescott (2005) model with a capacity utilization constraint. If capacity constraints are occasionally binding, aggregation theory leads to an aggregate production function that has a kink, which results in the labor income share falling when the capacity constraint is binding. It also implies that business cycle peaks will be flatter and smaller than troughs for the detrended data as they are. This is an improvement in theory but is of second-order importance.

2.7 A Digression on Methodology of Aggregate Analysis

Theory is a set of instructions for constructing a model economy to address a given question. The criterion for a good theory is that it is useful. Models are instruments used to draw scientific inference. What constitutes a good model depends on what question is being addressed. Reality is incredibly complex, and any model is necessarily an abstraction and therefore false.

The model economy selected in a particular application is not the one that best fits a particular set of economic statistics. It must fit along selected dimensions of reality given the question. To illustrate this idea, consider the question of how much of the higher average return on publicly traded stocks is a premium for bearing aggregate risk. The highly liquid short-term debt is called the safe asset. However, it is not a perfectly safe asset, as is the model economy's safe asset. A perfectly safe asset does not exist. Government debt is not safe because governments default fully or partially in extreme events. Therefore, the nature of the consumption process in the model economy used must not have the possibility of extreme events.

The model economy that Mehra and Prescott (1985) used to address this issue had only one type of infinitely lived households and a pure endowment process. We specified a Markov chain process on the growth rate of this endowment, which rules out extreme events. Equilibrium consumption was the output of the endowment process. The relation examined was the return on the endowment process and a security that paid one unit of consumption in the next market in the sequence with certainty in the sequence of market equilibria. Empirically, the difference in average yields on equity and short-term relatively risk-free liquid debt was over 6%. The finding was that only a small part of the difference in average yields on the two classes of securities was accounted for by a premium for bearing nondiversifiable aggregate risk.

Will a class of model economies with a richer class of processes on consumption growth rates resolve this puzzle? The answer is no because the abstraction used permits *any* stationary process on consumption growth rates. Our abstraction did rule out extreme events because truly risk-free assets do not exist.

This finding raised the question of what factors were giving rise to this big difference. McGrattan and Prescott (2005) subsequently learned that introducing taxes on distributions by corporations to owners reduced the premium by a third. Economic theory says it is after-tax distributions that should be considered in determining the return on different assets.

Another significant factor is the cost of managing assets. Pension funds have sizable costs that reduce the return on equity realized by households who are the indirect owners of the equity held by these funds. On the other hand, the cost of managing a portfolio of short-term liquid assets is small. The magnitude of the asset management and intermediation costs can be estimated using national income and product accounts. The aggregate value of the corporate equity held either directly or indirectly by the household sector can be estimated using aggregate balance sheet statistics. The annual costs are about 2% of the total value of the assets. This exercise was carried out in Mehra et al. (2011).

Most of the remainder of the difference in average yields is almost surely due to a liquidity premium for carry-out transactions. This leads to the conclusion that the equity premium puzzle is no longer a puzzle. Better measurement may identify a deviation from theory, but for the time being, theory is ahead of measurement with respect to the equity premium.

The model economy used to measure and estimate the premium for bearing nondiversifiable aggregate risk has no investment. In fact, investment is a sizable share of output. The model is not realistic along this dimension. However, this very simple model is sufficiently rich to address the question asked. The salient features of reality are incorporated into the model being used to address the given issue. The general principle is, if the question can be addressed with a simpler model, use the simpler one.

2.8 The Need for Discipline

A useful theory must have an associated discipline. Scientists, who employ the discipline and use the theory to answer a given question, reach the same conclusion as to what the theory says or does not say. Given the current state of the theory, the conclusion may state that the theory has to be extended before the question can be addressed. Or it may say that the answer depends on the magnitude of certain parameters, which have not yet been measured sufficiently accurately. The theory used in aggregate analysis is neoclassical growth theory. A crucial feature of this discipline is that when researchers extend the theory in order to resolve a deviation from theory or to expand its domain of applicability, the extended theory must be consistent with previously successful applications of the theory.

In the subsequent sections of this chapter, the development and use of neoclassical growth theory will be reviewed. This theory is applicable to virtually all substantive areas of economics including not only traditional business cycle fluctuations but also differences in per capita output levels across countries and across times. It is the theory in aggregate public finance, financial asset pricing, labor economics, monetary economics, environmental economics, and international finance.

The model economy used in an application is restricted by more disaggregated statistics. For example, the assumed time-to-build for new structures must be consistent with how long it typically takes to build a new structure. Econometricians have constructed statistical tests that rejected the Hansen (1985) model of business cycles. That model abstracted from time-to-build, because Hansen found this feature of reality to be of secondary importance in understanding business cycle fluctuations. Using data generated by the Kydland and Prescott (1982) model, which has a time-to-build technology, these statistical tests would lead to a rejection of the RBC model generating the data. It would be easy to come up with another test that would result in the rejection of the model with time-to-build. The implication is that statistical hypothesis testing is of little use in selecting a model to address some given question.

3. THE NATURE OF THE DISCIPLINE

3.1 The Back and Forth Between Theory and Measurement

The study of business cycle fluctuations led to the construction of dynamic stochastic general equilibrium models of these fluctuations. These early models had a quadratic household utility flow function and linear technology constraint. This research program did not produce models with national accounts that could be compared to the actual ones. Their use did not satisfy the Klein discipline. Examples of these early models include Sargent (1976) and Kydland and Prescott (1977). Another limitation was that using other observations in economics to restrict the choice of the model economy was difficult and, in some cases, impossible.

What turned out to be the big breakthrough was the use of growth theory to study business cycle fluctuations. A question is, why did it take so long before it was used for this purpose? The answer is that, based on micro theory reasoning, dynamic economic theory was viewed as being useless in understanding business cycle fluctuations. This view arose because, cyclically, leisure and consumption moved in opposite directions. Being that these goods are both normal goods and there is little cyclical movement in their relative price, micro reasoning leads to the conclusion that leisure should move procyclically when in fact it moves strongly countercyclically. Another fact is that labor productivity is a procyclical variable; this runs counter to the prediction of micro theory that it should be countercyclical, given the aggregate labor input to production. Micro reasoning leads to the incorrect conclusion that these aggregate observations violated the law of diminishing returns.

In order to use growth theory to study business cycle fluctuations, the investment-consumption decision and the labor-leisure decision must be endogenized. Kydland and Prescott (1982) introduced an aggregate household to accomplish this. We restricted attention to the household utility function for which the model economies had a balanced growth path, and this balanced growth path displayed the growth facts. With this extension, growth theory and business cycle theory were integrated. It turned out that the predictions of dynamic aggregate theory were consistent with the business cycle facts that ran counter to the conclusion of those using microeconomic reasoning.

That time-to-build model economy had only technology shocks, so the analysis was restricted to determining the consequences of different types of technological shock processes for the cyclical behavior of the neoclassical growth model. Kydland and Prescott (1982) found that if there are persistent technology shocks and the aggregate elasticity of labor supply is high, neoclassical growth theory can predict fluctuations of the business cycle variety. By construction, the model economy displayed the growth facts. However, the aggregate Frisch elasticity of labor supply is not tied down by the growth facts. Two questions needed to be answered before one could say that the neoclassical growth model displays business cycle fluctuations of the nature observed. The first question was whether the Frisch elasticity of the aggregate household labor supply was at least 3. The second question was whether technology shocks were highly persistent and of the right magnitude.

One criticism of Kydland's and my analysis was that empirically, cyclical labor productivity and total hours were roughly orthogonal during the period studied, whereas for the model economy, they were highly correlated. If productivity shocks were the only factor contributing to fluctuations, this would be a valid criticism, and business cycle fluctuations would be inconsistent with neoclassical growth theory. But productivity shocks were not the only factor giving rise to business cycle fluctuations during this period. To determine how much of the business cycle fluctuations were accounted for by productivity shocks, an estimate of the variance of these shocks was needed. This was provided by Prescott (1986). Given the estimate, labor productivity and aggregate hours worked should be roughly orthogonal, as they were during the period studied. The finding is that the US economy would have been 70% as volatile as it was during the period considered if productivity shocks were the only shocks.

The nature of the shock is important in the theory. If one thinks that all productivity change is due to the growth of knowledge useful in production, productivity shocks generally should be negative; in fact, however, productivity shocks are sometimes negative. One implication is that variations in the growth of the stock of useful knowledge cannot be the only reason for changes in productivity. Another factor giving rise to changes in productivity are changes in legal and regulatory constraints. Such changes can both increase and decrease productivity. The huge differences in productivity that are observed across countries provide strong evidence that the legal and regulatory systems are of great importance in determining the level of productivity.

3.2 Monopolistic Competition: Small Consequences for Business Cycle Accounting

Neoclassical growth theory assumes price taking in market transactions. Does abstracting from the fact that some businesses and groups of factor suppliers have market power and are not price takers alter the conclusions of the simple abstraction? Hornstein (1993) introduced monopolistic competition and found that for measuring the contribution of productivity shocks to business cycle fluctuations, it mattered little. He calibrated a monopolistic competitive model to the same set of statistics as those using the neoclassical growth model did. With monopolistic competition, the response to the shocks is greater, but this is offset by a smaller estimate of the variance of the underlying productivity shock. For this purpose, abstracting from market power mattered little for the estimate of the contribution of productivity shocks to business cycle fluctuations. For some other issues, this is probably not the case. This illustrates the way in which the theory progresses. A finding is successfully challenged by showing that introducing some feature of reality in a disciplined way changes the answer to the question. The results of unsuccessful challenges are of interest, for they add to the confidence in the original study.

3.3 Nonneutral Technological Change: Little Consequence in Basic Model

The relative price of the composite investment good and the composite consumption good has not been constant, as it is in the basic neoclassical growth model. Secularly, what is more or less constant is the value of investment goods produced relative to the value of all goods produced in nominal terms. A world in which the relative price of the investment good falls is one with the following aggregate production relation:

$$c_t + (1+\gamma)^{-t} x_t \leq A k_t^{\theta} h_t^{1-\theta},$$

where $\gamma > 0$. There is balanced growth with the relative price of the investment good to the consumption good falling at rate γ. Greenwood et al. (1988) show this. Another interesting finding in their paper concerns the nature of depreciation for the theory of business cycle fluctuations.

3.4 Nature of Depreciation: Matters

The standard abstraction for depreciation is the perpetual inventory assumption with a constant depreciation rate:

$$k_{t+1} = (1-\delta)k_t + x_t.$$

Greenwood et al. (1988) assume that the rate of depreciation increases with the intensity of the use of capital; that is, they assume a Taubman and Wilkinson (1970) depreciation technology. Let u_t denote the capital utilization rate. Capital services provided are $u_t k_t$. The depreciation rate is an increasing function of the utilization rate, $\delta_t = \delta(u_t)$. With this assumption, the response to productivity shocks is bigger and the aggregate elasticity of labor supply smaller for the model calibrated to the growth facts.

I am sure that this alternative theory of depreciation was considered by the national income and product accountants and found not to be important. It is true that during periods of high economic activity, some capital is utilized more intensely. However, for many capital goods, depreciation does not depend on the intensity of use. One reason is that during boom periods, machines are well maintained in order to keep them operating efficiently. Better maintenance lowers the depreciation rate. Higher occupancy rates of office buildings do not increase their depreciation rate. The national accounts stuck with the perpetual inventory method and useful life in calculating aggregate depreciation because it was consistent with the prices of used capital equipment. This is another example of micro evidence restricting the model economy being used to address an aggregate issue.

If this alternative theory of depreciation had passed the micro test, it would have introduced a number of discrepancies within the theory. Business cycle observations would imply a smaller aggregate labor supply elasticity, and this in turn would imply that the theory predictions for cross-country differences in aggregate labor supply arising from differences in the marginal tax rate on labor income would be much smaller than what they are. About the only way to resolve these discrepancies would be to assume country-specific differences in preferences that give rise to both higher marginal tax rates and lower labor supply. With this resolution, however, there would be big discrepancies between the predictions of theory for aggregate labor supply during growth miracles.

The important point is that preference and technology parameters, with the discipline reviewed here, must be consistent across applications.

3.5 Monetary Policy: Little Consequence for Business Cycle Fluctuations

The general view prior to the development of quantitative aggregate economic theory was that monetary policy had important real consequences for the behavior of real variables, in particular real output and employment. Once explicit transactions abstractions were developed that gave rise to a demand for money, it was possible to introduce them into the neoclassical growth theory and to assess their quantitative consequences for real variables. Cooley and Hansen (1995) did this and found that the real consequences were small for monetary policies that did not give rise to very high rates of inflation. This supported the empirical findings of Sargent and Sims (1977) that real movements were not the result of monetary factors in the postwar US economy.

Sticky wage and nominal staggered wage contracting arrangements were subsequently introduced into the neoclassical growth model and their quantitative consequences for real findings determined by Chari et al. (2000). The finding was that these mechanisms did not give rise to business cycle fluctuations of the nature observed.

Another bit of strong evidence for the unimportance of monetary policy is the fact that RBC models that abstract from monetary factors do not have large deviations from observations during periods with high variations in inflation rates, such as during the period 1978–82 in the United States.

3.6 Two Important Methodological Advances

In critiquing the use of neoclassical growth theory to study business cycle fluctuations, Summers (1986) asked a good question: What are these shocks? An important methodological advancement to the theory was needed before his question could be answered. The advancement was path analysis.

3.6.1 Path Analysis

Hansen and Prescott (1993) used path analysis when they addressed the question of whether technology shocks caused the 1990–91 recession. In that paper, the dynamic system for the model was used to generate time paths of the variables given the realized values of the stocks. The finding was that yes, productivity shocks did cause that recession.

That paper offered another interesting finding. A prediction of the technology-shock-only model is that the economy should have recovered in 1993–94, since productivity had returned to trend. Other factors had to be depressing the economy during this period. Subsequently, the factors were identified. They were increases in tax rates.

3.6.2 Distribution of Firms with Inventories a State Variable

A widely held view was that inventory behavior was important for understanding business cycle fluctuations given the large cyclical variability of inventory investment. The micro theory of inventory investment was developed, but introducing this feature into quantitative neoclassical growth theory was impossible given the lack of needed tools.

Fisher and Hornstein (2000) developed a way to introduce inventory investment when firms faced fixed resource costs when making an inventory investment. This made the stock of inventory a firm state variable and the distribution of firms as indexed by their inventory stock an aggregate state variable. This methodological advance was also used by Hornstein (1993) to assess the quantitative importance of monopolistic competition.

3.7 The Big Aggregate Economic Puzzle of the 1990s

A boom in output and employment in the United States began about 1994 and continued until the very end of the decade. This boom was puzzling from the perspective of what was then aggregate economic theory. In this boom, the corporate profit share of GNI was low. In other booms, this share was higher than normal. Another puzzling observation was that GDP per hour, the commonly used measure of productivity, was low in this boom. Normally, productivity accounts for about a third of the cyclical variation in GDP and market hours the other two-thirds. In this boom, the accounting was 125% due to market hours worked and negative 25% due to productivity. No changes in labor market policies or tax rates could account for these phenomena. This puzzle remained open for at least 6 years. One explanation consistent with general equilibrium theory was that Americans—as well as Europeans—experienced a contagious case of workaholism; that is, the rate at which people's willingness to substitute labor for leisure in the aggregate changed. Such explanations violate the discipline of dynamic aggregate theory reviewed in this essay.

To answer this question, two developments in quantitative aggregate theory were crucial. One was the use of an equilibrium condition for a class of economies that depend on current-period variables to account for the large differences in hours worked per working-age person across countries and across time. This equilibrium condition used was that the marginal rate of consumption and leisure is equal to the after-tax wage. A Cobb–Douglas production function was assumed, so the wage was just aggregate labor income divided by aggregate hours.[c] The elasticity of substitution between consumption and leisure for the aggregate household was the same as the one needed for the neoclassical growth model to display business cycle fluctuations.

The reason that Western Europeans now work 30% less than other advanced industrial countries is not that they are lazy or are better at making use of nonmarket productive time. It is that these countries have higher marginal tax rates on labor income and on consumption. These higher tax rates introduce a large tax wedge between the intratemporal marginal rate of substitution and the marginal rate of transformation between consumption and market time.

The second development was to use this methodology to account for the large secular movements in the value of corporations relative to GNP in the United States and the United Kingdom in the 1960–2000 period. The equilibrium relation used for the class of models considered was the following one. The market value of corporations is equal to the market value of the capital stocks owned by the firm. Given the importance of intangible capital in determining the value of corporations, this stock had to be included in the analysis. Brand names, organization capital, patents, and technology know-how embodied in the organization all contribute to the value of the business enterprise.

With these two developments, the stage was set for resolving the US hours boom of the 1990s.

4. MAJOR DEVELOPMENTS AND THEIR APPLICATIONS POST-1995

Important theoretical advancements in neoclassical growth theory have continued to occur and have expanded the theory's applicability. Also important was the development of new and better data sets that are easily accessible. These data sets are more uniform across countries, which facilitates the study of factors giving rise to international differences in economic aggregates. Increases in computing power made possible the introduction of demographics into models being used to draw scientific inference using the theory. The life cycle is crucial for understanding aggregate savings behavior as it gives rise to savings for retirement.

4.1 Clubs in the Theory and France's 35-Hour Workweek Policy

A development in valuation theory was the introduction of clubs. Clubs are arrangements that internalize externalities, whether they are positive or negative, within organizations that are small relative to the economy. One extremely important type of club is the household. In classical valuation theory, household clubs are a primitive. For each household, there is an agent that chooses an optimal point in a subset of the commodity

[c] This is the measure of wages used by Lucas and Rapping (1969) when they introduced labor supply into macroeconometric modeling.

space—that is, in that household's consumption possibility set—subject to its budget constraint. Business organizations are clubs as well. A firm is defined by its production possibility set, which is a subset of the commodity space, and the households' shares of ownership. Cole and Prescott (1997) extend valuation equilibrium theory to permit clubs.

To date, this development has been little used in quantitative aggregate analyses. To the best of my knowledge, I am aware of only one aggregate quantitative application using clubs. This application is due to Fitzgerald (1998), who uses this extension of the basic theory to predict the consequences of France's 35-hour workweek constraint. His framework has two types of households and two types of labor services: skilled and unskilled. Type 1 household can only supply unskilled labor. Type 2 household can supply either type. The important constraint is that for each firm, the work schedule of those performing the skilled and the unskilled tasks must be equal. The skilled workers' tasks include supervising, monitoring, and coordinating the unskilled workers.

The goal of the French 35-hour workweek policy was to help the unskilled and not the highly paid skilled workers. It turned out that the skilled are made better off under the 35-hour workweek and the unskilled worse off, counter to this objective. The legal constraint, which changed the technology set of a firm, had an unintended consequence. The program did have the intended consequence of increasing the employment rate of the unskilled.

4.2 Cartelization Policies and the Resolution of the US Great Depression Puzzle

Cole and Ohanian (1999) initiated a program of using the theory to study great depressions. They found a big deviation from the theory for the 1930–39 US Great Depression. This deviation was the failure of market hours per working-age person to recover to their predepression level. Throughout the 1930s, market hours per working-age person were 20–25% below their predepression level. The reasons for depressed labor supply were not financial. No financial crises occurred during the period 1934–39. The period had no deflation, and interest rates were low. This led Cole and Ohanian to rule out monetary policy as the reason for the depressed labor supply. Neither was the behavior of productivity the reason. Productivity recovered to trend in 1934 and subsequently stayed near the trend path.

These findings led Cole and Ohanian to search for an extension of the theory that would resolve this puzzling failure of the US economy to recover in the 1930s. They observed that relative wages in the cartelized industries increased relative to those in the noncartelized industries. Employment in the cartelized industries was the most depressed and did not recover. Those in the cartelized industries were the insiders and those in the competitive industries the outsiders. The problem Cole and Ohanian had to solve was to figure out how to introduce a cartelization arrangement into quantitative aggregate theory.

Eventually, Cole and Ohanian (2004) figured out a way and found that the cartelization policy was a major factor in accounting for the failure of the US economy to recover from the Great Depression subsequent to the recovery of productivity. They estimated that the cartelization policy alone accounted for over half of the depression in employment in the US Great Depression of the 1930s. It turned out that tax and wage policies can account for much of the remainder, so the Great Depression is no longer a puzzle.

McGrattan (2012) extended the theory to permit the consequences of expected future tax rate increases on the distributions from businesses to their owners. She found that they were important in accounting for the great decline in output in 1930. Businesses made large cash distributions to their owners rather than using cash to finance new investment. Fisher and Hornstein (2002) established that wage policies that set the wage above equilibrium value gave rise to the Great Depression in Germany from 1927 to 1932. The elimination of these policies late in 1932 resulted in rapid recovery from Germany's Great Depression, just as theory predicts.

4.3 Taxes and Country Labor Supply: Cross-Application Verification

The question is whether the theory used to study business cycle fluctuations accounts for the large difference in labor supply, as measured by market hours per working-age person, between Americans and Western Europeans. During the period 1993–96, Americans worked on average 40% more than did the French, Italian, and Germans. This was not always the case. In the period 1970–74, market hours per working-age person were comparable in both the United States and Western Europe and comparable to what they are now in the advanced industrial countries, with the notable exception of Western Europe.

The equilibrium relation used in Prescott (2004) to predict the difference in labor supply as a function of the effective tax rate on labor income was that the marginal rate of substitution between nonmarket productive time and consumption is equal to the after-tax real wage. A Cobb–Douglas aggregate production was assumed.

This equilibrium condition for country i can be written as

$$h_{it} = \frac{1-\theta}{1-\theta + \frac{c_{it}}{y_{it}} \frac{\alpha}{1-\tau_{it}}}.$$

Here, θ is the capital share parameter, α the value of leisure parameter, h_{it} the market hours per working-age person, τ_{it} the effective average marginal tax rate on labor income, and c_{it}/y_{it} the fraction of aggregate output consumed.

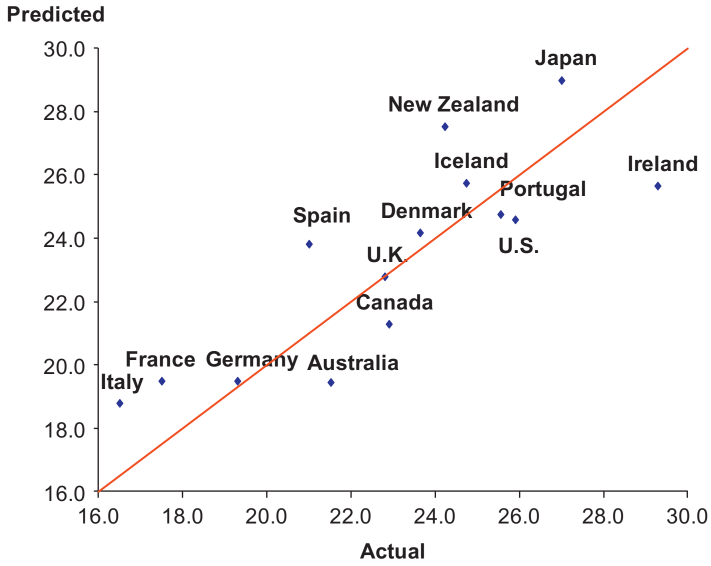


Fig. 1 Predicted and actual hours worked per working-age person, 1990–2002.

The analysis has only one free parameter, namely, the preference parameter α. This parameter is not tied down by the balanced growth facts. The capital income share parameter was nearly constant across countries and periods and was set equal to 1/3. The preference parameter α was picked so that the relation held for the United States.

The US boom in the 1990s was unlike previously studied booms and was at variance with the basic neoclassical growth model as discussed previously. Fig. 1 plots predicted and actual hours worked per working-age person for the period 1990–2002 using the model without intangible capital. It was a puzzle in the theory that remained open for 8 years. No alternative theory predicted this boom.

4.4 Use of the Overlapping Generations Abstract

For many issues, it does not matter whether the dynastic family or the overlapping generation structure is used. Before the great increase in computing capabilities, using the overlapping generation structure was not feasible. Braun et al. (2009) exploited this increase in computing capabilities and found that both the dynasty and the overlapping generation constructs are consistent with the fall in Japanese savings rates in the 1990s. However, the two constructs for aggregate households imply very different behavior for the Japanese savings rate post-2010. Because of Japan's large baby boom in the 1960s, the fraction of people who were dissaving to finance retirement would increase subsequent to 2010, and the aggregate savings rate would fall. Quantitatively, the savings rate did just what the theory with an overlapping generation structure predicted it would do.

5. INTANGIBLE CAPITAL EXPANDS THE APPLICABILITY OF THE THEORY

That intangible capital investment financed and owned by firms is big has never been in dispute. A question is why intangible capital was not incorporated into quantitative aggregate theory. The answer is that there was no disciplined way to incorporate this largely unreported component of output into the theory. The development of a consistent set of balance sheets for the household and business sectors was key to resolving this problem. Balance sheets, among other things, report the value of ownership of corporate equity.

5.1 The Value of Corporate Businesses

The price of capital good K_j is $q_j(\pi)$, where π specifies tax policy. Tax policy includes not only tax rates on corporate accounting profits but also the tax rate on distributions to owners, the nature of the capital consumption allowance, and the inflation rate. An important input to production is the services of human capital owned by the employees of the corporation. It is rivalrous and does not show up in the value of corporations. Consequently, it need not be included in the model used to determine the value of corporate businesses.

The aggregate corporate market value V, where subscript T denotes tangible capital and subscript I denotes firm-owned intangible capital, is

$$V = q_T(\pi)K_T + q_I(\pi)K_I.$$

If there were no capital income taxes, the prices of capital in units of the consumption good would be 1. But there are capital income taxes.

The price of one unit of tangible capital in terms of the consumption good, given that nearly all investment is financed through retained earnings, is

$$q_T = (1 - \tau_{\text{dist}}),$$

where τ_{dist} is the tax rate on distributions from corporations to owners. The average marginal tax rate on distributions is used. In the 1960s, virtually all distributions were in the form of dividends. The tax rate used was the average of the individual marginal tax rates weighted by the total dividends received by the group subject to that marginal tax rate. In the 1960s, this average tax rate was about 45%. Beginning in the 1980s, buybacks began to be used and permitted distributions to be deferred to when the capital gains were realized. This lowered the average tax on distributions.

Intangible capital was expensed, and as a consequence, its price to the owners of the businesses making the investment is smaller than the cost of producing it. The price of intangible capital is

$$q_I = (1 - \tau_{\text{dist}})\left(1 - \tau_{\text{corp profits}}\right).$$

In both the United States and the United Kingdom, there were large movements in V relative to annual GNI over the period studied by McGrattan and Prescott (2005) using this theory. The V/GNI number varied by a factor of 2.5 in the United States and by a factor of 3.0 in the United Kingdom during the period 1860–2000. This variation was not due to variation in the ratio of after-tax corporate income to GNI. This ratio varied little over the period. The theory found that the reason for the large secular changes was due to changes in taxes and regulations. Intangible capital was an important part of the value of corporations.

The big change in the tax system that increased the value of corporations was the deferred compensation individual savings account. These accounts permitted households to save for retirement free of capital income taxes. Insofar as the withdrawals are used to finance retirement consumption, there is no intertemporal wedge between the marginal rate of substitution between current and future consumption and the marginal rate of transformation between current and future consumption.

The added capital alone had little consequence for business cycle fluctuation accounting, so no new puzzles were created with this extension. An old puzzle that has not been resolved is the LeRoy and Porter (1981) and Shiller (1981) excess asset price volatility puzzle. Indeed, by looking at the values of the capital stocks owned by firms rather than at the present value of dividends, McGrattan and Prescott (2005) strengthened this excess volatility puzzle. These capital stocks vary smoothly, so the theory predicts their prices should as well.

In the model with intangible capital owned by business enterprises, we used an alternative aggregate production technology to the aggregate production function. There are three inputs: the services of tangible capital, the services of rival human capital, and the services of intangible capital. There are two output goods: one the composite output good less intangible capital investment and the other intangible capital investment. There were two *activities*: one producing intangible capital and one producing other final goods.

It is not a two-sector model because the services of intangible capital are not allocated between activities, as are the services of the other two inputs, but are used in both simultaneously by both activities. Otherwise, the production technology is standard. Letting Y_1 be output less intangible investment output, Y_2 intangible investment output, K_T tangible capital stock, K_I intangible capital stock, and L rival human capital services (labor), total output of the two activities is

$$\begin{aligned} Y_1 &= A_1 F_1(K_{T1}, K_I, L_1) \\ Y_2 &= A_2 F_2(K_{T2}, K_I, L_2) \\ K_T &= K_{T1} + K_{T2} \\ L &= L_1 + L_2 \end{aligned}$$

One unit of capital produces one unit of its services. All variables implicitly have a time subscript including the productivity parameters A_1 and A_2. The functions F_1 and F_2 have all the standard properties of the aggregate production function.

The important feature of the technology is that K_I has no activity subscript. A brand name can be used to produce a product sold in the market as well as in the development of a related product. The same is true of patents. The other two inputs are allocated between the activities. If productivity change is neutral in the sense that A_{1t}/A_{2t} stays constant, the implications for business cycles are the same. Thus, this technology works where the basic neoclassical growth model works. This part of the discipline is satisfied.

A problem is that most intangible capital investment made by firms and owned by firms is expensed and therefore not part of measured output. The question is how to incorporate this unobservable in a disciplined way. McGrattan developed a way (see McGrattan and Prescott, 2010b). The size of intangible capital net investment has implications for accounting profits of the corporate sector. Knowing the initial stock, the stocks can be computed from statistics reported in the national income and product accounts (NIPA).

5.2 US Hours Boom in the 1990s: A Crisis in RBC

The basic neoclassical growth theory model accurately predicted the behavior of the US economy prior to the 1990s, taking productivity taxes and demographics as exogenous. Theory was then ahead of measurement. In the 1990s it did not predict accurately. Market hours boomed while GDP per hour, the usual measure of productivity relative to trend, declined. The simple accounting was that the labor input accounted for 125% of the output and the standard measure of productivity for *minus* 25%. Typically, hours account for about two-thirds of the detrended change and productivity for the other third.

Taxes were not the answer, since the intratemporal tax wedge was, if anything, larger than before the boom. There were no major labor market reforms that improved the performance of the labor market. Economists were faced with the puzzle of why people were working so much. Fig. 2 plots the predicted and actual paths using the basic growth model without the introduction of intangible capital into the theory.

It was recognized that large investments in intangible capital were being made, and most were not reported as part of output because they were expensed. At the time, only computer software investment was reported.

Aggregate economics is not the only science with unobservable variables. A translation of a quote by Albert Einstein reads: “Not everything that counts can be counted, and not everything that can be counted counts.” The key relation is the accounting profit equation. The bigger the net unmeasured intangible investment, the smaller were these problems. This finding, along with the fact that accounting profits were a small share of GDP in this hours boom period, is consistent with intangible investment being large. Other evidence is from the National Science Foundation. The NSF provides estimates of private R&D expenditures, which are an important

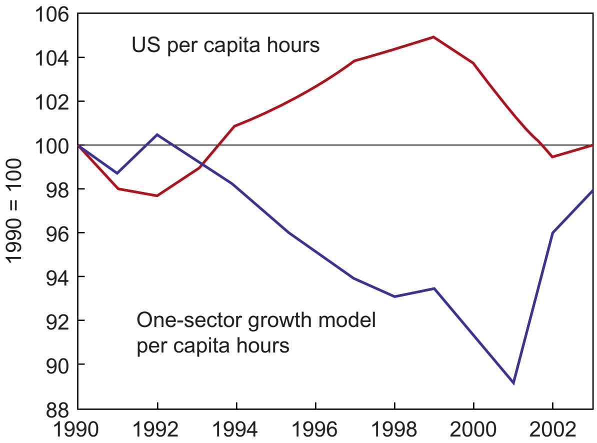


Fig. 2 Without intangible capital: big deviation from theory.

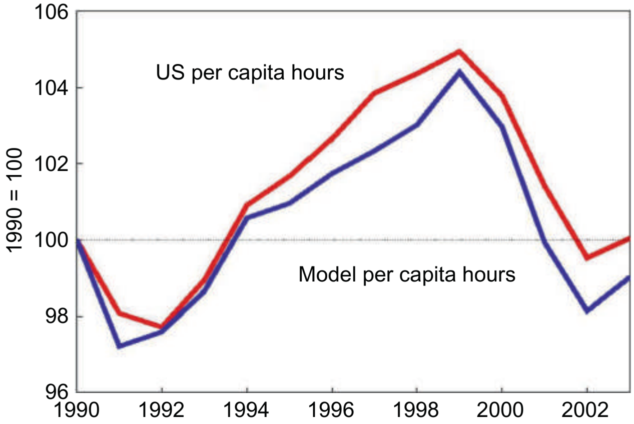


Fig. 3 With intangible capital: no deviation from theory.

component of intangible capital investment. These investment expenditures in percentage terms increased much more than measured investment expenditures during the 1990s boom.

With the introduction of intangible capital and nonneutral technology change in the production of GDP and intangible capital investment, measurement was again in conformity with theory. This is shown in Fig. 3.

The extended theory accounts for capital gains reported in the Federal Reserve System's flow of funds accounts. About half of these investments are financed by the owners of corporations subject to the corporate income tax and half by worker-owners of other businesses, which matches with micro observations.

5.3 Technology Capital

Intangible capital falls into different categories. Some are specific to the local production units and market. Some are assets with services that can be used at multiple locations. Virtually every metropolitan area in the United States has the same set of major retailers. Each of these major retailers uses the same know-how and name for all their retail outlets. The branches rely on their central headquarters for supply-side management, financial services, and advertising services. Intangible capital that can be used at multiple locations is technology capital. Investment in this type of capital is financed by location rents.

There are no increasing returns to scale, even though a closed economy with more locations will be richer than a closed economy with fewer locations, other things being equal. A production unit at a given location faces decreasing returns to scale. The production unit, being a price taker, realizes location rents. With technology capital, a reason for foreign direct investment (FDI) exists.

5.4 Use in Estimating Gains from Openness

Estimating gains from openness was originally introduced to study the role of openness in economic development (see McGrattan and Prescott, 2009). The observation was that for 50 years prior to World War II, the EU-6 GDP labor productivity was only a little more than half that of the United States, as it was in 1957 when the Treaty of Rome was signed. In the subsequent 30 years, EU-6 productivity caught up to that in the United States. This strongly suggests that openness fosters economic development. The role of trade can account for only one-ninth of the gain if the model used in the estimation is restricted to be consistent with the trade flows. Technology capital accounts for about one-third. This evidence indicates that other factors associated with openness are even more important. Two factors that have not yet been incorporated into the theory that empirically seem important are the faster diffusion of public knowledge and increasing competition reducing barriers to adopting more efficient technologies in production.

The technology extension has already permitted the theory to be used to assess China's direct foreign investment policy. Holmes et al. (2015) find that the Chinese policy of requiring access to technology capital of the foreign multinational making FDI in China in return for access to the huge Chinese market was in China's economic interest. In making these restrictions, China is violating the rules of the World Trade Organization. With the renminbi gaining reserve currency, interest in becoming more open to direct foreign investment will increase in China. This illustrates the usefulness of the theory in still another area, and, as stated earlier, usefulness is one criterion for a successful scientific theory.

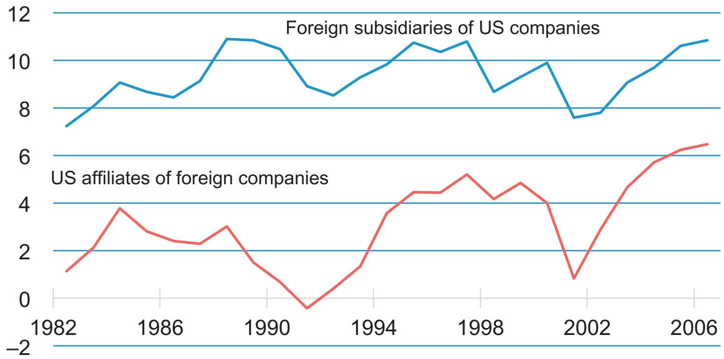


Fig. 4 BEA average FDI annual returns.

5.5 Use in Accounting for Features of US Current Accounts

A feature of US current accounts is the high reported earnings of US companies on their FDI and the low reported earnings of foreign companies' FDI in the United States. As reported by the Bureau of Economic Analysis (BEA), during the period 1982–2005, US companies earned an average return of 9.3 percentage points on their FDI, whereas foreign companies earned an average of 3.0 percentage points on their US FDI. Annual average returns for the period are plotted in Fig. 4. A question addressed by McGrattan and Prescott (2010a) naturally arises: why is the return differential so large and persistent?

The introduction of technology capital accounts for over 60% of the difference. Intangible capital investment stock is important because it increases profits but not the BEA stock of capital. It does increase the stock of capital, which lowers the economic return. US multinationals made large FDI earlier and, as a result, have relatively larger stocks of intangible capital than foreign multinationals have in their US subsidiaries. The age of the foreign subsidiaries matters because intangible investment is high and therefore BEA profits low when they are young. This micro evidence strongly supports the theory.

Using economic returns, the differential between the average return on US FDI and the average return on foreigners' FDI is reduced from 6.3 percentage points to about 2.5 percentage points. A question that naturally arises is, what accounts for the remaining 40% of the difference? Corporate tax rates differ across countries, and through transfer pricing, profits are shifted to countries where this tax rate is lower. Indeed, an important field of corporate finance is concerned with setting prices for goods and services transferred between multinationals and their foreign subsidiaries.

6. CONCLUDING COMMENTS

So much has been learned through the successful use of neoclassical growth theory and its extensions. This theory has directed the development of aggregate economics.

The availability of better data sets is fostering further development. As these better data sets become available, great progress is being made in incorporating features of the household sector,[d] which, like the business sector, is of great economic importance. In the earlier stages of the development and use of neoclassical growth theory, the household was a primitive. Now, however, its structure is becoming an endogenous element. The household sector has changed significantly over time and is not policy invariant.

In reporting household sector statistics, a household is the set of people residing at a dwelling—that is, a postal address. The size of households has changed significantly in the United States. Further, many households consist of married couples. Over time, the nature of matching has changed, as found by Greenwood et al. (2016). They find an important change is the increase in positive assortative matching. With more two-professional households, these changes have had major consequences for the distribution of household incomes.

Another important economic sector is the government sector. The question of how a group of people can set up sustainable collective government arrangements that result in outcomes preferred by the members of this group is an important one. Answering this question will require developments in pure theory.

Through the interaction of theory and measurement, the rapid development of quantitative aggregate economic theory is certain to continue. It will be interesting to see what these developments are.

REFERENCES

Adelman, I., Adelman, F., 1959. The dynamic properties of the Klein-Goldberger model. Econometrica 27 (4), 596–625.

Arkolakis, C., Costinot, C., Rodríquez-Clare, A., 2012. New trade models, same old gains? Am. Econ. Rev. 102 (1), 94–103.

Braun, R.A., Ikeda, D., Joines, D.H., 2009. The saving rate in Japan: why it has fallen and why it will remain low. Int. Econ. Rev. 50 (1), 291–321.

Chari, V.V., Kehoe, P., McGrattan, E.R., 2000. Sticky price models of the business cycle: can the contract multiplier solve the persistence problem? Econometrica 68 (5), 1151–1179.

Cole, H.L., Ohanian, L.E., 1999. The Great Depression in the United States from a neoclassical perspective. Fed. Reserve Bank Minneapolis Quart. Rev. 23 (1), 2–24.

Cole, H.L., Ohanian, L.E., 2004. New Deal policies and the persistence of the Great Depression: a general equilibrium analysis. J. Polit. Econ. 112 (4), 779–816.

Cole, H.L., Prescott, E.C., 1997. Valuation equilibrium with clubs. J. Econ. Theory 74 (1), 19–39.

Cooley, T.F., Hansen, G.D., 1995. Money and the business cycle. In: Cooley, T.F. (Ed.), Frontiers of Business Cycle Research. Princeton University Press, Princeton.

Fisher, J.D.M., Hornstein, A., 2000. (*S*,*s*) inventory policies in general equilibrium. Rev. Econ. Stud. 67 (1), 117–145.

Fisher, J.D.M., Hornstein, A., 2002. The role of real wages, productivity, and fiscal policy in Germany's Great Depression 1928–1937. Rev. Econ. Dyn. 5 (1), 100–127.

[d] McGrattan et al. (1997) introduce home production by the household into the theory. The implications for business cycles did not change.

Fitzgerald, T.J., 1998. Work schedules, wages, and employment in a general equilibrium model with team production. Rev. Econ. Dyn. 1 (4), 809–834.

Frisch, R., 1933. Propagation problems and impulse problems in dynamic economics. In: Kock, K. (Ed.), Economic Essays in Honour of Gustav Cassel. Allen & Unwin, London.

Frisch, R., 1970. From Utopian Theory to Practical Applications: The Case of Econometrics. (Lecture to the Memory of Alfred Nobel, June 17).

Greenwood, J., Hercowitz, Z., Huffman, G.W., 1988. Investment, capacity utilization, and the real business cycle. Am. Econ. Rev. 78 (3), 402–417.

Greenwood, J., Guner, N., Kocharkov, G., Santos, C., 2016. Technology and the changing family: a unified model of marriage, divorce, educational attainment and married female labor-force participation. Am. Econ. J. Macroecon. 8 (1), 1–41.

Hansen, G.D., 1985. Indivisible labor and the business cycle. J. Monet. Econ. 16 (3), 309–327.

Hansen, G.D., Prescott, E.C., 1993. Did technology shocks cause the 1990–1991 recession? Am. Econ. Rev. 83 (2), 280–286.

Hansen, G.D., Prescott, E.C., 2005. Capacity constraints, asymmetries, and the business cycle. Rev. Econ. Dyn. 8 (4), 850–865.

Hansen, G.D., Sargent, T.J., 1988. Straight time and overtime in general equilibrium. J. Monet. Econ. 21 (213), 281–304.

Hodrick, R.J., Prescott, E.C., 1980. Post-War U.S. Business Cycles: An Empirical Investigation. Northwestern University, Evanston, IL, pp. 1–28 (Discussion Paper No. 451).

Holmes, T.J., McGrattan, E.R., Prescott, E.C., 2015. Quid pro quo: technology capital transfers for market access in China. Rev. Econ. Stud. 82 (3), 1154–1193.

Hornstein, A., 1993. Monopolistic competition, increasing returns to scale, and the importance of productivity shocks. J. Monet. Econ. 31 (3), 299–316.

Hornstein, A., Prescott, E.C., 1993. The firm and the plant in general equilibrium theory. In: Becker, R., Boldrin, R., Jones, R., Thomson, W. (Eds.), General Equilibrium, Growth, and Trade II: The Legacy of Lionel McKenzie. Academic Press, San Diego, pp. 393–410.

Kaldor, N., 1957. A model of economic growth. Econ. J. 67 (268), 591–624.

Kondratieff, N.D., Stolper, W.F., 1935. The long waves in economic life. Rev. Econ. Stat. 17 (6), 105–115.

Kuznets, S., 1930. Secular movements in production and prices: their nature and their bearing upon cyclical fluctuations. Am. Econ. Rev. 20 (4), 787–789.

Kydland, F.E., Prescott, E.C., 1977. Rules rather than discretion: the inconsistency of optimal plans. J. Polit. Econ. 85 (3), 473–491.

Kydland, F.E., Prescott, E.C., 1982. Time to build and aggregate fluctuations. Econometrica 50 (6), 1345–1370.

LeRoy, S.F., Porter, R.D., 1981. The present-value relation: tests based on implied variance bounds. Econometrica 49 (3), 555–574.

Lucas Jr., R.E., 1972. Expectations and the neutrality of money. J. Econ. Theory 4 (2), 103–123.

Lucas Jr., R.E., 1976a. Econometric policy evaluation: a critique. Carnegie-Rochester Conference Series on Public Policy, vol. 1. Elsevier, Amsterdam, pp. 19–46.

Lucas Jr., R.E., 1976b. Understanding business cycles. Carnegie-Rochester Conference Series on Public Policy, vol. 5. Elsevier, Amsterdam, pp. 7–29.

Lucas Jr., R.E., Rapping, L.A., 1969. Real wages, employment and inflation. J. Polit. Econ. 77 (5), 721–754.

Lucas Jr., R.E., Sargent, T.J., 1979. After Keynesian macroeconomics. Fed. Reserve Bank Minneapolis Quart. Rev. 3 (2), 1–16.

MaCurdy, T.E., 1981. An empirical model of labor supply in a life cycle setting. J. Polit. Econ. 89 (6), 1059–1085.

McGrattan, E.R., 2012. Capital taxation during the U.S. Great Depression. Q. J. Econ. 127 (3), 1515–1550.

McGrattan, E.R., Prescott, E.C., 2005. Taxes, regulations, and the value of U.S. and U.K. corporations. Rev. Econ. Stud. 72 (3), 767–796.

McGrattan, E.R., Prescott, E.C., 2009. Openness, technology capital, and development. J. Econ. Theory 144 (6), 2454–2476.

McGrattan, E.R., Prescott, E.C., 2010a. Technology capital and the U.S. current account. Am. Econ. Rev. 100 (4), 1493–1522.

McGrattan, E.R., Prescott, E.C., 2010b. Unmeasured investment and the puzzling U.S. boom in the 1990s. Am. Econ. J. Macroecon. 2 (4), 88–123.

McGrattan, E.R., Rogerson, R., Wright, R., 1997. An equilibrium model of the business cycle with household production and fiscal policy. Int. Econ. Rev. 38 (2), 267–290.

Mehra, R., Prescott, E.C., 1985. The equity premium: a puzzle. J. Monet. Econ. 15 (2), 145–161.

Mehra, R., Piguillem, F., Prescott, E.C., 2011. Costly financial intermediation in neoclassical growth theory. Quant. Econ. 2 (1), 1–36.

Mitchell, W., 1913. Business Cycles. University of California Press, Berkeley.

Mitchell, W., 1927. Business Cycles: The Problem and Its Setting. National Bureau of Economic Research, New York.

Prescott, E.C., 1986. Theory ahead of business cycle measurement. Fed. Reserve Bank Minneapolis Quart. Rev. 10 (4), 9–22.

Prescott, E.C., 2004. Why do Americans work so much more than Europeans? Fed. Reserve Bank Minneapolis Quart. Rev. 28 (1), 2–13.

Prescott, E.C., Rogerson, R., Wallenius, J., 2009. Lifetime aggregate labor supply with endogenous workweek length. Rev. Econ. Dyn. 12 (1), 23–36.

Rogerson, R., 1984. Topics in the Theory of Labor Markets. (PhD Thesis). University of Minnesota.

Rosen, S., 1978. The supply of work schedules and employment. In: Work Time and Employment: A Conference Report. National Commission for Manpower Policy, Washington, DC.

Samuelson, P.A., 1939. A synthesis of the principle of acceleration and the multiplier. J. Polit. Econ. 47 (6), 786–797.

Sargent, T.J., 1976. A classical macroeconometric model for the United States. J. Polit. Econ. 84 (2), 207–238.

Sargent, T.J., Sims, C.A., 1977. Business cycle modeling without pretending to have too much *a priori* economic theory. In: Sims, C.A. (Ed.), New Methods in Business Cycle Research: Proceedings from a Conference. Federal Reserve Bank of Minneapolis, Minneapolis, pp. 45–110.

Shiller, R.J., 1981. Do stock prices move too much to be justified by subsequent changes in dividends? Am. Econ. Rev. 71 (3), 421–436.

Solow, R.M., 1956. A contribution to the theory of economic growth. Quant. J. Econ. 70 (1), 65–94.

Summers, L.H., 1986. Some skeptical observations on real business cycle theory. Fed. Reserve Bank Minneapolis Quart. Rev. 10 (4), 23–27.

Swan, T.W., 1956. Economic growth and capital accumulation. Econ. Rec. 32 (2), 334–361.

Taubman, P., Wilkinson, M., 1970. User cost, capital utilization and investment theory. Int. Econ. Rev. 11 (2), 209–215.

Thomas, J.K., 2002. Is lumpy investment relevant for the business cycle? J. Polit. Econ. 110 (3), 508–534.

Tinbergen, J., 1952. Business cycles in the United Kingdom, 1870–1914. Econ. J. 62 (248), 872–875.

Wicksell, K., 1918. Ett bidrag till krisernas teori. Review of Goda och daliga tider. Ekonomisk Tidskrift 20 (2), 66–75.

CHAPTER 23

Families in Macroeconomics

M. Doepke*, M. Tertilt†
*Northwestern University, Evanston, IL, United States
†University of Mannheim, Mannheim, Germany

Contents

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.04.006

Abstract

Much of macroeconomics is concerned with the allocation of physical capital, human capital, and labor over time and across people. The decisions on savings, education, and labor supply that generate these variables are made within families. Yet the family (and decision making in families) is typically ignored in macroeconomic models. In this chapter, we argue that family economics should be an integral part of macroeconomics and that accounting for the family leads to new answers to classic macro questions. Our discussion is organized around three themes. We start by focusing on short- and medium-run fluctuations and argue that changes in family structure in recent decades have important repercussions for the determination of aggregate labor supply and savings. Next, we turn to economic growth and describe how accounting for families is central for understanding differences between rich and poor countries and for the determinants of long-run development. We conclude with an analysis of the role of the family as a driver of political and institutional change.

Keywords

Family economics, Macroeconomics, Business cycles, Growth, Households, Fertility, Labor supply, Human capital, Gender

JEL Classification Codes

E20, E30, J10, J20, O40

1. INTRODUCTION

First impressions suggest that family economics and macroeconomics should be the two fields within economics at the greatest distance from each other: one looks at interactions between at most a handful of members of the same family, whereas the other considers the aggregated behavior of the millions of actors in an economy as a whole. Despite this contrast between the small and the large, we argue in this chapter that family economics and macroeconomics are in fact intimately related, and that much can be learned from making the role of the family in the macroeconomy more explicit.[a]

There are two different ways in which family economics and macroeconomics intersect. One side of the coin is to focus on questions that originate in family economics, but

[a] The basic point that family economics matters for macroeconomics was made by Becker in his AEA Presidential Address (Becker, 1988). At the time, Becker placed a challenge that inspired a sizeable amount of follow-up research. However, much of the early work at the intersection family economics and macroeconomics was focused on economic growth, whereas we argue in this chapter that family economics is equally relevant for other parts of macroeconomics.

use the methodology of dynamic macroeconomics to answer the questions. For example, macroeconomic models can be adapted to answer questions about how fertility rates, marriage rates, divorce rates, or the assortativeness of mating are determined and how they evolve over time. There is an active and exciting literature that takes this approach, but it is not the focus of this chapter.[b] Rather, our interest here is in the reverse possibility, namely that incorporating family economics into macroeconomics leads to new answers for classic macroeconomic questions. These questions concern, for example, the determination of the level and volatility of employment, the factors shaping the national savings rate, the sources of macroeconomic inequality, and the origins of economic growth.

We choose this path because, so far, it has been less traveled, yet we believe that it holds great promise. This belief is founded on the observation that many of the key decision margins in macroeconomic models, such as labor supply, consumption and saving, human capital investments, and fertility decisions, are made in large part within the family. The details of families then matter for how decisions are made; for example, the organization of families (eg, prevalence of nuclear vs extended families or monogamous vs polygynous marriage) changes the incentives to supply labor, affects motives for saving and acquiring education, and determines possibilities for risk sharing. Yet typical macroeconomic models ignore the family and instead build on representative agent modeling that abstracts from the presence of multiple family members, who may have conflicting interests, who might make separate decisions, and who may split up and form new households.

One might argue that subsuming all family details into one representative household decision maker constitutes a useful abstraction. This would perhaps be the case if the structure and behavior of families were a given constant. However, the structure of the family has changed dramatically over time and is likely to continue to do so in the future. Large changes have occurred in the size and composition of households. Fertility rates have declined, divorce risk has increased (and then decreased), the fraction of single households has grown steadily, and women have entered the labor force in large numbers. Given these trends, the nature of family interactions has changed dramatically over time, and so have the implications of family economics for macroeconomics.

There is a small, but growing, literature that opens the family black box within macro models. The goal of this chapter is to survey this literature, to summarize the main results, and to point to open questions and fruitful avenues for future research. We also aim to introduce macroeconomists to the tools of family economics.

There are multiple ways in which families can be incorporated into macroeconomics. The first generation of macroeconomists who took the family more seriously added home production to business cycle models (eg, Benhabib et al., 1991; Greenwood and Hercowitz, 1991). The insight was that home production cannot be ignored if

[b] See Greenwood et al. (2016b) for an excellent recent survey of that kind of family economics.

the cyclicality of investment and labor supply is to be understood. A large part of investment happens within the household in the form of consumer durables, a large part of time is spent on home production, and both vary over the cycle. The interaction of market time and business investment with these variables that are decided within the family is therefore important for understanding business cycles. In the home production literature, the family is a place of production, but decision making is still modeled in the then-standard way using a representative household with a single utility function.

In this chapter, we take the notion of families a step further. We emphasize that families consist of multiple members and that the interaction between these multiple members is important. We look at both horizontal interactions in the family, ie, between husband and wife, and vertical interactions, ie, between parents and children. Family members may have different interests, resources, and abilities. How potential conflicts of interests within the family are resolved has repercussions for what families do, including macrorelevant decisions on variables such as savings, education, fertility, and labor supply.

This chapter has three parts. We first consider how the family matters for short- and medium-run fluctuations. Second, we turn to economic growth. Third, we consider the role of families for understanding political and institutional change.

Our discussion of short- and medium-run fluctuations uses the US economy as an example to demonstrate how changes in family structure feed back into macroeconomics. We start by documenting how US families have changed in recent decades, including a decline in fertility rates, a large increase in the labor force participation especially of married women, and changes to marriage and divorce. We then analyze how these changes affect the evolution of aggregate labor supply over the business cycle and the determination of the savings rate. With regard to labor supply, we emphasize that couples can provide each other with insurance for labor market risk. For example, a worker may decide to increase labor supply if the worker's spouse becomes unemployed, and couples may make career and occupation choices that minimize the overall labor market risk for the family. The extent to which such insurance channels operate depends on family structure (eg, the fraction of single and married households and divorce rates) and on the relative education levels and labor force participation rates of women and men. We argue that recent changes to family structure have likely changed the volatility of aggregate labor supply and contributed to the "Great Moderation" in economic fluctuations observed between the 1980s and the Great Recession. We also discuss research that suggests that changes in female labor force participation are the main reason behind the recent phenomenon of jobless recoveries. Regarding savings rates, we emphasize how changes to divorce risk affect couples' incentives to save. We conclude this part of the chapter by discussing alternative models of the family and their use within macroeconomics. We argue that there is a need for more detailed dynamic modeling of family decision making, an area where methods widely used in macroeconomics may be fruitfully applied to family economics.

The second part of the chapter focuses on the long run, ie, economic growth. Here education, human capital accumulation, and fertility are the key choices of interest. We start by documenting sharp correlations between measures of family structure and measures of economic development in cross-country data. In a series of simple growth models, we then show how different family dimensions affect the growth rate. The first dimension is the interaction between parents and children, noting that, typically, parents make education decisions for their children. We then add fertility choice and discuss government-imposed fertility restrictions such as the one-child policy in China. Next we move from one-gender to two-gender models by first adding a second person in decision making and then adding a distinction between the two in technology. We use the framework to discuss the implications of the widely observed son preference for economic growth. We conclude the section with a discussion on the importance of nonwestern family structures (such as polygyny) and endogenous marriage.

The third part examines the role of the family in the context of political economy. We argue that the family is an important driver of political and institutional change in the course of development. Throughout the development process, all of today's rich countries (except a few countries whose wealth is built on oil) went through a similar series of reforms. Democracy was introduced, public education was initiated, child labor laws were implemented, the legal position of women was improved, and welfare and social security systems were established. Two important questions are why these reforms were implemented at a particular stage of development, and why many poorer countries failed to introduce similar reforms. We emphasize that most of these reforms concern the nature of the family. Public schooling moved the responsibility of education from the family to the public sphere, and public pension did the same for old age support. Child labor laws put constraints on the power parents have over their children. The introduction of women's rights changed the nature of the interaction between husband and wife. We discuss mechanisms linking the family and political change and the possibility of a two-way feedback between economic development and political reform. We then focus on the political economy of two specific reforms, namely the expansion of women's economic rights and the introduction of child labor laws.[c]

Throughout this chapter, we point out promising directions for future research. In line with the overall theme of the chapter, most of these research directions concern using family economics to generate new answers for questions that originate from macroeconomics. However, we also see a lot of potential for intellectual arbitrage in the opposite direction, namely using tools that are widely used in macroeconomics to build improved models of the family. In particular, a striking difference between the fields is that almost all macroeconomic models are dynamic, whereas in family economics static modeling is still

[c] The political economy of women's rights is addressed in more detail by Doepke et al. (2012).

common. In reality, dynamic considerations should be just as important in family economics as in macroeconomics. For example, if a woman decides to stay at home with her children, she will usually be aware that her absence from the labor market decreases her outside option. Similarly, when a woman and a man decide on whether to have a child, how the child will affect their future interactions will be an important consideration. There is a small literature that documents the importance of dynamics for the family. In particular, Mazzocco (2008) shows empirically that Euler equations hold at the individual but not the household level, and Mazzocco (2007) and Lise and Yamada (2015) provide evidence suggesting that bargaining power within the household evolves over time. To capture such phenomena and to better understand the link between family decisions and aggregate outcomes, more dynamic family bargaining models are needed. Tools that are widely used in macroeconomics, such as dynamic contracting under limited commitment and private information constraints, should prove useful for building such models.

In the following section, we start our analysis by considering the implications of the family for macroeconomic outcomes in the short and the medium run. In Section 3, we investigate the role of the family for economic growth, and Section 4 puts the spotlight on the family as a driver of political change. Section 5 concludes by discussing yet other dimensions in which the family matters for macroeconomics and by providing thoughts on promising directions for future research. Proofs for propositions are contained in the Appendices.

2. THE FAMILY AND THE MACROECONOMY IN THE SHORT AND MEDIUM RUN

Ever since micro-founded modeling became dominant in the 1970s and the 1980s, explicit models of household decision making have been a standard ingredient in macroeconomic models. Depending on the application, the household may face a variety of decisions, such as choosing labor supply, accumulating assets, or investing in human capital. However, within macroeconomics comparatively few attempts have been made to explicitly model families. By modeling families, we mean to account for the fact that households may contain multiple members, who may have different interests, who may make separate decisions, and who may split up in divorce or join others and form new households.

In the following sections, we argue that modeling families can make a big difference in understanding aggregate household behavior in the short and the medium run. We focus on the most basic role of the household sector in macroeconomic models, namely to provide a theory of labor supply and savings.

2.1 The Point of Departure: Representative Households

Traditional macroeconomic models used for business cycle and monetary analysis are populated by an infinitely lived, representative household, who derives utility from consumption and leisure and derives income from supplying labor and accumulating savings. A prototype household problem looks like this (eg, Cooley and Prescott, 1995):

$$\max_{\{c_t,l_t\}} E\left\{\sum_{t=0}^{\infty}\beta^t U(c_t,l_t)\right\} \tag{1}$$

subject to:

$$\begin{aligned} c_t + a_{t+1} &= w_t l_t + (1+r_t)a_t, \\ a_{t+1} &\geq -B, \\ a_0 &= 0, \\ 0 \leq l_t &\leq T. \end{aligned}$$

Here c_t is consumption, l_t is labor supply, w_t and r_t are the wage and the interest rate (taken as given by the household), β is a discount factor that satisfies $0<\beta<1$, and $B>0$ defines a slack borrowing constraint that rules out running a Ponzi scheme. The first-order conditions for the household's maximization problem are:

$$-\frac{U_l(c_t,l_t)}{U_c(c_t,l_t)} = w_t, \tag{2}$$

$$U_c(c_t,l_t) = \beta E\{(1+r_{t+1})U_c(c_{t+1},l_{t+1})\}. \tag{3}$$

Here (2) is the requirement that the marginal rate of substitution between labor and leisure is equal to the wage, and (3) is the intertemporal Euler equation for consumption. Condition (2) pins down average labor supply and the elasticity of labor supply as a function of the relative wage and overall wealth, and (3) determines savings as a function of wealth, interest rates, and expectations over future leisure and consumption.

A representative household based on a problem similar to (1) underlies most of the macroeconomic modeling in the real business cycle literature, the monetary DSGE literature, and many other subfields of macroeconomics. A theory of labor supply and savings that is build on a representative household has a number of limitations, including the obvious one that such a theory has nothing to say about questions that involve heterogeneity and inequality across households. Of course, there is nothing wrong with simplifying assumptions in principle; after all, models are intended to be simplified representations of reality. The limitations of the representative household become a bigger concern, however, when some of the driving forces the model abstracts from are subject to changes over time that substantially alter macroeconomic behavior.

There is already a sizeable literature that extends the representative-household framework in other key dimensions, in particular by accounting for heterogeneity in age (ie, allowing for the life cycle) and heterogeneity in wealth and income.[d] This literature has characterized some of the macroeconomic changes brought about by the changing economic environment in recent decades, such as the large rise in income inequality and returns to education since the 1970s, and the population aging in industrial societies that resulted from rising life expectancy and low fertility. There is much less work on the dimension that this chapter focuses on, namely allowing for the fact that many households have multiple members, ie, accounting for families.

In the following sections, we argue that accounting for families is just as important as the existing extensions of the representative-agent framework. The main reason for this is that families have changed substantially in recent decades; for example, there have been large changes to rates of marriage and divorce, to female labor force participation, and to fertility rates. We start by outlining the main facts of changing families in the United States (to the extent that they are relevant from a macroeconomic perspective), and we then outline channels for how these changes are relevant for determining aggregate labor supply and savings. We note that while there is a lot of existing work documenting and explaining the family trends, there are few papers that focus specifically on the implications of these changes for macroeconomics. In our view, this presents a high-return area for future research, with a lot of low-hanging fruit.

2.2 The Facts: Changing Families in the United States

Throughout the 20th century, the major industrialized countries underwent large changes in the composition and behavior of families. We illustrate this transformation with statistics from the US economy as an example. In the following sections, we explain the relevance of these trends for macroeconomics.

The first transformation concerns changes in fertility over time. Fig. 1 displays the number of children ever born to US women by birth cohort (ie, the horizontal axis is the year in which a mother is born; the corresponding births mostly take place 20–40 years later). As in all industrialized countries, the main trend associated with long-run development is declining fertility. In the case of the United States, fertility fell almost threefold from the cohorts born in the mid-19th century to those born in the late 20th century. The trend was not uniform, however. In the middle of the 20th century there was a phase of rising fertility: the US baby boom. In the course of the baby boom, fertility rose from about two to about three children per woman, and then sharply reversed course to fall back toward two again. These changes have led to large variations

[d] Much of this literature is surveyed by Heathcote et al. (2009) and in the chapter “Macroeconomics and household heterogeneity” by Krueger, Mitman, and Perri (in this volume).

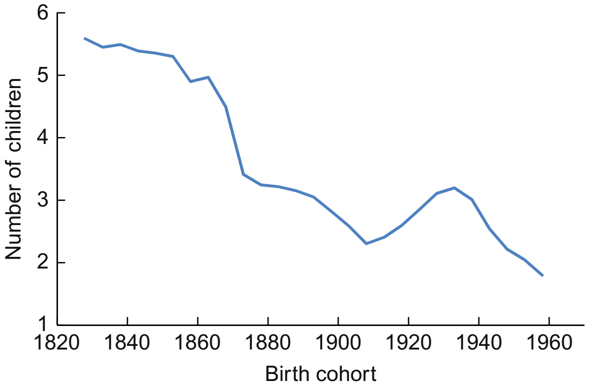


Fig. 1 Children ever born by cohort, United States (ie, average number of children for women born in a given year). *Jones, L.E., Tertilt, M., 2008. An economic history of the relationship between occupation and fertility—U.S. 1826–1960. In: Rupert, P. (Ed.), Frontiers of Family Economics, vol. 1. Emerald Group Publishing Limited, Bingley, UK (Table 1A).*

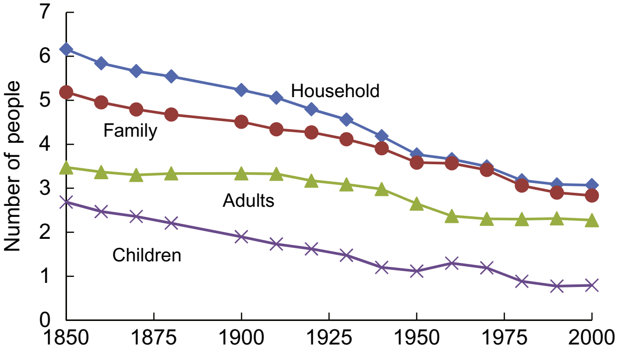


Fig. 2 Household size over time, United States. *Salcedo, A., Schoellmann, T., Tertilt, M., 2012. Families as roommates: changes in US household size from 1850 to 2000. Quant. Econ. 3 (1), 133–175 (Figure 1).*

in cohort sizes, which will affect the macroeconomy for decades to come now that the baby boom cohorts (ie, the babies, not the mothers) are reaching retirement age.

Fig. 2 displays a closely related change: a secular decline in the average size of households. Fertility decline is a main driver of this change; ie, the decline in fertility resulted in fewer children per household and thus a lower household size. However, there are additional factors because the number of adults per household also declined over time. This is in part due to fewer adults within families; ie, a smaller fraction of families include multiple generations of adults, and more families are headed by a single adult. In addition, fewer households include adults who are not related to each other.

Fig. 3 shows that there is not just a decline in the size of households but also a dramatic change in the composition of household types. As recently as 1950, most households

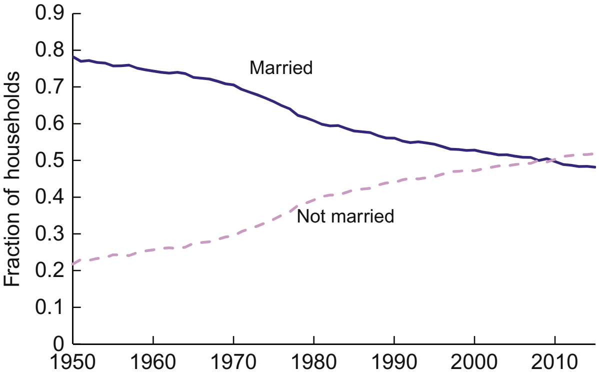


Fig. 3 Proportion of households including a married couple vs all other households over time, United States. *US Census Bureau, Historical Time Series, Current Population Survey, March and Annual Social and Economic Supplements, 2015 and earlier.*

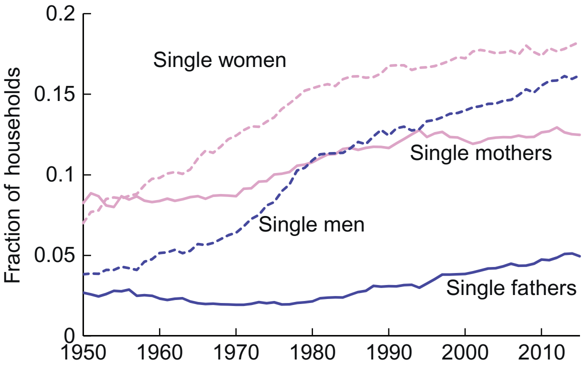


Fig. 4 Nonmarried households by type over time, United States. *US Census Bureau, Historical Time Series, Current Population Survey, March and Annual Social and Economic Supplements, 2015 and earlier.*

(about 80%) included at least one married couple. Now, married-couple households are no longer the majority. Fig. 4 breaks down the nonmarried households into further subcategories, with increases in every subcategory. The figures for single women and single men rise most, indicating primarily lower marriage rates, a higher age at first marriage, and a higher divorce rate. Single mother and single father households have also increased since the 1970s. Fig. 5 looks specifically at the role of marriage and divorce. The figure shows that the decline in the fraction of married women is due in almost equal parts to a rise in the number of never married women and a rise in the number of divorced women. Fig. 6 shows the divorce rate (defined as the number of divorces per 1000 women). Apart from the spike after World War II, the divorce rate was roughly constant from 1940 until the late 1960s and then increased sharply over the course of a decade. It has been relatively constant since the early 1980s, albeit at a much higher level.

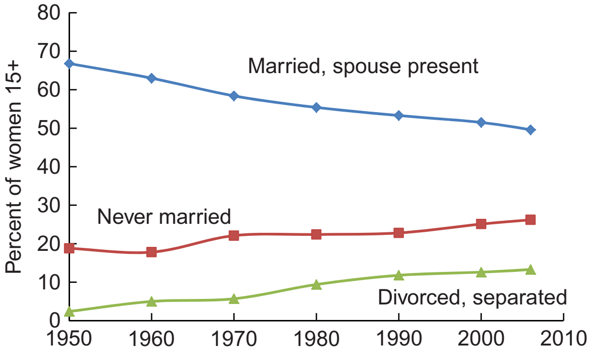


Fig. 5 Breakdown of marital status of women age 15+ over time, United States. *US Census Bureau, Families and Living Arrangements, Current Population Reports.*

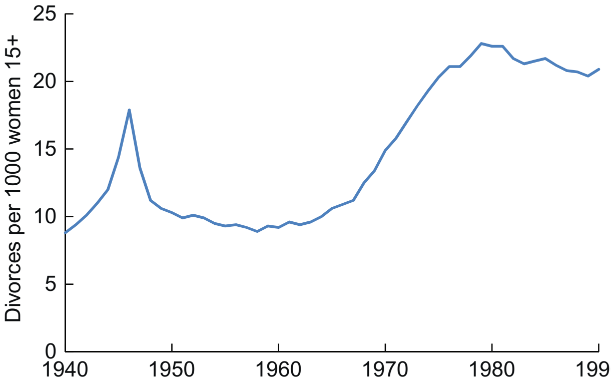


Fig. 6 Divorce rate over time, United States. *US Vital Statistics; Clarke, S.C., 1995. Advance report of final marriage statistics, 1989 and 1990. Monthly Vital Stat. Rep. 43 (12), 3–5.*

Another key trend linking family economics and macroeconomics is the rise in female labor force participation in the postwar era. From the beginning of the 20th century until the 1950s, for married households the single male breadwinner model was the norm. Since then, female labor force participation has risen steadily over a number of decades. As Fig. 7 shows, overall female participation rose from about 30% to more than 60% of the adult population between 1950 and 1990. In the late 1990s, female participation flattened out and declined a little in the current century. Female participation still falls short of male participation, but by a small margin compared to the 1950s. As we will see later (Fig. 13), the rise in female participation is predominantly due to married women. There is also a compositional effect due to the increase in the share of single women coupled with the fact that single women are more likely to work than married women are.

A trend closely related to the rise in female labor market participation is a decline in time spent on home production by women. Figs. 8 and 9 display the average hours men

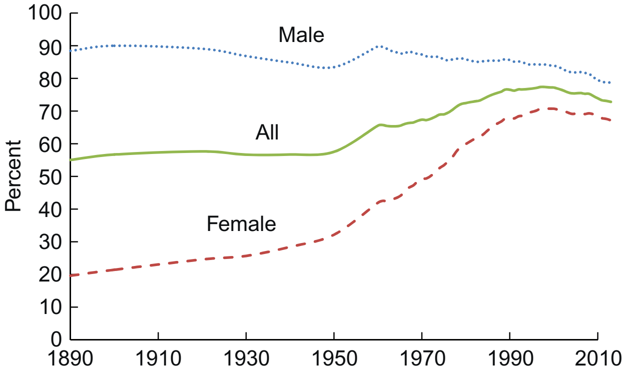


Fig. 7 Labor force participation by gender over time, all persons 16–64 years old, United States. *OECD LFS Sex and Age Indicators and US Department of Commerce, Bureau of the Census, "Historical Statistics of the United States: Colonial Times to 1970," Bicentennial Edition, Part 1, 1975, Tables A119–134 and D29–41.*

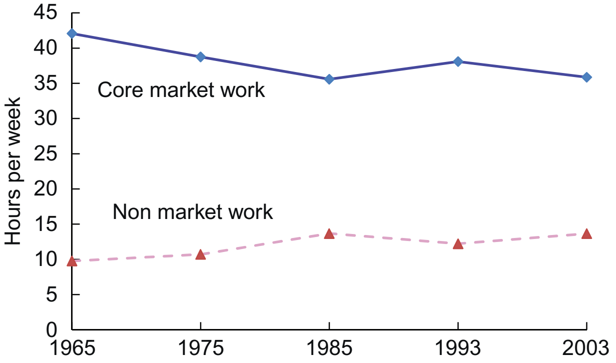


Fig. 8 Men's weekly market vs nonmarket (ie, home) work hours over time, United States. *Aguiar, M., Hurst, E., 2007. Measuring trends in leisure: the allocation of time over five decades. Q. J. Econ. 122 (3), 969–1006 (Table II).*

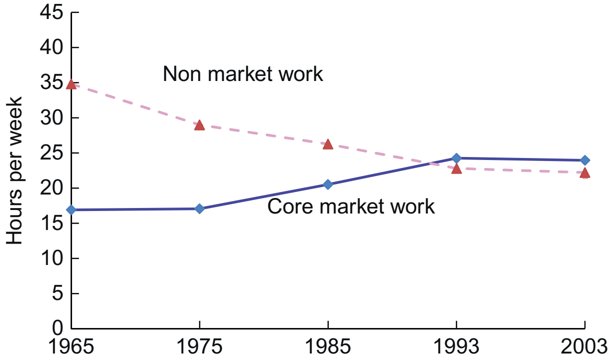


Fig. 9 Women's weekly market vs nonmarket (ie, home) work hours over time, United States. *Aguiar, M., Hurst, E., 2007. Measuring trends in leisure: the allocation of time over five decades. Q. J. Econ. 122 (3), 969–1006 (Table II).*

and women spent per week on market work vs nonmarket work, ie, home production (activities such as child care, cleaning, and preparing food). For men, there is a small decline in market work and an equally small corresponding rise in nonmarket work. For women, in contrast, since 1965 there has been a major transformation in time use: time spent on nonmarket work has dropped sharply while market work has risen, and now exceeds nonmarket time use.

Another closely related fact is the change in relative wages of men and women. Over the course of the 20th century, women have been catching up dramatically in terms of pay. Fig. 10 displays women's median earnings relative to men's earnings. In both cases only full-time, year-round workers are considered. As the figure shows, at the beginning of the 20th century, women earned less than half of what men earned. The ratio increased steadily and had reached 65% by 1955. There was a drop in the late 1960s and 1970s, but from the 1970s onward, the ratio continuously increased again. Today, female relative earnings have reached an all-time high of 80%.

While our focus here is on changes over time in the United States, an interesting pattern in cross-country data is that there is a positive correlation between the fertility rate and the female labor-force participation rate across industrialized countries (Fig. 11). That is, the OECD countries with the highest fertility rates (the United States, France, and the Scandinavian countries) all have relatively high female labor force participation rates, whereas in low fertility countries (such as Italy and Spain) fewer women work in the labor market. The pattern is important because it goes against the relationship between these variables in time-series data: within most countries, the trend through the last 100 years or so has been toward lower fertility and higher female participation. Working in the market and caring for children are alternative uses of women's time. If a single force (say, a rise in

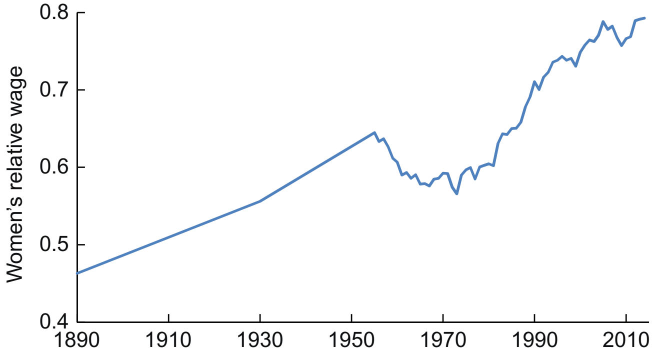


Fig. 10 Gender wage gap: median earnings of full-time, year-round, female workers 15 years and older, relative to men, United States. *US Census Bureau, Historical Income Tables. Numbers for 1890 and 1930 are from Goldin, C., 1990. Understanding the Gender Gap: An Economic History of American Women. Oxford University Press, Oxford (Table 3.2).*

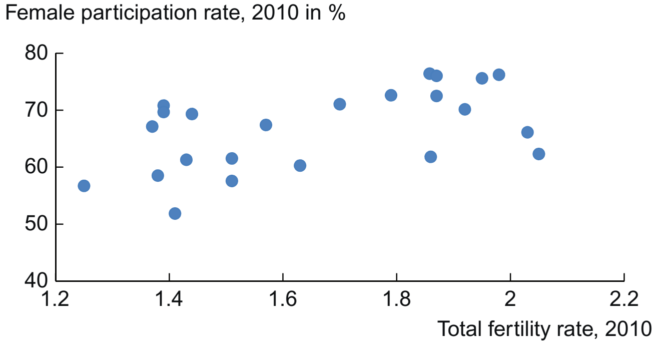


Fig. 11 Fertility vs female labor force participation across European OECD countries. *OECD LFS sex and age indicators and world development indicators.*

relative female wages) was responsible for changes to both labor force participation and fertility, we would expect these variables to always move in opposite directions. The observation in Fig. 11 that, across countries, these variables are positively correlated suggests that such a one-dimensional explanation is at odds with the data and is informative for which kind of theories can explain the family trends described here.

2.3 Explaining the Facts

There is a large literature (spanning family economics, labor economics, development economics, and macroeconomics) that provides explanations for the transformation of the family described above. We keep our discussion of this literature brief, since the goal of this chapter is not explaining these family facts but rather studying their importance for macroeconomic analysis. For a comprehensive survey of the literature on the drivers of changes in the family, we refer the reader to Greenwood et al. (2016b).

The best-known explanations for the historical fertility decline are based on the quantity–quality trade-off together with the idea that returns to education were increasing over time due to technological progress (see also Section 3.3). The more recent fertility decline that followed the baby boom is often connected to the increasing value of female time. The baby boom itself still presents a bit of a puzzle. The conventional wisdom of women catching up on their fertility after the war is clearly not the main driver, as it was young women (not of child-bearing age during the war) who had most children during the baby boom, as Fig. 1 shows. Doepke et al. (2015) suggest that the increase in labor force participation during the war was a major driver for the baby boom. The war generation of women accumulated valuable labor market experience, and after the war these women provided strong competition in the labor market for younger women who lacked that experience. Doepke, Hazan, and Maoz argue that many of these younger women were crowded out of the labor force and decided to start having children earlier instead.[e] Other papers provide a complementary explanation by attributing part of the baby boom to a decline in the cost of child bearing, for example, due to medical progress that made childbirth less risky to mothers (Albanesi and Olivetti, 2014) or improvements in household technology that lowered the time cost of children (Greenwood et al., 2005a).[f]

The causes for the secular increase in female participation have also been widely explored. Some of the explanations focus on the alternative uses of female time and argue that the time required for home production (such as child care, preparing food, or cleaning the home) fell, freeing up time for work. Greenwood et al. (2005b) attribute the reduction in time required for home production to technological progress, and in particular the introduction of time saving appliances. Even if technology had stayed as it was, home production time would have fallen because of the large reduction in the average fertility rate from the baby boom period of the 1950s to the present. Figs. 8 and 9 show that time use data indeed display a large reduction in nonmarket work (ie, home production) for women that closely mirrors the rise in market work. We also observe a small rise in home production for men, suggesting that some of the reduction in female home production arises from substitution within the household. However, the rise in male home production is quantitatively small compared to the decline in female home production. A related theory put forth by Albanesi and Olivetti (2016) is based on technological advances in health. Innovations such as infant formula made it much easier to reconcile work and motherhood and thus were an important contributor to the contemporaneous increase in fertility and female participation between 1930 and 1960.

Another factor contributing to the rise in female participation in the labor market is the decline in the gender wage gap between men and women, as shown in Fig. 10. While some of the overall rise in relative female pay is due to endogenous decisions such as education and the accumulation of work experience, other factors such as the disappearance of marriage bars can be regarded as exogenous driving forces.[g] The gender gap may also have narrowed because of technological change in the market sector that made male and female work more similar. If men have the comparative advantage in brawn and women in brain, then as knowledge becomes more important, female relative wages go up.[h] The role of the

[e] See also Goldin (1990) and Goldin and Olivetti (2013) for other perspectives on the long-run impact of World War II on the female labor market.

[f] Yet another possibility is a link between economic and demographic cycles; Jones and Schoonbroodt (2015) provide a model in which the baby boom arises due to the recovery from the Great Depression in terms of both income and fertility.

[g] Eckstein and Lifshitz (2011) decompose the effect of rising education and the decline in the gender gap conditional on education and find that rising female education accounts for a larger fraction of the increase in female participation.

[h] This idea was first formally modeled by Galor and Weil (1996). See Albanesi and Olivetti (2009) for an alternative theory of how a gender wage gap can arise from private information on work effort and specialization within the household.

declining gender gap in explaining the rise in participation is emphasized by Jones et al. (2015), who also allow for technological improvements in home production, but find them not to be quantitatively important. Attanasio et al. (2008) study the life-cycle labor supply of three cohorts of American women, born in the 1930s, 1940s, and 1950s. Their model allows for a number of potential determinants of labor supply, including changes in the gender wage gap, the number and cost of children, and changes in the returns to labor market experience. They find that for the cohorts considered, both a reduction in the costs of children and a decrease in the gender wage gap need to be allowed to explain the rise in participation. More recent contributions connect the decline in the gender wage gap explicitly with the rise of the service sector (Rendall, 2010; Ngai and Petrongolo, 2014).

Another channel that can affect relative male and female labor supply is endogenous bargaining within the household. In explicit household bargaining models (see Section 2.5), the outside options of the spouses are usually important determinants of bargaining power. Improved labor market opportunities for women (through whichever channel they occur) improve women's outside options and thus should improve women's bargaining power in marriage. Using a quantitative model, Knowles (2013) argues that an endogenous increase in female bargaining power is important in explaining the rise in female labor supply over the 1970–2000 period without implying a (counterfactual) large decline in male labor supply. Eckstein and Lifshitz (2015) estimate a labor supply model in which couples differ in how bargaining takes place (eg, cooperative vs noncooperative bargaining) and find that bargaining has a large impact on female, but not male labor supply.

The link between fertility and employment decisions is likely to have become more important throughout the last few decades. Before the 1960s, in industrialized countries most mothers were not in the labor force, so that for many the employment margin was not operative as far as decisions on additional births were concerned. Today, in the United States and other industrialized countries, most mothers are in the labor force. Hence, having children interacts with employment more directly, through margins such as deciding to work full or part time or the choice between career paths that differ in flexibility for dealing with child care needs. Recently, Adda et al. (2016) have provided a detailed study of the costs of children in terms of mother's careers based on a detailed life cycle model of female employment and fertility matched to German data. They show that the career costs of having children are substantial and that realized and expected fertility can account for a large fraction of the gender wage gap.[i] Based on the same data,

[i] See also Miller (2011) who estimates the career costs of children, using US data on biological fertility shocks as instruments. Guvenen et al. (2014) provide a recent analysis of the gender pay gap at the very top of the income distribution. They argue that a large part of the underrepresentation of women among top earners is due to the "paper floor," ie, a higher likelihood of women dropping out of the top pay percentiles, part of which may be due to fertility decisions.

Bick (2016) provides a quantitative analysis of the importance of the availability of market-based child care for fertility and female labor supply.

As discussed in Section 2.2, if a single force was responsible for both the upward trend in female labor force participation and the downward trend in fertility, we would expect these variables to always move in opposite directions. However, if we look at the cross section of industrialized countries, a positive correlation between female labor force participation and fertility emerges (see Fig. 11).[j] A number of recent studies have developed theories that are consistent with this pattern. The general intuition for these results is that many women now want to have both children and careers. In places where policies (or cultural expectations) are such that mothers can easily combine having children and careers, fertility and female labor force participation will both be high. In contrast, if there are obstacles to combing motherhood with working, many women will choose one or the other, and both fertility and participation will be lower. One of the first papers to formalize this intuition is Da Rocha and Fuster (2006), who focus on differences in labor market frictions across countries. Using a quantitative model, they find that in countries where unemployment risk is high, women both work less and are more likely to postpone births. Similarly, Erosa et al. (2010) find that more generous parental leave policies can increase both fertility and female labor force participation. Another source of variation can be cultural expectations for the roles of mothers and fathers in raising children. Doepke and Kindermann (2015) show that in European countries with exceptionally low fertility rates, women bear a disproportionately large share of the burden of caring for children. In a model of household bargaining over fertility decisions, they show that this leads to many women being opposed to having (additional) children. Hence, once again fertility will be lower, while at the same time many mothers are not able to work due to their child care duties.

The causes behind the decline in marriage, rise in divorce, and increase in single motherhood (as shown in Figs. 3–6) are likely related to the increase in female labor force participation. For a discussion of the causes behind these changes in the family structure, see Greenwood et al. (2016b).

2.4 Changing Families and Aggregate Labor Supply

We now turn to the main focus of this section, namely how changes to the family affect how labor supply and savings are determined in the aggregate. We start with aggregate labor supply, where the role of changes in female labor market behavior takes center stage.

A common thread through the studies of the rise in female participation is that the female participation decision is qualitatively different than the male participation decision. At least in part, this is due to a higher fixed cost of participation for women,

[j] A similar phenomenon has emerged recently in cross-sectional data in the United States. Hazan and Zoabi (2015a) document a U-shaped relationship between female education and fertility.

who often bear the primary responsibility for child care. The different nature of female labor supply suggests that today, aggregate labor supply is determined in a qualitatively different fashion compared to a few decades ago. We now consider a deliberately simplified model to illustrate the main channels through which the joint determination of female and male labor supply within a family affects the macroeconomic properties of labor supply.

2.4.1 Joint Labor Supply in the Family

To focus on the extensive margin, we consider a setting where an individual can either work full time or not at all.[k] The utility function of an individual of gender $g \in \{f,m\}$ is given by:

$$U_g(c_g, l_g) = \log(c_g) - \eta_g l_g,$$

where $l_g \in \{0,1\}$ is labor supply and c_g is consumption.[l] The relative weight of leisure in utility η_g varies in the population. People can live either as singles or as married (or cohabiting) couples. The budget constraint for a single individual is:

$$c_g + \psi l_g = w_g l_g + y_g,$$

where w_g is the wage for gender g, y_g is unearned income (ie, endowment or transfer income), and ψ represents the fixed cost of running a household conditional on working. The implicit assumption is that a person who does not work can replace the cost ψ through costless home production. We assume that ψ is a scalar that satisfies $0 < \psi < \min(w_f, w_m)$. The model is static, but alternatively we can interpret the decision problem as representing the labor-supply decision of a long-lived individual/household with exogenous saving in a given period, in which case y_g represents exogenous net saving/dissaving in the period.

For a married couple, the same fixed cost of running a household applies, but only if both spouses are working.[m] The joint budget constraint for a couple then is:

$$c_f + c_m + \psi \min(l_f, l_m) = w_f l_f + w_m l_m + y, \tag{4}$$

[k] We focus on the extensive margin for tractability. However, similar forces will be effective at the intensive margin as well.

[l] Here we assume that consumption is a private good. Many family models assume that consumption in the family is a public good. We consider pure public goods in Section 3. In reality, there are some private and some public elements in household consumption (see Salcedo et al., 2012 for a detailed analysis of this point).

[m] See Cho and Rogerson (1988) for an early contribution on the implications of this type of fixed cost of participation for the elasticity of labor supply.

where $y = y_f + y_m$. In this setting, the decision problem for a single person is straightforward. Comparing the utility conditional on working vs not working, an individual chooses to work if the condition,

$$\log(w_g + y_g - \psi) - \eta_g \geq \log(y_g),$$

is satisfied, or, equivalently, if the opportunity cost of working is sufficiently low:

$$\eta_g \leq \log\left(\frac{w_g + y_g - \psi}{y_g}\right).$$

For a married couple, we have to take a stand on how the inherent conflict of interest between the spouses given their different preferences is resolved. We assume cooperative bargaining, ie, the household solves a Pareto problem with welfare weights λ_f and λ_m for the wife and the husband, with $\lambda_f + \lambda_m = 1$. The problem solved by a married couple is then given by:

$$\max\{\lambda_f[\log(c_f) - \eta_f l_f] + \lambda_m[\log(c_m) - \eta_m l_m]\} \tag{5}$$

subject to the budget constraint (4). The maximization problem can be solved by using first-order conditions to characterize the consumption allocation conditional on a given pattern of labor supply, and then comparing utilities to determine optimal labor supply. To simplify notation, we focus on the case where husbands always work as long as $w_m > 0$. If the wife does not work, household income is given by $w_m + y$ and the consumption allocation is $c_f = \lambda_f(w_m + y)$, $c_m = \lambda_m(w_m + y)$. If the wife also works, household income net of the participation cost is $w_f + w_m + y - \psi$, and the consumption allocation is $c_f = \lambda_f(w_f + w_m + y - \psi)$, $c_m = \lambda_m(w_f + w_m + y - \psi)$. Denote by $V(l_f, l_m)$ the value of the objective function of the household (5) given labor supply and the optimal conditional consumption allocation. The wife will work if $V(l_f = 1, l_m = 1) \geq V(l_f = 0, l_m = 1)$, which can be written as[n]:

$$\begin{aligned}\log(w_f + w_m + y - \psi) + \lambda_f \log(\lambda_f) + \lambda_m \log(\lambda_m) - \lambda_f \eta_f - \lambda_m \eta_m \\ \geq \log(w_m + y) + \lambda_f \log(\lambda_f) + \lambda_m \log(\lambda_m) - \lambda_m \eta_m.\end{aligned}$$

Simplifying, women will work if and only if:

$$\eta_f \leq \frac{1}{\lambda_f} \log\left(\frac{w_f + w_m + y - \psi}{w_m + y}\right).$$

Hence, women are more likely to work if the participation cost ψ or male wages w_m are low, and if female wages w_f are high. A low bargaining power for women λ_f also translates into higher participation because households then place less value on the wife's leisure.

[n] For now we assume full commitment, ie, people get married before disutilities from working are realized, and they stay together even if being single would provide higher utility.

Note that the assumption of full commitment is important here. If the bargaining power of women is low, women pay the utility cost of working and consume little. Such a woman may prefer not to be a married at all. Later we endogenize the bargaining weights to ensure that participation constraints hold.

We can now consider the implications of the simple model for the variability of labor supply. Consider, first, the own-wage elasticity of labor supply. Consider the case where the only dimension of heterogeneity in the population is in leisure preference η_g, the distribution of which is described by the distribution function $F(\eta_g)$ with continuous marginal density $f(\eta_g) = F'(\eta_g)$. We assume that the density satisfies the assumptions $F(0)=0$, $F'(\eta_g)>0$ for $\eta_g > 0$, $\lim_{\eta_g \to 0} f(\eta_g) = 0$, and $\lim_{\eta_g \to \infty} f(\eta_g) = 0$. That is, all individuals place at least some value on leisure and the distribution thins out at each tail (one example is a log-normal distribution for η_g). For singles of gender g, the fraction working N_g^s given wage w_g is given by:

$$N_g^s = F\left(\log\left(\frac{w_g + y_g - \psi}{y_g}\right)\right).$$

The aggregate wage elasticity of labor supply is then given by:

$$\frac{\partial N_g^s}{\partial w_g}\frac{w_g}{N_g^s} = \frac{w_g}{w_g + y_g - \psi}\frac{F'\left(\log\left(\frac{w_g + y_g - \psi}{y_g}\right)\right)}{F\left(\log\left(\frac{w_g + y_g - \psi}{y_g}\right)\right)}.$$

Note that this elasticity focuses on the extensive margin and hence is different from what is typically measured in the micro data (eg, Pistaferri, 2003 measures only the intensive margin elasticity).[o]

Consider now married couples. By assumption, we focus on the case where married men always work if they are able to. The fraction of married women working is then given by:

$$N_f^m = F\left(\frac{1}{\lambda_f}\log\left(\frac{w_f + w_m + y - \psi}{w_m + y}\right)\right)$$

and the elasticity of their labor supply is:

$$\frac{\partial N_f^m}{\partial w_f}\frac{w_f}{N_f^m} = \frac{w_f}{\lambda_f(w_f + w_m + y - \psi)}\frac{F'\left(\frac{1}{\lambda_f}\log\left(\frac{w_f + w_m + y - \psi}{w_m + y}\right)\right)}{F\left(\frac{1}{\lambda_f}\log\left(\frac{w_f + w_m + y - \psi}{w_m + y}\right)\right)}.$$

[o] Recent contributions that explicitly consider the extensive margin include Chetty et al. (2011, 2012) and Attanasio et al. (2015).

The relative size of single and married women's labor supply elasticity cannot be unambiguously signed, because this depends on the shape of the distribution function F and the size of unearned income. However, married women's labor supply will be more elastic than the labor supply of single women if unearned income y_f is sufficiency small:

Proposition 1 (Labor Supply Elasticity of Single vs Married Women) *If unearned income* $\mathrm{y_f}$ *is sufficiently small, married women's labor supply elasticity is higher than that of unmarried women.*

Intuitively, if unearned income is small, singles have to work if they want to consume, whereas a married woman can rely in part on her spouse's income. This result is in line with the empirical observation that married women's labor supply is much more elastic than that of married men or single women at the microlevel (see, eg, the survey by Blundell and MaCurdy, 1999). Of course, if the labor supply of married men were endogenized, they would also have more scope for variability in supply compared to single men. In practice, as long as the gender wage gap was sizeable and social expectations were that women do more child care and home work, the assumption that men are the default earners was broadly realistic. But as gender roles have become more equalized over time, we can expect the labor supply behavior of men and women to converge also.

Ultimately we would like to assess the implications of changes in the family for the behavior of aggregate labor supply. The results so far may seem to suggest that a higher proportion of married households should make aggregate labor supply more variable. However, so far we have only considered the own wage elasticity of female labor supply. Another important dimension of the family is the possibility of insurance within the family. Specifically, if in a marriage the working husband experiences a negative shock such as a layoff, the wife may be able to offer insurance by starting to work. Hence, in the aggregate, the variable labor supply of married women may dampen fluctuations in total labor supply, by offsetting shocks experienced by men.[p]

To analyze the possibility of insurance within the family, consider an extension of the environment with unemployment shocks. With probability u, a given individual is unable to work, or equivalently, the potential wage is zero. The realization of the shock is independent across spouses. We can now consider how aggregate labor supply reacts to changes in u, where an increase in u can represent a recession.

As before, we start by considering singles. Their aggregate labor supply is:

$$N_g^s = (1-u)F\left(\log\left(\frac{w_g + y_g - \psi}{y_g}\right)\right).$$

[p] An early study of this insurance channel is provided by Attanasio et al. (2005).

For singles, the elasticity of labor supply with respect to the probability of employment $1 - u$ is unity:

$$\frac{\partial N_g^s}{\partial(1-u)}\frac{1-u}{N_g^s}=1.$$

For married couples, labor supply is driven by two different thresholds for the wife's leisure preference, depending on whether the husband is working or not. Denote these thresholds by:

$$\hat{\eta}_e=\frac{1}{\lambda_f}\log\left(\frac{w_f+w_m+\gamma-\psi}{w_m+\gamma}\right),$$
$$\hat{\eta}_u=\frac{1}{\lambda_f}\log\left(\frac{w_f+\gamma}{\gamma}\right).$$

The average labor supply per married couple is then:

$$N^m=(1-u)(1+(1-u)F(\hat{\eta}_e))+u(1-u)F(\hat{\eta}_u).$$

Here the first term corresponds to employed husbands, and the second term corresponds to unemployed husbands. Wives of unemployed husbands work with a strictly higher probability than wives of employed husbands, because the cost ψ does not have to be paid (a substitution effect) and overall income is lower (an income effect working in the same direction). The derivative of labor supply with respect to $1 - u$ for the married couples is:

$$\frac{\partial N^m}{\partial(1-u)}=(1+(1-u)F(\hat{\eta}_e))+(1-u)F(\hat{\eta}_e)+uF(\hat{\eta}_u)-(1-u)F(\hat{\eta}_u),$$

$$\frac{\partial N^m}{\partial(1-u)}=(1+2(1-u)F(\hat{\eta}_e))-(1-2u)F(\hat{\eta}_u).$$

The elasticity of married labor supply with respect to $1 - u$ is then:

$$\frac{\partial N^m}{\partial(1-u)}\frac{1-u}{N^m}=\frac{1+2(1-u)F(\hat{\eta}_e)-(1-2u)F(\hat{\eta}_u)}{1+(1-u)F(\hat{\eta}_e)+uF(\hat{\eta}_u)}.$$

If it were the case that $F(\hat{\eta}_u)=F(\hat{\eta}_e)$, the expression once again would yield an elasticity of unity as for the singles. However, in fact we have $\hat{\eta}_u>\hat{\eta}_e$ and hence $F(\hat{\eta}_u)>F(\hat{\eta}_e)$, so that the elasticity of labor supply by married couples is strictly smaller than one. Intuitively, there is a fraction of women (given by $F(\hat{\eta}_u)-F(\hat{\eta}_e)$) who do not work if their husband is working, but choose to enter the labor force if the husband is unemployed. Hence, there is insurance in the family that dampens fluctuations in aggregate employment. Even though married female labor supply is more elastic at the microlevel, it contributes to a dampening of the volatility of aggregate labor supply due to this intrafamily insurance effect.[q]

In the data, married female employment rose massively in the second half of the 20th century (see Fig. 7), and there were also large shifts in the composition of household types (see Figs. 3 and 4). The model suggests that these changes should affect the volatility of aggregate labor supply. The following proposition summarizes the main results.

Proposition 2 (Family Determinants of Volatility of Aggregate Labor Supply) *Consider a population of measure one consisting of* M *married households (with two members each) and 1–2*M *single households. We then have:*

1. *The elasticity of aggregate labor supply* N *with respect to* $1-\mathrm{u}$ *(the fraction of workers not affected by the unemployment shock) is equal to one if the fraction of married people is* $\mathrm{M}=0$ *and decreases with* M *for* $\mathrm{M}>0$.
2. *For a fixed* $\mathrm{M}>0$*, the elasticity of aggregate labor supply* N *with respect to* $1-\mathrm{u}$ *is strictly smaller than one, but approaches one when* $\mathrm{w_f}$ *converges to zero or to infinity.*

The first premise suggests that the large shifts in the composition of households in the past few decades may have had a marked effect on the response of aggregate labor supply to shocks. The second premise suggests that, in addition, the increase in female labor supply should also affect the behavior of aggregate labor supply, albeit in a nonmonotone way. Regarding the married households, what is at stake is the potential for insurance within the family. When conditions are such that women do not work even if their husbands are unemployed (captured here by the case of a female wage close to zero), there is no potential for insurance, and hence the labor supply of married households will be just as elastic as that of single households. Conversely, when conditions are such that all women work regardless of the employment status of their husbands (captured by the case of the female wages approaching infinity), there is no potential for insurance either. Insurance does play an important role when there is a sizeable group of women who do not work if their husbands are employed, but are willing to enter the market when the husband loses his job. Hence, the mechanism would predict the greatest role for insurance at a time when the rise in female employment is well underway, but still not close to being completed.

Fig. 12 displays how the elasticity of total labor supply by married households with respect to the unemployment shock depends on relative female wages in a computed example.[r] The male wage is normalized to one, and the source of variation is the relative

[q] There is an active debate in the literature on how micro- and macroestimates of labor supply elasticities can be reconciled (see Chetty et al., 2011, 2012; Keane and Rogerson, 2012 for recent contributions).

[r] Parameter values: $w_m=1$, $\gamma=0.1$, $\psi=0.1$, and $\lambda_f=0.5$. The distribution of leisure preferences is log-normal with $\mu=0.5$ and $\sigma=1$, and the elasticity of labor supply is evaluated at an unemployment rate of $u=0.1$.

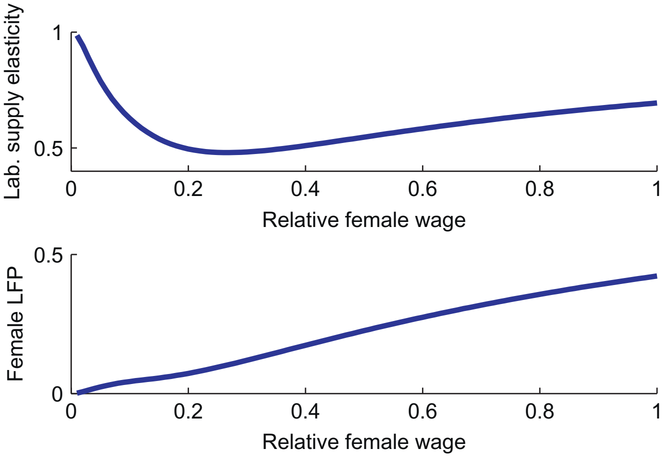


Fig. 12 Aggregate labor supply elasticity and female labor force participation (LFP) as a function of relative female wage in labor supply model.

female wage. The lower panel shows female labor supply as a function of the relative female wage. Not surprisingly, at a zero female wage, female labor supply is zero as well. However, even with very low wages some women work, namely those whose husbands are unable to work and who have a low leisure preference. The upper panel shows that this implies that the aggregate elasticity is U-shaped in relative female wages. In light of the observed decline in the gender wage gap and the increase in female labor force participation in US data (see Figs. 7 and 10), the findings suggest that the aggregate labor supply elasticity should have changed substantially in recent decades.

2.4.2 Endogenous Bargaining

The analysis of married couples' decisions has been carried out so far under the assumption of exogenous bargaining weights and full commitment. As mentioned above, if female bargaining power is low and female wages are high, women are likely to work a lot and consume little, and hence such women may prefer not to be married at all. Without full commitment, ie, if women were allowed to leave such a marriage, efficient bargaining subject to the limited commitment constraint would dictate that bargaining weights adjust to ensure that married women get at least as much utility as they would if they were single. Adjusting bargaining weights in this way is possible as long as the surplus from marriage is positive, which is guaranteed in our setting as long as $\psi > 0$ (married couples economize on the cost of running a household).[s]

[s] Other reasons for a positive marital surplus include consumption being a public good (see Section 3) and a utility benefit from being married (see Section 2.5).

Now consider how bargaining weights would adjust to changing wages w_g in this setting.[t] The utility of a single female is the maximum value between working and not working as a single:

$$U_f^s = \max\{\log(w_f + y_f - \psi) - \eta_f, \log(y_f)\}.$$

Assume that w_f is high enough (or y_f low enough) so that as a single, she always prefers to work. Comparing her utility as a single with that when married, she will prefer to be single if:

$$w_f + y_f > \frac{\lambda_f}{1-\lambda_f}(w_m + y_m) + \psi.$$

This condition will hold, for example, when her wages are high or her bargaining power is low. In such a case, the bargaining power in marriage should adjust to guarantee her at least her reservation (ie, single) utility:

$$\lambda_f = \frac{w_f + y_f - \psi}{w_m + y_m + w_f + y_f - \psi}.$$

Of course, any λ_f higher than the expression above would also guarantee that her participation constraint is satisfied.

We can use this logic to assess what would happen in a dynamic model with shocks to wages and participation cost. Suppose the couple starts out with a large marital surplus and bargaining weights such that neither participation constraint is binding. Suppose now that her wage increases unexpectedly such that, holding λ_f constant, her participation constraint would be violated. In response, her bargaining weight will increase. Similarly, a fall in the participation cost ψ may also lead to a tightening of the participation constraint and hence a shift in bargaining weights.[u] Bargaining positions will also be affected by changes in unearned income such as lottery winnings or an inheritance.

Now consider how such changes in bargaining weights affect the elasticity of labor supply. Qualitatively, the effects described in Propositions 1 and 2 rely on the possibility of insurance within the family and do not depend on the assumption of fixed bargaining weights. However, endogenous bargaining may well matter for the quantitative size of the effects. Both Knowles (2013) and Voena (2015) examine this issue, although their

[t] The model is static of course so there is no adjustment over time. Rather, one should think of bargaining weights differing across couples in an economy with heterogeneity in relative wages. However, the basic logic would carry over to a dynamic model with limited commitment where similar forces would lead to adjustments in the bargaining weights over time, see Mazzocco (2007) and Voena (2015).

[u] Since a decline in ψ affects both the male and female participation constraint, the direction of the change will depend on the details and in particular the status quo bargaining weight. Suppose her constraint is exactly binding before the shock lowering ψ is realized. Then, clearly, since he is currently reaping the entire surplus, her weight will have to go up to ensure continued participation in marriage by the female.

analyses are concerned with longer-term changes rather than with the business cycle. Nevertheless, the forces they identify should also be active at the business cycle frequency. If a higher wage increases bargaining power, it also increases the weight in the bargaining process on the leisure of the spouse who is receiving the raise. This effect lowers the response of labor supply to wage changes. Indeed, Knowles (2013) argues that the overall response of aggregate labor supply to the increase in female wages is dampened because of shifts in bargaining power. Whether such shifts in bargaining power also dampen aggregate labor volatility is less clear, as the opposite effect will apply to the other spouse. We view this as a fruitful area for future research.

2.4.3 Linking Changes in the US Labor Market to Family Labor Supply

We now relate the theoretical channels linking the family to variations in aggregate labor supply outlined above to empirical evidence on fluctuations in employment and output in the United States. We are interested in how the variability of aggregate labor supply varies between men and women and single and married individuals, and how these factors changed over time. Our analysis is based on annual data from the Current Population Survey (CPS) for the years 1962–2014. We focus on average weekly hours worked per person for the population aged 25–65.[v] Fig. 13 shows how this measure of labor supply

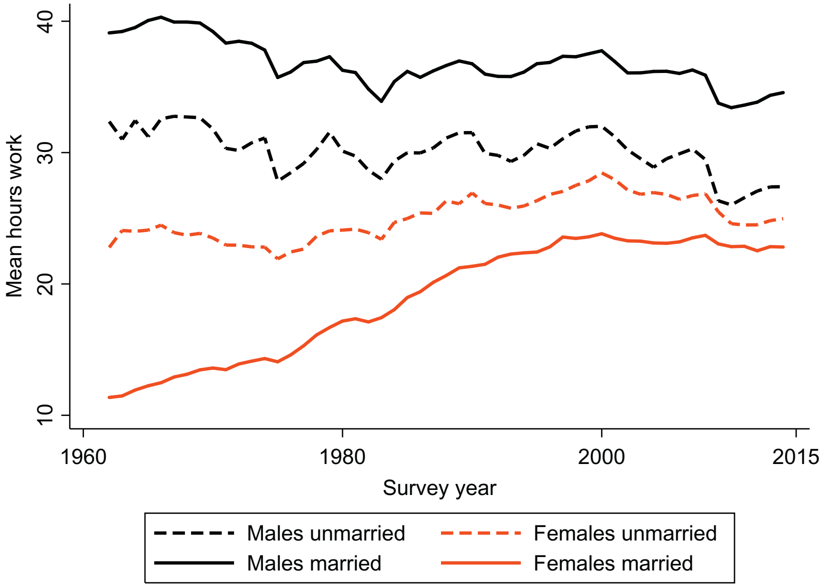


Fig. 13 Average weekly work hours by gender and marital status over time, United States. *Current Population Survey, March and Annual Social and Economic Supplements, 1962–2014.*

[v] The sample includes self-employed individuals.

evolves over time by gender and marital status. The sharp upward trend in married women's labor supply from the 1960s to the 1990s is apparent, as well as the comparatively larger drop in male labor supply since the Great Recession of 2008.

To focus on fluctuations at the business cycle frequently, we compute the cyclical component as the residual after subtracting a Hodrick–Prescott trend from the logarithm of each series (with a smoothing parameter of 6.25). The cyclical component of labor supply by gender and marital status is displayed in Fig. 14. It is immediately apparent that aggregate male labor supply is more volatile than aggregate female labor supply. Single men experience the largest fluctuations in labor supply over the cycle, whereas the smallest fluctuations are observed for married women.

The large differences in the volatility of female and male labor supply together with the large increase in female labor supply suggest that family trends may have had repercussions for the cyclical properties of aggregate labor supply over the observed period. To examine this possibility more formally, Table 1 provides detailed information on fluctuations in aggregate labor supply in the United States in relation to gender and marital status. In the table, the total volatility of a given series is the percentage standard deviation of the cyclical component of average labor supply per person in the group. Cyclical volatility is the percentage standard deviation of the predicted value from a regression of the cyclical component of employment in each group on the cyclical component of real GDP per capita (also computed using the HP filter). Cyclical volatility captures the

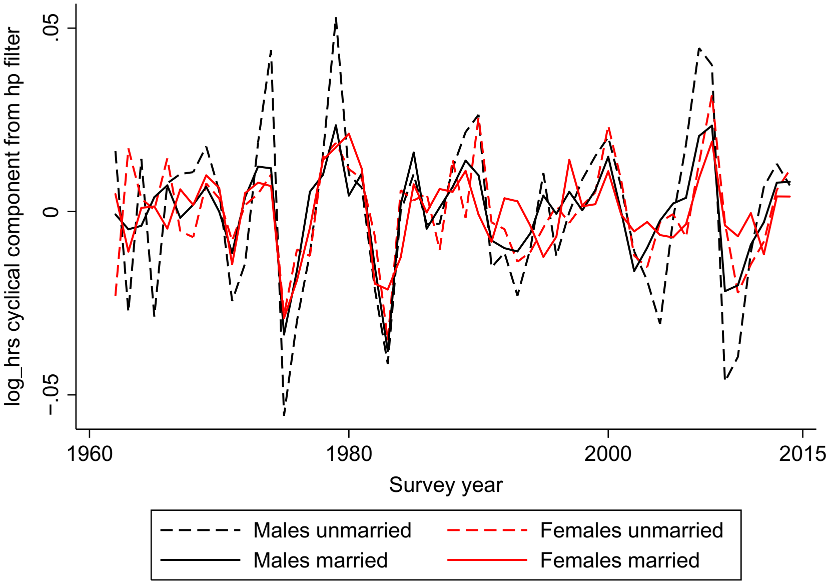


Fig. 14 Cyclical component of average weekly work hours by gender and marital status over time, United States (cyclical component is deviation from Hodrick–Prescott trend, smoothing parameter 6.25). *Current Population Survey, March and Annual Social and Economic Supplements, 1962–2014.*

Table 1 Volatility of hours worked in the United States, by gender and marital status

	All			Married		Single	
	Total	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**
				1962–2014			
Total volatility	1.25	1.04	1.46	1.04	1.25	1.33	2.33
Cyclical volatility	0.99	0.72	1.18	0.67	1.01	0.74	1.68
Hours share		38.09	61.91	23.90	47.71	14.19	14.20
Volatility share		27.22	72.78	16.20	48.98	10.64	24.17
				1962–88			
Total volatility	1.35	1.19	1.48	1.26	1.36	1.37	2.44
Cyclical volatility	1.08	0.87	1.19	0.87	1.09	0.79	1.65
Hours share		33.71	66.29	21.99	55.29	11.72	11.00
Volatility share		27.14	72.86	18.02	56.29	8.67	17.02
				1989–2014			
Total volatility	1.15	0.87	1.47	0.79	1.16	1.30	2.25
Cyclical volatility	0.91	0.51	1.23	0.38	0.95	0.70	1.82
Hours share		42.64	57.36	25.89	39.83	16.75	17.53
Volatility share		23.68	76.32	10.80	41.51	12.88	34.81

Notes: All data from Current Population Survey, March and Annual Social and Economic Supplements, 1962–2014. Total volatility is the percentage standard deviation of the Hodrick-Prescott residual of average labor supply per person in each group. Cyclical volatility is the percentage deviation of the predicted value of a regression of the HP-residual on the HP-residual of GDP per capita. Hours share is the share of each component in total hours. Volatility share is share of each group in the cyclical volatility of total hours.

component of employment volatility that is related to aggregate economic fluctuations. The hours share and volatility share break down the contribution of each component to aggregate hours and to the cyclical volatility of aggregate labor supply.[w]

The first column displays the volatility of aggregate labor supply (women and men combined), and the next two columns break down labor supply between women and men. Over the entire sample, women's labor supply is less volatile than men's labor supply. Moreover, for women cyclical volatility is a smaller fraction of total volatility compared to men; ie, less of the variation in female labor supply is related to aggregate economic fluctuations. As a consequence, even though over the entire sample women contribute close to 40% of total hours, they account for less than 30% of volatility in aggregate labor supply.

A key observation is that female labor supply is less variable than male labor supply in the aggregate, even though at the microlevel women have a much higher labor supply

[w] The computation of cyclical volatility and hours and volatility shares follows the methodology used by Jaimovich and Siu (2009) and Jaimovich et al. (2013) to characterize the contributions of the young and the old to aggregate fluctuations.

elasticity than men. These facts can be reconciled if some of the microvariability in female labor supply is due to adjustments that move in the opposite direction of aggregate changes, such as women increasing labor supply in a recession. We would expect such movements to be especially likely to arise among married households, where the spouses can provide each other with some insurance. To evaluate this possibility, in the further columns the fluctuations in labor supply are further broken down into married vs single individuals. Consistent with a role for insurance, we see that, for both women and men, fluctuations are much smaller for the married than for the single individuals.

At first sight, the lower variability of married labor supply may appear to contradict Proposition 1, which states that married women should have a higher wage elasticity of labor supply than single women. However, Table 1 captures macroeconomic fluctuations rather than microelasticities, and we would expect married women to have lower aggregate volatility precisely if their higher microelasticity arises from a fraction of married women adjusting their labor supply countercyclically in response to changes in their husbands' earnings.[x]

Some of the lower variability of female labor supply is related to the fact that a larger share of women is employed in the service sector, which is less cyclical than the manufacturing sector where men dominate. However, when we compare workers employed in manufacturing and services, we find that within each sector women experience a lower cyclical volatility than men. Moreover, the link to the sector of employment does not contradict a role for insurance within the family, because the choice of sector (and also occupation) is endogenous and may be made in part precisely to offset risk encountered by a worker's spouse.[y]

The theoretical mechanisms outlined in the previous section suggest that the aggregate elasticity of labor supply should respond to changes in female labor force participation. To explore this possibility, the remainder of Table 1 compares fluctuations during the first half of our sample (1962–88), when female labor supply was rising quickly from an initially low level, to the period 1989–2014, when female labor supply had reached a

[x] A second factor driving the higher aggregate volatility of single labor supply (which is not captured in the model) is that singles tend to be younger than married people, and the young generally have more variable labor supply for other reasons (such as a more important education margin, see Jaimovich et al., 2013). We can control for the effect of age by considering narrower age brackets. For example, among people aged 25–30, the total volatility of the labor supply of married and single women is about the same.

[y] The special role of the service sector in the rise of female employment is analyzed by Buera et al. (2013), Ngai and Petrongolo (2014), and Rendall (2015). Olivetti and Petrongolo (2016) provide an empirical study of the role of industry structure for trends in female employment, working hours, and relative wages in a cross-section of developed economies, and argue that the rise of the service sector accounts for at least half of the long-term variation in female hours. Albanesi and Şahin (2013) study the role of industry composition for male-female differences in cyclical fluctuations in employment in the United States, and show that that industry composition was not important for pre-1990 recessions, but mattered more once female participation flattened out in the 1990s.

higher plateau. The most important observation here is that whereas the volatility of male labor supply is essentially unchanged, the volatility of female labor supply has substantially decreased, and particularly so the cyclical volatility. The breakdown by marital status shows that this change is driven primarily by married women. Married women already have a low total volatility of about 0.8% in the second half of the sample, and less than half of this total volatility is accounted for by cyclical volatility. These numbers suggest, as predicted by the simple theoretical model in the previous section, that the rise in female labor force participation had a substantial dampening effect on the volatility of total labor supply. In contrast, there are no substantial changes in the cyclical volatility of the labor supply of singles, with a small decrease in volatility for single women and a small increase for single men.

The overall result of the changes is that at the same time women increased their share of total hours (from 34% to 43%), they accounted for a smaller share of total volatility (24% in 1989–2014 compared to 27% in 1962–88). As a consequence, the total volatility and cyclical volatility of aggregate labor supply fell substantially (see first column), even though the volatility of male labor supply slightly increased over the period. Hence, the rise in female participation dampened the volatility of aggregate labor supply over the cycle, in line with Proposition 2 and the declining portion of the aggregate elasticity in Fig. 12. Rising female participation may thus have been one of the driving forces of the "Great Moderation" in US aggregate fluctuations observed from the mid-1980s to the onset of the Great Recession in 2007.[z] Of course, the Great Recession appears to have brought the Great Moderation to an end, and hence one may wonder whether this dampening effect is still operative. The data suggest that female labor supply continues to partially offset aggregate fluctuations. A division of the sample into three periods shows that the most recent era displays the lowest volatility of female labor supply, with a cyclical volatility for married women of only 0.37%. The dampening role of married women's labor supply was particularly pronounced during the Great Recession itself. From 2007 to 2010, the average labor supply by married men declined by more than 8%, whereas the decrease was less than 3% for married women.

If the trend toward more gender equality continues, according to Proposition 2 the volatility of female and male labor supply should ultimately become more similar again (see also Fig. 12). In part, as married women become even more strongly attached to the labor force (eg, in the sense of more women being the main breadwinner for their family), their labor supply will become less elastic (this can already be observed at the

[z] See Galí and Gambetti (2009) for an overview of the discussion on the Great Moderation, and Jaimovich and Siu (2009) for an explanation that focuses on changes in the age composition of the labor force. Mennuni (2015) also considers the impact of demographic trends on the Great Moderation (although without considering the distinction of single and married individuals), and finds that demographics (including the rise in female participation) account for about 20% of the Great Moderation in the United States.

microlevel, eg, Heim, 2007). Conversely, men will become more able to rely on their wives' incomes, which should make their labor supply more elastic at the microlevel but also less cyclical in the aggregate. Hence, family trends will continue to play a role in shaping aggregate fluctuations.

2.4.4 Jobless Recoveries

A phenomenon that has received a lot of attention recently in business cycle research is the so-called jobless recoveries. This term refers to a recent change in the employment response to recessions in the United States. Before the 1990s, most postwar recessions were characterized by a strong rise in employment from the trough of the recession. In contrast, since the 1990s the increase in employment during the recovery has been anemic.

A variety of explanations have been proposed for the recent jobless recoveries, including structural change (Groshen and Potter, 2003), an increase in "job polarization" (the disappearance of jobs in the middle of the skill distribution in recessions; see Jaimovich and Siu, 2014), and fixed costs of labor adjustment (Bachmann, 2012). However, in recent work, Albanesi (2014) makes a strong case for jobless recoveries at least in part being due to changes within families, and more specifically to changes in female labor force participation. In a nutshell, Albanesi argues that employment differed in the aftermath of pre-1990 and post-1990 recessions because the earlier recessions took place in the context of a strong secular upward trend in female labor force participation, whereas the more recent ones did not. As Fig. 7 shows, female labor force participation in the United States followed a sharp upward trend, but participation leveled out after about 1990, and even declined somewhat in the last 15 years.

Table 2 summarizes the employment response to recent recessions and breaks them down by male vs female employment. Each entry in the table is a percentage change in the employment to population ratio (E/P) in the 4 years following the trough of the recession. The first column reproduces the basic fact of jobless recoveries. In the pre-1990 recessions, employment had fully recovered (and even increased a little) 4 years after the downturn, whereas for the post-1990 recessions the E/P ratio is on average close to 3% lower at that point of the recovery (1.35% if the Great Recession is excluded). Hence, it appears that recoveries after 1990 are qualitatively different from earlier recoveries. The next two columns break down the overall employment change into changes in the E/P ratio for women and men. The main message from these data is that, statistically, the jobless recoveries are due to changes in the behavior of female but not male employment. For men, recoveries have been "jobless" even before 1990, in the sense that the E/P ratio is down by 2.62% on average 4 years after the trough. The decline in E/P after 1990 is of a similar order of magnitude, and in fact a little smaller when the Great Recession is excluded. In contrast, we see a dramatic change for women. In the pre-1990 recessions, the female E/P ratio recovers strongly after each recession, with an average increase of

Table 2 Jobless recoveries: change in employment/population ratio in 4 years after peak in unemployment rate, in percentage points, by gender (includes three pre-1990 and tree post-1990 recessions)

	Change in E/P		
Period	**Total**	**Men**	**Women**
Pre-1990	0.65	−2.62	5.85
Post-1990	−2.78	−3.94	−1.41
Post-1990, excl. Great Recession	−1.35	−2.47	−0.07

Notes: Pre-1990 recessions include the 1969, 1973, and 1981 recessions. Post-1990 recessions include the 1990, 2001, and 2007 recessions.

close to 6% after 4 years. In contrast, in the post-1990 downturns female employment declines and now follows a pattern similar to that of male employment.

Table 2 suggests that, in a statistical sense, the change in the trend in female labor supply is responsible for jobless recoveries. Specifically, for men recoveries have always been jobless, whereas for women, before 1990 recession-related job losses were quickly made up by the secular upward trend in female participation. Of course, the empirical findings alone are not conclusive evidence in favor of such an explanation. For example, it is conceivable that if in the pre-1990s recessions female employment had risen more slowly, male employment would have suffered fewer losses. To fully evaluate the role of the changing trend in female labor supply for explaining jobless recoveries, one needs to spell out an economic model. Albanesi (2014) considers a model in which the increase in female participation is driven by gender-biased technological change, ie, tasks at which women have a comparative advantage become more important compared to those that favor men (such as those relying on physical strength). Albanesi shows that the model can reproduce both the long-run trend in female participation and the occurrence of jobless recoveries after female employment levels out.

2.4.5 Additional Notes on Related Literature

Whereas few papers explicitly consider how family trends change business cycle dynamics, there is a larger literature that incorporates at least some of the features of the family labor supply model described above into business cycle research. An early example is the literature on home production in macroeconomics (see Greenwood et al., 1995 for an early overview of this work). The first models did not explicitly distinguish between male and female labor supply, but by incorporating the possibility of working in the home (on child care, food production, and so on), the literature took implicit account of the different nature of female labor supply. Benhabib et al. (1991) is an early contribution focusing on the importance of home production for explaining business cycle facts. In their model, households derive utility from home and market consumption and supply both home and market hours. They find that the model with home production is much better at matching various volatilities and correlations over the business cycle than standard macro models. Closely related arguments are made by Greenwood and Hercowitz (1991) and Ríos-Rull (1993).

The role of family labor supply in the context of search models of the labor market has been analyzed by Guler et al. (2012). Spouses who are both in the labor force can provide each other insurance in the case of unemployment. They find that the possibility of insurance lowers the search effort of unemployed workers and also provides higher welfare compared to a setting where all workers are singles. Ortigueira and Siassi (2013) use a quantitative model to assess the importance of risk sharing within the family, and find that insurance through spousal labor supply is particularly important for wealth-poor households who lack access to other insurance mechanisms.

Family labor supply also plays a central role in a recent macroeconomic literature on the effects of tax reform. Using a quantitative life-cycle model with single and married households calibrated to US data, Guner et al. (2012a) explore the economic consequences of revenue-neutral tax reforms that adopt either a flat income tax or separate taxation of married couples (ie, separate filing). In either case, the reform generates a large increase in labor supply, which is mostly driven by married women (see also Guner et al., 2012b). Guner et al. (2014) extend this work to consider the effects of child care subsidies. They find that such subsidies have large effects on female labor supply, in particular at the bottom of the skill distribution. Bick and Fuchs-Schündeln (2014) document differences in labor supply of married couples across 18 OECD countries, and find that variation in tax systems (in particular joint vs separate taxation) can account for most of the differences.[aa]

In the labor literature, the phenomenon of a wife entering the labor market in response to her husband's unemployment that partly underlies Proposition 2 is known as the "added worker effect" (Lundberg, 1985). Empirical studies using data from the early 1980s or earlier have generally only found weak evidence in favor of the added worker effect. Using CPS data over a long time period, Juhn and Potter (2007) find evidence in support of the added worker effect but also argue that it has diminished in strength recently, in part because assortative mating has led to a higher intrahousehold correlation of the labor market shocks faced by wives and husbands.

The large differences in the cyclical volatility of the labor supply of single and married women and men documented above suggest that insurance within the family goes beyond a narrow added worker affect (which specifically concerns wives entering the labor force *after* their husbands become unemployed). Other forms of insurance include entering employment already in response to higher unemployment risk for the spouse (rather than the actual realization of unemployment, when entering the labor force quickly may be difficult), and adjustments on the intensive margin when both spouses

[aa] See also Chade and Ventura (2005) for an analysis of the welfare consequences of different tax treatments for married couples.

are in the labor force. Hyslop (2001) and Shore (2010, 2015) provide evidence in favor of a more general sharing of labor market risk in terms of the correlation of earnings within couples. Using a structural model of life cycle decisions, Blundell et al. (2016) similarly find strong evidence in favor of insurance within the family. Using CPS data, Mankart and Oikonomou (2015) document a substantial response of female labor force participation to spousal unemployment, where the response is more drawn out over time compared to early tests of the added worker effect. Moreover, Shore (2010) provides evidence that intrahousehold risk sharing is particularly strong within recessions. Our findings of a shift over time in the aggregate behavior of labor supply by gender and marital status suggest that it would be productive to expand on these findings by examining whether insurance within the family has undergone similar shifts at the microlevel.[ab]

Our analysis of family labor supply has focused on the interaction between husbands and wives. Another dimension of insurance within the family concerns the interaction between young and old family members. Quantitative studies that focus on this dimension include Jaimovich et al. (2013), who aim to explain age differences in the volatility of labor supply, and Kaplan (2012), who quantifies the role of the option of moving in and out of the parental home as an insurance mechanism for young workers. Building on this work, Dyrda et al. (2016) develop a business cycle model that allows for the option of young people moving in with their parents. They find that living arrangements matter a lot for labor supply elasticities: the elasticity is three times larger for young people who live with their parents compared to those who live alone. Accounting for household formation also implies that the aggregate labor supply elasticity is much larger than the microelasticity for stable households.

2.5 Changing Families and Aggregate Savings

In addition to providing a theory of labor supply, the representative household that populates baseline macroeconomic models also provides a theory of savings. In this section, we argue that models that go beyond representative households by explicitly modeling families have important implications for the determination of savings in the macroeconomy.

There are a few different channels through which families matter for savings; they relate to the life cycle savings motive and the precautionary savings motive. First, changes in the size of the household over time (eg, through marriage, divorce, and having children) imply that consumption needs vary over the life cycle, which is reflected in the optimal level of saving. Second, the precautionary savings motive also plays an important role in macroeconomic models (at least since Aiyagari, 1994). The strength of the precautionary motive depends on the insurance mechanisms people have access to. Similar to our analysis of labor supply above, we will argue that insurance within the family plays

[ab] Some evidence in this direction is provided by Blau and Kahn (2007), who show that married women's labor supply has become less responsive to their husbands' wages since the 1980s.

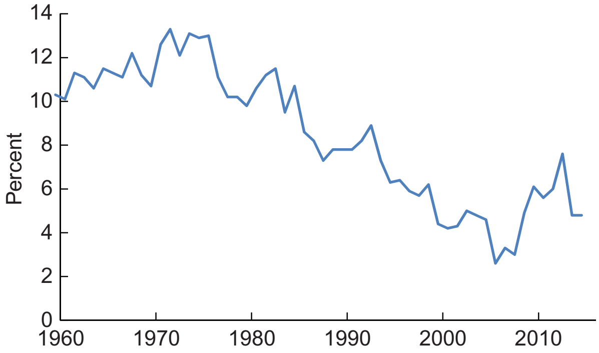


Fig. 15 Personal savings rate, United States. *Bureau of Economic Analysis, retrieved from FRED, St. Louis Fed.*

an important role in the sharing of income risk and hence in the determination of savings. Third, not only do families affect the sharing of existing sources of risk, but accounting for families also introduces new sources of risk. Getting married and having children can lead to (sometimes large) additional expenses, and to the extent that people face uncertainty over marriage and fertility, this should affect their precautionary savings. Equally important is the probability that a family dissolves: divorce is common and in many cases represents a sizeable financial risk.

The large shifts in fertility, marriage, and divorce over the last few decades suggest that the family determinants of savings may have been responsible for some of the changes in aggregate savings behavior over time. In particular, in the United States the personal savings rate has declined steadily from more than 10% in the late 1970s to less than 5% in the mid-2000s (see Fig. 15). Various explanations have been proposed for this change, although no single explanation is widely accepted (see Guidolin and Jeunesse, 2007 for an overview and discussion). In this section, we examine the possibility that changes at the family level may have played a role.

As far as the life cycle savings motive is concerned, there is a substantial literature within macroeconomics that accounts for the life cycle using a unitary model of the household, ie, without making an explicit distinction between the interests of different household members. Life cycle models were first introduced to modern business cycle research by Attanasio and Browning (1995) and Ríos-Rull (1996). In such models, the varying consumption needs due to changes in family composition over the life cycle can be incorporated through consumption equivalence scales.[ac] There is a small literature that uses life cycle models to quantify the impact of population aging on savings (Miles, 1999; Ríos-Rull, 2001). Depending on future population growth, these effects on the

[ac] See, for example, Fernández-Villaverde and Krueger (2007) and Fernández-Villaverde and Krueger (2011).

savings rate can be large, although they generally occur too slowly to explain much of the rapid decline in the savings rate in recent decades.

Given that there is already a sizeable literature on the life-cycle motive for saving, our discussion here is focused primarily on the implications of marriage and divorce for aggregate savings, a topic on which relatively few papers exist.

2.5.1 Savings and Divorce

In the models discussed in Section 2.4, we examined differences in the behavior of single and married households, while taking the existence of these different types of households as given. In reality, most adults start out as singles, marry at some point in their life, and many return to being single, eg, due to divorce. We now consider the implications for savings of the possibility of divorce. We start by taking marital bargaining power as given and by modeling divorce as an exogenous shock; endogenous bargaining and endogenous divorce will be considered below.

We consider a married couple whose life extends over two periods. The couple is married in the first period, and in the second period the union continues with probability $1-\pi$, whereas with probability π a divorce occurs. The divorce regime is that in the case of a divorce the wife retains fraction κ_f of assets, and the husbands receives $\kappa_m = 1 - \kappa_f$.

We focus on implications for savings and take as given that both spouses work in both periods.[ad] Let d' denote savings. The couple bargains cooperatively with bargaining weights given by λ_f and $\lambda_m = 1 - \lambda_f$. The couple's decision problem in the first period can be formulated as follows:

$$\max_{c_f, c_m, d'} \Big\{\lambda_f \log(c_f) + \lambda_m \log(c_m) \\ + \beta[\lambda_f(\pi V_f^D(d') + (1-\pi)V_f(d')) + \lambda_m(\pi V_m^D(d') + (1-\pi)V_m(d'))]\Big\}$$

subject to the budget constraint:

$$c_f + c_m + d' = w_f + w_m.$$

Here $V_g(d')$ is the second period value function for spouse $g \in \{f,m\}$ if the union continues, and $V_g^D(d')$ is the value function in the case of divorce.

In the case of divorce, in the second period each spouse simply consumes earnings and savings, which earn interest at rate r. We therefore have:

$$V_g^D(d') = \log(w_g' + (1+r)\kappa_g d').$$

[ad] Clearly, the possibility of divorce also affects the incentive to work, in part by altering the marginal utility of wealth, and in more complex environments also through the accumulation of individual-specific labor market experience.

In contrast, if the marriage continues, consumption shares are given by bargaining weights:

$$V_g(a') = \log(\lambda_g(w'_f + w'_m + (1+r)a')).$$

We can now consider the savings problem in the first period. The first-order condition for a' is given by:

$$\begin{aligned}\frac{1}{w_f + w_m - a'} = \beta\pi &\left[\frac{\lambda_f(1+r)\kappa_f}{w'_f + (1+r)\kappa_f a'} + \frac{\lambda_m(1+r)\kappa_m}{w'_m + (1+r)\kappa_m a'}\right] \\ &+ \beta(1-\pi)\frac{1+r}{w'_f + w'_m + (1+r)a'}.\end{aligned} \tag{6}$$

The optimal savings in the case of no divorce risk ($\pi = 0$) are:

$$\tilde{a} = \frac{\beta(1+r)(w_f + w_m) - w'_f - w'_m}{(1+\beta)(1+r)}.$$

Now consider the case $\pi > 0$. The optimal savings will be unchanged at $\tilde{a}$ if the following condition is satisfied:

$$\frac{w'_g + (1+r)\kappa_g\tilde{a}}{\lambda_g} = w'_f + w'_m + (1+r)\tilde{a}$$

for $g \in \{f, m\}$, or:

$$\kappa_f = \widetilde{\kappa f} \equiv \frac{-\lambda_m w'_f + \lambda_f w'_m + \lambda_f(1+r)\tilde{a}}{(1+r)\tilde{a}},$$

$$\kappa_m = \widetilde{\kappa m} \equiv \frac{\lambda_m w'_f - \lambda_f w'_m + \lambda_m(1+r)\tilde{a}}{(1+r)\tilde{a}},$$

where we have $\widetilde{\kappa}_f + \widetilde{\kappa}_m = 1$ as required. Intuitively, this specific divorce regime recreates the same consumption allocation that would have been obtained had the marriage continued, and hence savings incentives are unchanged. What happens when κ_f does not equal $\widetilde{\kappa}_f$ depends on relative female and male bargaining power. The derivative of the right-hand side of (6) with respect to κ_f is given by:

$$\beta\pi(1+r)\left(\frac{\lambda_f w'_f}{(w'_f + (1+r)\kappa_f a')^2} - \frac{\lambda_m w'_m}{(w'_m + (1+r)\kappa_m a')^2}\right)$$

Evaluating this expression at $a' = \tilde{a}$, $\kappa_f = \widetilde{\kappa}_f$, and $\kappa_m = \widetilde{\kappa}_m$ gives:

$$\frac{\beta\pi(1+r)}{\left(w'_f + w'_m + (1+r)\tilde{a}\right)^2}\left(\frac{w'_f}{\lambda_f} - \frac{w'_m}{\lambda_m}\right).$$

Hence, the derivative is positive if $w'_f/\lambda_f > w'_m/\lambda_m$, which is equivalent to $\widetilde{\kappa}_f < \lambda_f$. A positive derivative, in turn, implies that when $\kappa_f > \widetilde{\kappa}_f$, the optimal savings a' satisfy

$a' > \tilde{a}$, ie, the presence of divorce risk increases savings. More generally, divorce risk increases savings if for the spouse who is made worse off by divorce the asset share in divorce exceeds the relative bargaining power in marriage. Intuitively, under this condition increasing savings lowers the additional inequality across spouses brought about by divorce, which generates a precautionary demand for savings.[ae] If the couple starts out with equal bargaining power and there is an equal division divorce regime $\lambda_f = \lambda_m = \kappa_f = \kappa_m = 0.5$, the possibility of divorce always leads to precautionary savings, except in the knife edge case where the divorce regime that exactly reproduces the married allocation. The intuition is the same as for the usual motive for precautionary savings with preferences that display prudence. Under divorce, one spouse ends up with less consumption and the other one with more consumption compared to the married state. Due to the curvature in utility, the outcome of the less fortunate spouse receives higher weight when savings are determined in the first period, leading to an increase in precautionary savings.

We derived these results under the assumption that the divorce leaves the consumption possibilities of the couple unchanged. Realistically, there are also direct costs of divorce and forgone returns to scale from having a joint household. Hence, the possibility of divorce would also induce a negative income effect, which further increases desired savings.

To summarize the results, the effect of divorce risk on savings depends on the divorce regime (ie, the property division rule in divorce) and also on the relative bargaining power of the spouses. In practice, the most common divorce regimes in the data are the title-based regime and the equitable distribution regime.[af] Under the title-based regime, each spouse gets to keep the marital assets that are already in her or his name; ie, real estate goes to the owner listed in the title, bank accounts go to the account owner, and so on. Under the equitable distribution regime, judges have discretion in dividing assets in divorce. Often an equal division of marital assets is a starting point, but judges can make allowances for different needs (eg, the spouse with custody for children may receive more assets). When men are the main breadwinners and also hold title to major assets such as real estate, cars, and bank accounts, we would expect divorce under the title-based regime to lead to a precautionary demand for savings, because the wife is likely to be worse off in divorce compared to marriage. However, the precautionary demand only arises if the wife is able to save in her own name, because otherwise she would not be able to increase her outcome in divorce. Predictions are more ambiguous under the equitable distribution regime, because in this regime the wife may obtain more consumption in divorce compared to marriage. Comparing across regimes for a given divorce rate, as long as equitable distribution is more advantageous for the spouse with less power than the title-based system (as seems likely), a switch to equitable distribution (which occurred

[ae] This is a local result close to the marriage allocation.

[af] Additional possibilities include an equal division regime, and a regime where the division of assets is set through enforced prenuptial agreements.

in most US states in the 1970s) will weaken the precautionary motive and hence lead to lower savings.

What is more, individual labor earnings are likely to make up a large fraction of income in divorce. The rise in married women's earnings over time also implies that women are better able to support themselves after divorce (under either divorce regime). Hence, for a given divorce risk, the rise in married women's labor force participation and the decline in the gender pay gap are likely to have lowered the precautionary demand for savings associated with divorce over time.

2.5.2 Savings and Divorce with Endogenous Bargaining Power

The analysis so far suggests that divorce may have a substantial impact on a country's personal savings rate. Divorce is one of the largest and most common risks people face today (along with unemployment and ill health). Moreover, changes in the divorce rate, the divorce regime, and female labor force participation all affect how much precautionary saving arises from divorce risk, and thus may be in part responsible for changes in the savings rate over time.

In the preceding analysis, we introduced divorce as an exogenous shock, and the impact of divorce risk on couples' behavior was proportional to the probability with which this shock occurred. In this setting, the possibility of divorce has large effects only if the divorce rate is high. We now extend our analysis by endogenizing the divorce decision and the evolution of bargaining power within the marriage. We will see that in this extended model, the mere possibility of divorce can affect household behavior, so that large impacts on behavior can arise even if few couples divorce in equilibrium. Hence, the extension further amplifies the potential role of divorce for explaining how a country's savings rate is determined.

We consider a variant of the model above in which bargaining and divorce are endogenous. The ability of the spouses to commit to future allocations is limited by the ability to divorce, so that divorce functions as a threat point that informs bargaining during the marriage. In the first period, the couple is married and starts out with initial bargaining power $\lambda_f$and λ_m, where $\lambda_f + \lambda_m = 1$. In the second period, the couple experience marriage quality shocks ξ_f, ξ_m, which can be positive or negative. There is a unilateral divorce regime; that is, the marriage continues in the second period only if both spouses are at least as well off married compared to being divorced.

In the first period, the couple's decision problem can be written as:

$$\max\left\{\lambda_f \log(c_f) + \lambda_m \log(c_m) + \beta\left[\lambda_f E(V_f(a', \xi_f, \xi_m)) + \lambda_m E(V_m(a', \xi_f, \xi_m))\right]\right\},$$

subject to the budget constraint:

$$c_f + c_m + a' = w_f + w_m.$$

Here $V_g(d', \xi_f, \xi_m))$ is the expected utility of spouse g in the second period as a function of the state variables d', ξ_f, and ξ_m.

In the second period, the decision problem of the couple is constrained by the possibility of divorce. If a divorce takes place, existing property is divided with share κ_f for the wife and $\kappa_m = 1 - \kappa_f$ for the husband. Utilities conditional on divorce are therefore given by:

$$V_g^D(d') = \log(w_g' + (1+r)\kappa_g d').$$

The full decision problem in the second period can then be written as:

$$\max_{D \in \{0,1\},\, c_f,\, c_m} \Big\{\lambda_f \Big[(1-D)\big(\log(c_f) + \xi_f\big) + DV_f^D(d')\Big] + \lambda_m\big[(1-D)(\log(c_m) + \xi_m) + DV_m^D(d')\big]\Big\} \tag{7}$$

subject to:

$$c_f + c_m = w_f' + w_m' + (1+r)d', \tag{8}$$

$$(1-D)\big(\log(c_f) + \xi_f\big) + DV_f^D(d') \geq V_f^D(d'), \tag{9}$$

$$(1-D)(\log(c_m) + \xi_m) + DV_m^D(d') \geq V_m^D(d'). \tag{10}$$

Here $D \in \{0,1\}$ denotes the endogenous divorce decision and c_f, c_m is the consumption allocation conditional on staying married. Clearly, by setting $D = 1$ (divorce) the constraints (9) and (10) can always be met. However, divorcing is optimal only if there is no consumption allocation that leaves both spouses at least as well off married compared to divorced.

The decision problem in the second period can be solved by first considering a spouse who ends up just indifferent between divorce and staying married. Let $\tilde{\lambda}_g$ denote the consumption share that would make spouse g indifferent between these options, for a given ξ_g. The indifference condition is:

$$\log\Big(\tilde{\lambda}_g\Big(w_f' + w_m' + (1+r)d'\Big)\Big) + \xi_g = \log\Big(w_g' + (1+r)\kappa_g d'\Big),$$

which can be solved to give:

$$\tilde{\lambda}_g = \frac{w_g' + (1+r)\kappa_g d'}{\exp(\xi_g)\Big(w_f' + w_m' + (1+r)d'\Big)}.$$

The second period outcome can now be determined by comparing the implicit bargaining weights $\tilde{\lambda}_f$ and $\tilde{\lambda}_m$ to the actual ex ante bargaining weights λ_f and λ_m. In particular:

Proposition 3 (Divorce and Bargaining Power in Limited Commitment Model)

The outcome of the couple's decision problem in the second period can be characterized as follows:

- *If $\widetilde{\lambda}_f \leq \lambda_f$ and $\widetilde{\lambda}_m \leq \lambda_m$, the couple stays married (D = 0), and consumption is:*

$$c_f = \lambda_f \left(w'_f + w'_m + (1+r)d' \right),$$
$$c_m = \lambda_m \left(w'_f + w'_m + (1+r)d' \right).$$

- *If $\widetilde{\lambda}_f > \lambda_f$ and $\widetilde{\lambda}_f + \widetilde{\lambda}_m \leq 1$, the couple stays married (D = 0), but the wife's consumption share is increased to satisfy her participation constraint. Consumption is:*

$$c_f = \widetilde{\lambda}_f \left(w'_f + w'_m + (1+r)d' \right),$$
$$c_m = w'_f + w'_m + (1+r)d' - c_f.$$

- *If $\widetilde{\lambda}_m > \lambda_m$ and $\widetilde{\lambda}_f + \widetilde{\lambda}_m \leq 1$, the couple stays married (D = 0), but the husband's consumption share is increased to satisfy his participation constraint. Consumption is:*

$$c_m = \widetilde{\lambda}_m \left(w'_f + w'_m + (1+r)d' \right),$$
$$c_f = w'_f + w'_m + (1+r)d' - c_m.$$

- *If $\widetilde{\lambda}_f + \widetilde{\lambda}_m > 1$, the couple divorces (D = 1), and consumption is:*

$$c_f = w'_f + (1+r)\kappa_f d',$$
$$c_m = w'_m + (1+r)\kappa_m d'.$$

The implications of the possibility of divorce for savings are similar to those of the exogenous-divorce model above, but savings are affected already when one of the spouses' participation constraints is binding, even if the marriage continues.

Fig. 16 presents a computed example to show how the trend toward higher labor market participation of married women would affect divorce and the savings rate in the model with endogenous bargaining and divorce.[ag] Male earnings are normalized to $w_m = 1$, and the equilibrium savings rate and divorce rate are shown for female earnings varying from $w_f = 0.1$ to $w_f = 0.8$. The divorce regime is unilateral divorce with an equal division of marital assets upon divorce. Given that total earnings are constant and the interest rate equals the inverse of the discount factor, if there was no possibility of divorce, the savings rate would be equal to zero regardless of female earnings. Hence, any positive savings are due to the precautionary motive generated by the possibility of divorce.

[ag] The parameter values used are $\lambda_f = 0.4$, $\lambda_m = 0.6$, $r = 0.05$, and $\beta = 1/(1 + r)$. The divorce regime features equal division of assets, $\kappa_f = \kappa_m = 0.5$, and the marriage quality shocks ξ_f and ξ_m are uniformly distributed on the interval [−0.2, 1] and are independent across the spouses.

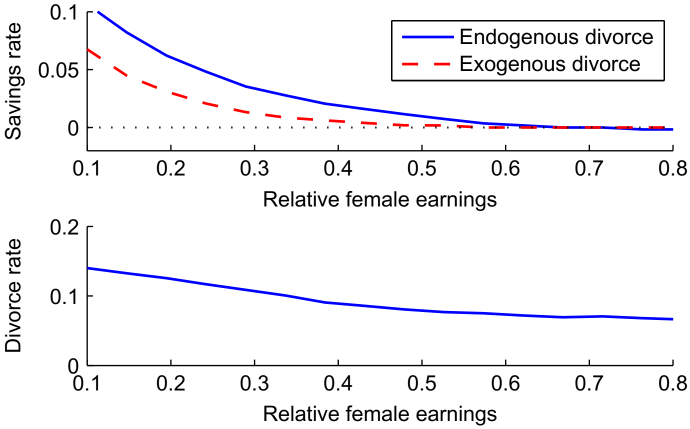


Fig. 16 Savings rate and divorce rate as a function of relative female earnings.

With endogenous bargaining and divorce, we see that the savings rate and divorce rate are both positive, and sharply decreasing in relative female earnings. Once female earnings are above 60% of male earnings, the savings rate approaches zero (the value that would be obtained without the possibility of divorce). The intuition for these findings is that for low female earnings, divorce leaves women much worse off compared to marriage. The equal division of assets only provides limited insurance, because most of the second period income of the couple is due to the husband's earnings. Thus, the possibility of divorce leads to a precautionary demand for savings primarily to insure women against the possibility of divorce. Own earnings provide an alternative route of insurance and also increase the overall share of income that women can claim in divorce. Hence, as earnings rise, precautionary savings are much reduced and ultimately disappear.

The picture also displays the savings rate in the exogenous divorce model when the equilibrium divorce rate (displayed in the lower panel) is fed as an exogenous variable into the model of the previous section (ie, the exogenous divorce rate varies together with female earnings). The exogenous divorce model generates qualitatively similar findings, but the impact on savings is much smaller in size. In the exogenous divorce model, as long as the couple stays married, bargaining power stays at the initial value. In contrast, in the endogenous divorce model, there are couples where, say, the husband is at the participation constraint (the realization of ξ_m is low), so that the wife has to offer additional compensation to the husband for the husband to stay. This need to compensate the other spouse generates an additional need for precautionary savings. Hence, the endogenous divorce model generally leads to a larger impact on the savings rate and can generate a feedback from the possibility of divorce on aggregate variables even if the realized divorce rate is low.

2.5.3 Additional Notes on Related Literature

There are only a few papers that use models of the type outlined here to address macroeconomic questions. Dynamic models of marriage under limited commitment with the possibility of divorce have been introduced by Mazzocco (2007), Mazzocco et al. (2013), and Voena (2015). In these models, the shifts in bargaining power that are necessary when one of the spouses' participation constraint is binding have persistent effects on the marital allocation. By specifically addressing how divorce law affects incentives for saving, Voena (2015) is the closest to the questions addressed here. Voena finds (using an estimated structural model) that the introduction of unilateral divorce (in states with an equal division of property) leads to higher savings and lower female employment. Intuitively, the introduction of unilateral divorce removes spouses' veto power in the divorce decision, which reduces risk sharing and increases precautionary savings. To our knowledge, there are no studies that analyze how the possibility of divorce (in a given divorce regime) affects the private savings rate (and other aggregate variables) in light of other observed changes to the family, such as the rise in female labor force participation and relative female earnings and the decline in fertility.

An early study that considers the role of divorce as an exogenous shock is Cubeddu and Ríos-Rull (2003). They assess the potential role of divorce for asset accumulation by comparing counterfactuals that differ in when (or if) people marry and divorce, and in how costly divorce is. Unlike in the model outlined above, consumption within marriage is constrained to be equal across spouses. They find that the impact of marriage and divorce can be large in their setting, but they do not directly relate this finding to observed changes in macro variables.[ah]

Love (2010) documents empirically (and analyzes in a quantitative model) how asset allocations change with marital-status transitions. As in Cubeddu and Ríos-Rull (2003) and Hong and Rios-Rull (2012), changes in marital status are modeled as exogenous shocks, and there is only public consumption in marriage. The theoretical model predicts that portfolio shares (ie, the fraction of wealth invested in stocks vs bonds) should react sharply to fertility, marriage, and divorce. Empirical results based on the Health and Retirement Study and the Panel Study on Income Dynamics are supportive of some of the predictions of the model, although not for all groups of households.

Fernández and Wong (2014a,b) use a quantitative life cycle model with exogenous divorce to study the importance of the likelihood of divorce for explaining the rise in female labor force participation from the 1960s to the 1990s. They argue that the increase in divorce risk accounts for a substantial fraction of the increase in female labor force participation. The main reason for this finding is that women (who often have lower wages

[ah] A similar framework is used by Hong and Rios-Rull (2012) in a setting that also accounts for the arrival of children, stochastic survival, and bequest motives, and uses information on life insurance holdings to infer how the utilities of different household members interact.

than their husbands and need to provide for their children) face lower consumption possibilities after a divorce, which increases desired savings. One way of increasing savings is to work more during marriage, which raises the total resources of the household and facilitates the smoothing of consumption between the married and divorced states. In Fernández and Wong (2014c) this analysis is extended to a setting with endogenous divorce.

In addition to increasing savings and increasing labor supply, another insurance mechanism that is likely to be relevant in the data is education. In Guvenen and Rendall (2015), women acquire education in part as insurance against a bad marriage. Guvenen and Rendall argue that the introduction of unilateral divorce increases this insurance motive, accounting for a sizeable fraction of the increase in female education and helping rationalize the observation that women now obtain more higher education than do men.[ai]

2.6 Private Information in the Household

Throughout Section 2, we have used a number of different approaches for modeling husband–wife interactions. We now step back from the applied questions to discuss the relative advantages of different models of the family and their uses within macroeconomics. The pioneering work of Gary Becker was largely based on the so-called unitary model of the family. A unitary model distinguishes between, say, male and female labor supply, but does so in the context of a single household utility function rather than allowing for separate preferences for each spouse. This approach is also how the family was first introduced into macroeconomics in the literature on home production and the business cycle (eg, Benhabib et al., 1991; Greenwood and Hercowitz, 1991). The limitation of the unitary approach is that since it does not distinguish individual utility functions, it does not allow for conflict of interest between spouses. This restricts the range of questions that can be addressed by the unitary model. Moreover, there is a sizeable literature in family economics that empirically tests the unitary model against richer alternatives that allow for bargaining, and finds strong evidence against the unitary model.[aj]

To go beyond the unitary model, one needs to start with women and men (characterized by separate utility functions) as primitives and then analyze how they act either together as couples or as singles. Within couples, one has to specify some form of bargaining process that determines how the couple resolves the conflict of interest between the spouses. Two broad classes of bargaining models that can be used for this purpose are

[ai] Another perspective on higher premarital investments by women is provided by Iyigun and Walsh (2007a), who focus on the impact of investments both on sorting of spouses and on bargaining power within marriage (see also Iyigun and Walsh, 2007b; Chiappori et al., 2009).

[aj] See Alderman et al. (1995) for an early summary of the evidence, and Attanasio and Lechene (2002) for an influential contribution based on Progresa data from Mexico.

noncooperative bargaining models (where the interaction between the spouses is modeled as a noncooperative game, using standard game theory tools) and cooperative bargaining models (where the spouses are able to achieve an outcome that is at least statically efficient). A common argument in favor of cooperative bargaining is that marriage is usually a sustained long-term relationship, which suggests that the spouses should be able to avoid major inefficiencies. However, while the majority of recent work in family economics uses a cooperative approach, other authors provide evidence in favor of inefficient bargaining outcomes within the family,[ak] and noncooperative models have been used by Lundberg and Pollak (1994), Konrad and Lommerud (1995), and Doepke and Tertilt (2014), among others.

Within the literature on cooperative bargaining in the family, many papers use explicit bargaining models such as Nash bargaining subject to divorce as the outside option.[al] Another popular approach, introduced by Chiappori (1988; 1992), is to only impose that the couple reaches a statically efficient outcome, but to remain agnostic about the details of the bargaining process. Empirical implementations of this approach often allow bargaining power to be a function of observables (called "distribution factors") such as the relative education or the relative age of the spouses, without specifying the mechanism through which these variables matter.[am] The advantage of this approach, labeled the "collective model," is its generality, because all (static) efficient allocations can be characterized in this way. The labor supply model employed in Section 2.4 is an example of a collective model (albeit with fixed bargaining power).

The collective approach is less suitable for dynamic contexts, because it does not provide an explicit theory for how bargaining within a couple evolves. This would perhaps not matter much if bargaining weights were constant over time, which would also imply ex-ante efficiency, ie, full insurance in the household. Yet there is plenty of empirical evidence of limited risk sharing in couples. For example, based on data from Kenya, Robinson (2012) documents that private expenditures increase in own labor income. Duflo and Udry (2004) use data from the Ivory Coast to show that the composition of household expenditure is sensitive to the gender of the recipient of a rainfall shock that affects male and female income differentially. The evidence is not exclusive to developing countries. Cesarini et al. (2015) document a larger fall in labor earnings after winning a lottery for the winners relative to their spouses in Sweden. One could rationalize such findings in a collective model where the bargaining weights move due to shifts in relative income, wages, or lottery winnings. However, this approach has the downside of

[ak] See, eg, Udry (1996), Duflo and Udry (2004), and Goldstein and Udry (2008).

[al] The classic papers are Manser and Brown (1980) and McElroy and Horney (1981). Another classic is the "separate spheres" bargaining model of Lundberg and Pollak (1993), which is an interesting hybrid between a cooperative and a noncooperative model.

[am] See, for example, Attanasio and Lechene (2014).

violating ex-ante efficiency without being explicit about the underlying bargaining friction. Moreover, the approach precludes transitions to a (presumably) noncooperative state such as divorce, which is an important limitation given that divorce is commonplace (see Fig. 6).

A more fruitful avenue in our view is to take a stand on the friction that prevents couples from achieving full insurance and model it explicitly. One obvious friction is limited commitment. Since spouses usually have the option to walk away from each other (ie, divorce or separation), at any point in time each spouse should get at least as much utility as his or her outside option. This is what we alluded to at the end of Section 2.4 and modeled more explicitly in the endogenous bargaining model of Section 2.5. A limited literature on dynamic household decisions pursues this avenue.[an] A model based on limited commitment will lead to endogenous shifts in bargaining power over time, namely whenever the commitment constraint becomes binding. When divorce is the outside option, limited commitment implies shifts in bargaining power only when a couple is close to divorce. An alternative is to consider an outside option of noncooperation within marriage as in Lundberg and Pollak (1993). Doepke and Kindermann (2015) is a recent example of a dynamic bargaining model with such an outside option. Such limited commitment models are consistent with the empirical evidence on continuously shifting bargaining power within couples provided by Lise and Yamada (2015).

An alternative friction that so far has received much less attention is private information within the household. Before showing how this friction can be modeled, let us discuss some indications that private information may indeed be relevant for bargaining between spouses. There are many things that spouses may not precisely know about each other, such as income, assets, consumption, work effort, or preferences. Contrary to the belief that love and altruism will lead to perfect information sharing between spouses, the evidence suggests otherwise. The most obvious example may be that people do not typically tell their partner when they are having an extramarital affair. Relatedly, some people do not disclose that they have HIV or other sexually transmitted diseases to their partner. Women sometimes hide from their partners that they are using birth control (or, depending on the context, that they are not using birth control).[ao] More directly related to the context of this chapter is that people do not always disclose income, spending, and savings behavior to their spouse. de Laat (2014) shows that husbands in split-migrant couples in Kenya invest significant resources into monitoring the spending behavior of their wives. When given the option, people often prefer to put money into private (and possibly secret) accounts.[ap]Hoel (2015) finds in Kenyan data that 31% of people say their spouse was not aware of any income they had received the preceding week.

[an] See in particular Mazzocco (2007) and Voena (2015).

[ao] For example, Ashraf et al. (2014) show that women in Zambia hide the use of birth control from their husbands when given the chance.

[ap] See Anderson and Baland (2002), Ashraf (2009), and Schaner (2015).

Further, evidence from lab and field experiments suggests that information treatments affect intrahousehold allocations, suggesting that information frictions are important.[aq] Most of this evidence is from developing countries and in some dimensions (such as uncertainty about a spouse's income) couples in industrialized countries with joint checking accounts and tax filings may be less affected by information frictions. However, private information about preferences and hidden effort is likely to be equally relevant all around the world.

In sum, there is ample evidence that private information plays an important role in household bargaining. Nevertheless, hardly any work has been done on this issue in terms of explicit models of the bargaining process. We believe this is an important area for future work. While most of this chapter concerns applying family economics to macroeconomics, the issue of information frictions presents an opportunity for intellectual arbitrage in the opposite direction: while in family economics static models are still common, in macroeconomics dynamic contracting models that make the underlying frictions explicit have been widespread for many years. In particular, it should be possible to apply some of the tools to analyze informational frictions currently used in theoretical macroeconomics and public finance to issues in family economics.[ar] Some work of this kind exists in development economics (eg, Townsend, 2010; Karaivanov and Townsend, 2014; Kinnan, 2014), but the question is a different one as the degree of insurance within a village—as opposed to within a couple—is analyzed.

We currently explore how to account for information frictions in household bargaining in ongoing work (Doepke and Tertilt, 2015). As a simple example for modeling such a friction, consider a variant of the model analyzed above under private information about each spouse's labor income w_g. To simplify the exposition, we assume that there is a private income realization only in the first period, whereas there is no income in the second period, $w'_f = w'_m = 0$. Bargaining is assumed to be efficient subject to the constraints imposed by private information, with initial welfare weights λ_f and λ_m. The constrained efficient allocation can be computed as a mechanism design problem. The revelation principle can be applied and implies that we can restrict attention to truth-telling mechanisms with truth-telling constraints imposed. Hence, the spouses will simultaneously report their income w_f and w_m to each other, and consumption is given by functions $c_g(w_f, w_m)$ and $c'_g(w_f, w_m)$, which depend on the reports. For simplicity, we

[aq] When income is private information in dictator games, less is transferred to the partner Hoel (2015). Migrants send home less cash to family members when their choice is not revealed to the recipients (Ambler, 2015). More is spent on goods that are hard to monitor or difficult to reverse and less on household public goods when a transfer is given privately to one spouse relative to a full information transfer (Castilla and Walker, 2013).

[ar] See Atkeson and Lucas (1992) and the follow-up literature for applications of models with information frictions in macroeconomics. For a survey of the literature incorporating information frictions into public finance, see Golosov et al. (2006).

assume that each income is drawn from a finite set $w_g \in W_g$ with independent probability distributions denoted by $p(w_g)$.

With these preliminaries, the optimization problem faced by the household can be written as follows:

$$\begin{aligned}\max E \quad \{&\lambda_f[\log(c_f(w_f,w_m)) + \beta \log(c_f'(w_f,w_m))]. \\ &+ \lambda_m[\log(c_m(w_f,w_m)) + \beta \log(c_m'(w_f,w_m))]\},\end{aligned}$$

subject to the budget constraints:

$$\begin{aligned} c_f + c_m + a' &= w_f + w_m, \\ c_f' + c_m' &= (1+r)a. \end{aligned}$$

The maximization problem is also subject to truth-telling constraints. Consider first the wife. For each w_f and each alternative $\widetilde{w}_f \in W_f$, we impose:

$$\begin{aligned}&\sum_{w_m} p(w_m)[\log(c_f(w_f,w_m)) + \beta \log(c_f'(w_f,w_g))] \\ &\qquad \geq \sum_{w_m} p(w_m)[\log(c_f(\widetilde{w}_f,w_m) + w_f - \widetilde{w}_f) + \beta \log(c_f'(\widetilde{w}_f,w_m))].\end{aligned}$$

Similarly, for the husband we have:

$$\begin{aligned}&\sum_{w_f} p(w_f)\left[\log(c_m(w_f,w_m)) + \beta \log(c_m'(w_f,w_g))\right] \\ &\qquad \geq \sum_{w_f} p(w_f)\left[\log(c_m(w_f,\widetilde{w}_m) + w_m - \widetilde{w}_m) + \beta \log(c_m'(w_m,\widetilde{w}_m))\right].\end{aligned}$$

A direct implication of this model is that consumption is more responsive to a change in own income than to a change in the spouse's income. The reason is that incentives need to be provided to tell the truth about own income shocks. Other frictions (such as unobservable effort or unobservable preference shocks) can be modeled along similar lines.

Models of bargaining with limited commitment frictions and private information frictions have distinct implications for how consumption and leisure depend on bargaining power. Consider, for example, a limited commitment model where the outside option responds to income shocks. In such a setting, a positive income shock for a given spouse increases this spouse's bargaining weight, which (all else equal) tends to increase leisure and lower labor supply. In contrast, in a hidden effort model it is costly to distort the effort of a productive spouse; hence, a more productive spouse may be provided more incentives to work and end up working more. This example shows that the underlying friction matters for how household bargaining reacts to family trends such as the increase in women's labor market attachment. We believe that further work on incorporating methods for dealing with dynamic contracting frictions into family economics will be productive for improving our understanding of these issues.

3. THE FAMILY AND ECONOMIC GROWTH

The most fundamental questions in macroeconomics concern economic growth. As Robert Lucas put it, once one starts to think about the determinants of cross-country income differences and policies that may allow poor countries to catch up with rich ones, "it is hard to think about anything else" (Lucas, 1988, p. 5).

Early theorizing on the sources of economic growth was focused on firms rather than families. The Solow model, for example, puts investment in physical capital by the business sector into the spotlight, coupled with exogenous improvements in productivity. To be sure, even in a model driven by capital accumulation families matter for growth; after all, investment has to be financed by savings, and savings are determined within the family. Both husband-wife and parent-child interactions are relevant for savings. First, as already shown in Section 2.5, a couple's savings rate responds to the possibility of divorce. More generally, if husbands and wives disagree about the consumption-savings trade-off (eg, because they differ in their degree of patience), then how spouses negotiate affects the savings rate. Second, a large part of long-run wealth accumulation is due to bequests, for which interactions between parents and children are crucial.

Family decisions have become even more central to growth theory with more recent developments that emphasize the importance of human capital accumulation and endogenous population growth. The importance of human capital accumulation for growth has been well recognized since the work of Lucas (1988). To fix ideas, consider a simple endogenous growth model based on accumulation of human capital H and physical capital K. Final output is produced using physical capital and effective units of labor as inputs. Effective units of labor depend both on time spent working u and the stock of human capital. Assuming a simple Cobb–Douglas production function, output is:

$$Y = K^{\alpha}(uH)^{1-\alpha}.$$

Human capital is accumulated by spending time studying. The higher the level of human capital and the more time spent in school $(1 - u)$, the higher is tomorrow's human capital,

$$H' = B(1-u)H, \tag{11}$$

where B is a technology parameter. In the simplest model, the fraction of time spent in school is given exogenously. Then, the growth rate of output in the balanced growth path is simply $B(1-u)$. Growth thus depends not only on technology but also on the time spent in school.

So far we have taken u to be an exogenous parameter. But clearly the time spent on education is a choice. Who makes the choice? A large part of education happens during childhood and hence, leaving mandatory schooling laws aside, it is parents who make education decisions for their children. In other words, education is a family decision. Note also that the formulation of the human capital production function above assumes past human capital enters into next period's human capital. Intuitively, the initial human capital stock of a new member in society is proportional to the level already attained by older members of the family. As Lucas put it, "human capital accumulation is a social activity, involving groups of people in a way that has no counterpart in the accumulation of physical capital" (Lucas, 1988, p. 19). Much of the time, the group in which the accumulation happens is the family, where children learn from parents both by imitating them and by being actively taught.

Understanding the human capital accumulation process is an active research area. Many open questions remain, but what is understood by now is that education and skill formation are complex processes that involve many ingredients. Inputs both in forms of time (own time, teacher time, parental time) and goods (textbooks, school buildings) are important, as is the age at which specific investments take place. For example, Jim Heckman and coauthors have emphasized the importance of early childhood education for long-run outcomes (Heckman, 2008). Citing Cunha and Heckmann (2007), "The family plays a powerful role […] through parental investments and through choice of childhood environments." Recent research captures such links in formal models of human capital investments within families (eg, Caucutt and Lochner, 2012; Aizer and Cunha, 2012). Del Boca et al. (2014) find that both paternal and maternal time input are essential inputs into child development.

So far, we have motivated the importance of families for growth based on the intuitive argument that human capital and savings decisions are made in the family. An equally compelling argument for the importance of families can be made on the basis of empirical findings. As we will document in the next section, cross-country data show strong correlations between development indicators such as GDP per capita and measures of family structure. While such findings constitute no proof of causality, they suggest a close link between family structure and development. After documenting these facts, we will show in a sequence of simple growth models how modeling increasingly complex family interactions can affect economic growth in an economy. While the most straightforward link from families to growth concerns fertility decisions, we emphasize that there are many dimensions to families, their role in producing new people being only one of them. Families typically consist of many family members (husband, wife, sons, daughters), who may differ in preferences and skills. When preferences differ, the exact nature of the decision process in the family becomes important. When skills differ, ie, when men and women are not perfect substitutes in production, then the details of how they enter differently into the human capital and goods production functions will also matter for growth. Further, families may have different attitudes toward sons and daughters, affecting human capital investment, and institutions such as polygyny may also affect incentives for investing in human and physical capital.

3.1 Cross-Country Family Facts

In this section, we report strong correlations between indicators of economic development and measures of family structure. Perhaps the most well-known example is the close link between the fertility rate and development. Fig. 17 displays a strong negative relationship between the total fertility rate and GDP per capita across countries.[as] Fertility, in turn, is strongly negatively correlated with measures of schooling (Fig. 18).

Many other measures of family structure are related to development as well. Fig. 19 displays the fraction of teenage girls (15–19 years) that has ever been married. The figure reveals a striking negative relationship between GDP per capita and early marriage. In poor countries, such as Ghana and Malawi, almost 50% of 15–19 year old girls are

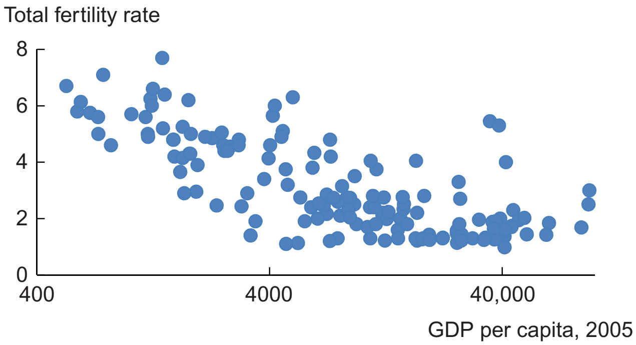


Fig. 17 TFR and GDP per capita across countries. *GID 2006 and World Development Indicators 2005.*

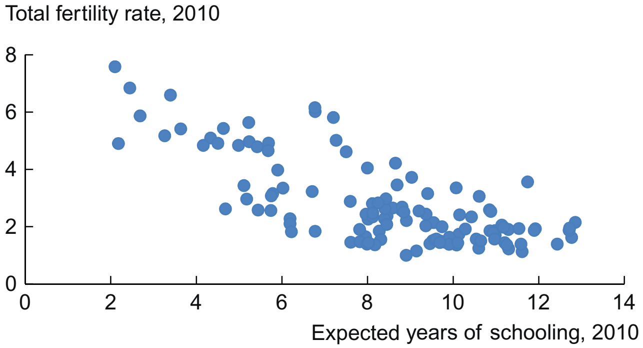


Fig. 18 Schooling and TFR across countries. *World Development Indicators.*

[as] A similar relationship can be observed over time within countries: in most cases, the demographic transition took place during times of rapid economic growth. For the United States, the decline in children ever born by birth cohort of mothers is shown in Fig. 1.

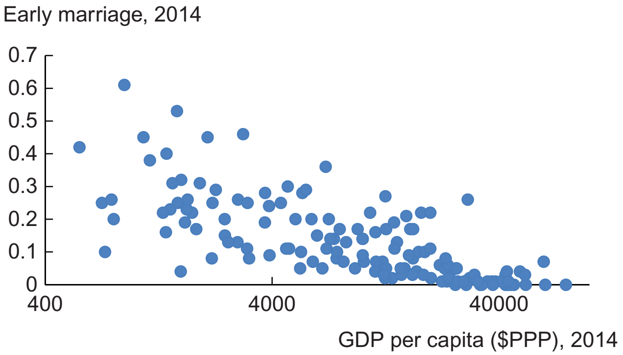


Fig. 19 Early marriage and GDP per capita across countries. *OECD Gender Statistics 2014 and World Development Indicators.*

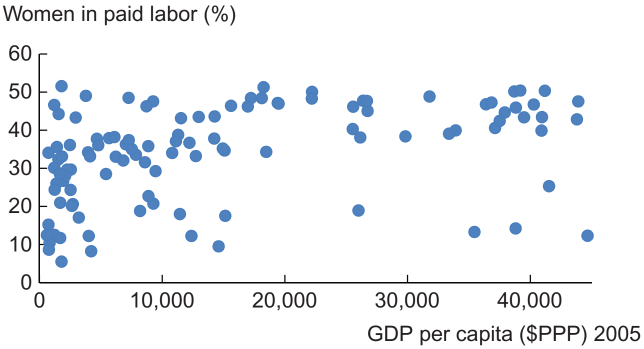


Fig. 20 Women in paid labor and GDP per capita across countries. *OECD Gender Statistics 2006 and World Development Indicators.*

married, compared to less than 5% in countries with a GDP per capita of more than $25,000 (in 2005 PPP terms). Fig. 20 plots the relationship between female labor force participation and GDP per capita. Since rates of formal employment are low for women and men alike in many poor countries, rather than plotting the absolute participation rate, Fig. 20 depicts the fraction of formal employment accounted for by women. In virtually all countries with a GDP per capita higher than $20,000, women make up 40% or more percent of the paid labor force, while in many poor countries women account for less than 20%.[at]

The figures discussed so far were chosen to highlight a few particularly interesting and pronounced relationships between family structure and development. Yet, essentially all indicators of family structure are related to development, including both measures of outcomes and measures of legal differences between men and women. Table 3 gives

[at] The few rich countries with low female labor force participation are oil-rich countries such as Saudi Arabia and the United Arab Emirates.

correlations of family variables with two measures of economic development, GDP per capita and the share of the agricultural sector in GDP (which is typically low in developed countries). The first three rows are about children: Fertility rates are high, child mortality is high, and schooling is low in poor countries. The next two rows show that a preference for sons is systematically related to development. First, people in poor countries are more likely to state that when resources are scarce, educating boys is more important than educating girls. Second, inheritance laws favor sons over daughters. The next three rows are about the education and work of women relative to men. Women are more likely to be illiterate than men in poor countries. They work less in the market and provide a larger burden of unpaid family care work, such as taking care of children and the elderly. The next set of indicators show that the legal position of women is negatively related to development. Women obtained access to politics (through representation in national parliaments) earlier in today's rich countries. They also have better access to land ownership and usage. There is also a tight relationship between the United Nations' Gender Empowerment Measure and GDP per capita. The last set of indicators show that the position specifically of married women is weaker in poor countries. Women in poor countries marry earlier than in rich countries and wife beating is more accepted. The legal position also favors men in poor countries: inheritance laws are more likely to favor

Table 3 Correlations between family variables and GDP per capita and share of agriculture across countries

Variable	GDP p.c.	Share agric.
Total fertility rate, GID 2006	−0.49	0.71
Child mortality rate, WDI 2014	−0.54	0.75
Average years of schooling, WDI 2003	0.76	−0.79
Son preference in education, GID 2014	−0.26	0.33
Inheritance discrimination against daughters, GID 2014	−0.24	0.45
Female literacy relative to male, GID 2006	0.37	−0.65
Percent females in paid labor force, GID 2006	0.32	−0.52
Unpaid care work by women, GID 2014	−0.37	0.43
Year first woman in parliament, UN 2004	−0.58	0.36
Women's access to land, GID 2014	−0.41	0.54
Gender empowerment measure, UN 2004	0.70	−0.60
Early marriage, GID 2014	−0.50	0.65
Agreement with wife beating, GID 2014	−0.42	0.57
Inheritance discrimination against widows, GID 2014	−0.21	0.42
Laws on domestic violence, GID 2014	−0.16	0.46

Notes: Data are from OECD gender, institutions, and development data base (GID 2006 and 2014), the world development indicators (WDI 2003, 2005, and 2014), and the UN Development Report 2004. Correlations are computed with GDP per capita and percentage of value-added in agriculture from the WDI in two different years: 2005 and 2014. See the Appendices for variable definitions and further details.

Table 4 Differences between polygynous countries and monogamous countries close to the equator

	Polygynous	Monogamous
Total fertility rate	6.8	4.6
Husband–wife age gap	6.4	2.8
Aggregate capital–output ratio	1.1	1.9
GDP per capita (dollars)	975	2798
Number of countries	28	58

Notes: Data are either from 1980 or an average for the 1960–85 time period. Details and sources are given in Tertilt (2005). Polygynous countries defined as countries with at least 10% of men in polygamous unions. Monogamous countries are all other countries within 20 degrees of latitude from the equator, to control for the fact that most polygynous countries are in sub-Saharan Africa.

widowers over widows, and laws against domestic violence (if they exist in the first place) are less strict compared to developed countries.

A family structure that has long been illegal in most developed countries but is still practiced in many poorer countries is polygyny, which is the practice of men being married to multiple wives. Table 4 shows that polygynous countries are among the poorest in the world, display extremely high fertility rates, invest little, and are characterized by large age gaps between husbands and wives.

3.2 Parents and Children

The strong empirical association between economic development and measures of family structure suggests that changes to the family are an integral part of the growth process. We now analyze a series of simple growth models to highlight a number of specific channels that tie development and families together.

We start with a simple view of the family. In this first version of the model, each family consists of a parent and a child. Parents care about children in a warm-glow fashion. Specifically, they derive utility from their children's full income.[au] Fertility is exogenous. In other words, we start with a single sex model where each parent has exactly one child. Since the children themselves will have children again, the model is an overlapping generations model. The difference to the standard OLG setup is that generations are explicitly linked through parent-child relationships.

Preferences are given by the utility function

$$u(c)+\delta u(y'),$$

where c is the parent's consumption and y' is the child's full income (as an adult in the next period). For simplicity, we assume consumption goods are produced at home with a

[au] Models with true altruism would yield qualitatively similar results, but are less tractable.

production function that uses effective units of time as the only input.[av] Let H denote the human capital of the parent and ℓ the units of time the parent devotes to production. Then consumption, or equivalently GDP (per adult), is given by:

$$c = A\ell H,$$

where A is a technology parameter. We define full income as the income that would be obtained if the parent was working full time:

$$y = AH.$$

Not all time will be devoted to production, because the parent will also spend some time educating the child. Let e denote this education time. Human capital of the child is given by the following production function:

$$H' = (Be)^{\theta} H,$$

where B and θ are technology parameters. Here θ is an especially important parameter as it captures the returns to education. Each parent is endowed with one unit of time. Thus, the parent faces the following time constraint: $\ell + e \leq 1$. Assuming log utility, we can write the objective function of the parent as follows:

$$\max \log(c) + \delta\theta \log(e).$$

The equilibrium is characterized by the optimal education choice $e^* = \dfrac{\delta\theta}{1+\delta\theta}$. The equilibrium growth rate (for both human capital and consumption) is:

$$\frac{H'}{H} = \left(B\frac{\delta\theta}{1+\delta\theta}\right)^{\theta}. \tag{12}$$

As in the simple Lucas model at the beginning of this section (Eq. 11), the human capital accumulation technology in part determines the growth rate. What is different from the Lucas model is that how much parents care about their children's well-being also enters. In contrast, in standard growth models that abstract from intergenerational links, it is the individual's discount factor that matters. There is no reason for the rate of time preference across periods for a given person to coincide with the intergenerational discount factor. A related point is that the intergenerational elasticity of substitution may differ from the intertemporal elasticity of substitution (IES). In other words, estimates of the IES in the business cycle context are not necessarily relevant for calibrating growth models based on trade-offs across generations.[aw] There is a need for empirical research in this

[av] This is isomorphic to a model with market production. The home production formulation has the advantage that we do not need notation for wages and, later, interest rates.

[aw] See Cordoba and Ripoll (2014) for a formal treatment of this point.

area, as good estimates of the intergenerational discount factor and the intergenerational elasticity of substitution are currently not available.

The model as written assumes that all families accumulate human capital independently from each other. An alternative vision of the process of human capital accumulation is that much of the increase in people's productivity over time is due to the dissemination of productive ideas, implying that exchange of knowledge between different families is crucial for growth. In a setting that makes this engine of growth explicit, de la Croix et al. (2016) examine the role of institutions that organize the exchange of knowledge for growth. They compare both family-based institutions (knowledge exchange within nuclear families or families/clans) and market-based institutions, and argue that institutions that facilitated the exchange of ideas across families were crucial for the economic ascendency of Western Europe in the centuries leading up to industrialization.

3.3 Adding Fertility Choice

Next, we enrich the model by endogenizing fertility choice. The analysis of fertility choices in explicit dynamic growth models was pioneered by Becker and Barro (1988) and Barro and Becker (1989). These papers assume an altruistic utility function (ie, the children's utility enters the parent's utility), whereas we will stick to the warm-glow motive for investing in children. This distinction makes no difference for most qualitative results and allows more closed form solutions. In contrast to Barro and Becker (1989), which features exogenous technological progress, our focus is on human capital as the engine of growth.

For simplicity (and in line with the majority of existing analyses of fertility in dynamic models), we stick with one-parent families. However, conceptually it is straightforward to consider fertility decisions in a two-parent model (see Doepke and Tertilt, 2009 for an example).[ax]

To give the parent a reason to want children, we modify the utility function as follows:

$$u(c) + \delta^n u(n) + \delta u(y'),$$

where n is the number of children chosen by the parent. It takes ϕ units of time to raise a child in addition to the e units of education time devoted to each child. Note that ϕ is a fixed cost, while e is a choice variable. The time constraint is thus

[ax] Doepke and Kindermann (2015) document empirically that spouses often disagree about whether to have another child and present a bargaining model of fertility decisions to analyze the implications of this fact.

$$\ell + (\phi + e)n \leq 1.$$

We keep everything else (ie, production and human capital accumulation) as before. Assuming log utility, the objective function can be written as

$$\max \log(c) + \delta^n \log(n) + \delta\theta \log(e).$$

To guarantee that the problem is well defined, we assume $\delta^n > \delta\theta$.

The equilibrium is characterized by the following education and fertility choices:

$$e^* = \frac{\delta\theta}{\delta^n - \delta\theta}\phi,$$

$$n^* = \frac{(\delta^n - \delta\theta)}{\phi(1 + \delta^n)}.$$

The equilibrium growth rate is:

$$\frac{H'}{H} = \left(B\frac{\delta\theta\phi}{\delta^n - \delta\theta}\right)^{\theta}. \tag{13}$$

Comparing the expression for n^* and the equilibrium growth factor given in (13), it becomes apparent that many of the same features leading to high fertility, such as a low cost of children and low returns to education, also lead to a low growth rate. The negative dependence of fertility on growth was already a feature in Barro and Becker (1989), albeit in a model of exogenous growth. The importance of human capital as an engine for growth in a model with endogenous fertility was first analyzed by Becker et al. (1990). While the exact expression is different, they also derive a growth rate that depends positively on the returns to education, the fixed cost of children, and an altruism parameter.

Comparing the growth rate given in (13) with the growth rate in the model without fertility choice (12), two points emerge. First, two types of intergenerational preference parameters appear now: δ and δ^n. In other words, how much parents care about the quality vs the quantity of children is a determinant of the growth rate. Second, the return to human capital enters positively into the optimal education choice and negatively into the optimal fertility choice.

These results may help in understanding some empirical regularities, such as the negative relationship between fertility and schooling, on the one hand, and fertility and GDP per capita, on the other hand (Figs. 17 and 18). In the model, these relationships would arise if countries differ in the return to skill θ or the cost of children ϕ. Similarly, within most countries fertility decreased, while education increased over time. The model can generate this pattern if the return to education increases gradually from generation to generation. The resulting theory interprets the demographic transition to low fertility as driven by a move from investing in child quantity to emphasizing child quality (ie, education).

There is a substantial literature aiming to account for the historical relationship between fertility and growth based on this mechanism. Before the onset of industrialization in the 18th century, living standards around the world were stagnant, and fertility rates were high. In most countries, this "Malthusian" stage was followed by a transition to growing incomes and declining fertility rates. The first theory to fully account for such a transition is Galor and Weil (2000), which is based on the quantity–quality trade-off, a Malthusian constraint due to the role of land in agriculture, and human capital as an engine for growth. The role of structural change in the transition is highlighted by Hansen and Prescott (2002), who model the endogenous transition from a stagnant land-intensive technology to a capital-intensive growth technology. Population growth changes with growing incomes in their model. However, rather than explicitly modeling fertility choice, the authors assume a particular dependence of population growth on consumption. Greenwood and Seshadri (2002) introduce explicit fertility preferences when analyzing a similar transition from an agricultural to a manufacturing society. Doepke (2004) also models fertility preferences explicitly to analyze the importance of education and child labor policies for the transition from stagnation to growth. Some authors argue that the transition was triggered by declines in mortality, which increased the incentive to educate children. Soares (2005) provides a model where gains in life expectancy lead to reductions in fertility and increases in human capital accumulation, leading to an endogenous transition from a Malthusian to a long-run growth equilibrium.[ay] However, Hazan and Zoabi (2006) show that the impact of increasing longevity on human capital investment is mitigated by the fact that higher longevity also raises the incentive to have more children, which works against human capital investment through the quantity–quality trade-off.

One could also use variants of this setup to understand cross-country fertility differences today. For example, Manuelli and Seshadri (2009) study international fertility differences using a life-cycle version of the Barro–Becker model with human and health capital. They find that differences in productivity, social security, and taxes can go a long way in explaining the observed differences.

The empirical regularities that characterize differences across countries are also visible across families. There is a sizeable empirical literature documenting that in the cross section of families in a given country, quantity and quality of children are negatively related.[az] An augmented version of the model with heterogeneity across families in δ^n (or, similarly, δ) would deliver this empirical regularity. The overall economy-wide

[ay] The importance of changes in mortality for development is also analyzed in Cervellati and Sunde (2005).

[az] See, for example, Rosenzweig and Wolpin (1980) and Bleakley and Lange (2009). Vogl (2016) argues that the negative relationship of quantity and quality may be a relatively recent phenomenon. He documents that in many developing countries there was a reversal in the education-fertility relationship from positive to negative. Baudin et al. (2015) provide an analysis that also allows for the possibility of childlessness, and argue that childlessness is U-shaped as economies develop.

growth rate would then depend on how many parents of each type exist, and also on whether such preferences are passed on from parents to children or randomly distributed in the population.[ba] de la Croix and Doepke (2003) explore the association between inequality and growth based on the differential fertility channel and argue that it explains a large part of the observed relationship between inequality and growth across countries.[bb]

3.3.1 Fertility Restrictions

The link between fertility and human capital accumulation suggests that countries may be able to speed up economic development by limiting fertility rates. Out of the many policies that can affect a country's fertility rate, the most direct is a hard limit on how many children a couple can have. Several countries have implemented such fertility restrictions, the most famous example of which is the one-child policy of China. Another examples are forced sterilization policies implemented by the Indian government in the 1980s. Other countries have used more subtle family planning policies, either through monetary incentive schemes or in the form of media campaigns, often advocating a two-child norm.

We can incorporate such policies into the model by adding a fertility limit $\bar{n}$. Whenever the constraint is binding, the optimal education decision is:

$$e^* = \frac{\delta\theta[\frac{1}{\bar{n}} - \phi]}{1 + \delta\theta}.$$

Education increases with a tighter fertility restriction. Thus, fertility restrictions do speed up economic growth in our model. Yet, they are not the panacea one might have hoped for, as fertility restrictions also come with a cost. Fig. 21 illustrates these effects in a computed example of our model.[bc] The top panels show how fertility and education change with different levels of fertility restrictions, while the bottom panels depict the growth rate and steady state utility as a function of the restrictions. The optimal (unrestricted) fertility rate in the example is 3. Thus, only restrictions below 3 are binding. Tighter restrictions lead to higher levels of education and higher growth rates, but they lower equilibrium utility. In our simple model, this negative effect on utility comes from parents being deprived of (part of) the enjoyment they

[ba] Thus, whether differential fertility increases or decreases the growth rate depends on many factors. See Vogl (2016) for an analysis of this point. The specific role of preference transmission in the context of the British Industrial Revolution is analyzed by Doepke and Zilibotti (2008).

[bb] de la Croix and Doepke (2004, 2009) analyze the importance of this mechanism in the context of education policies.

[bc] The parameters in the example are: $\delta^n = 0.8$, $\delta = 0.5$, $\phi = 0.1$, $B = 1$, $\theta = 0.5$, $A = 10$. The initial level of human capital is normalized at $H = 1$ and the fertility restriction ranges from 1 to 5.

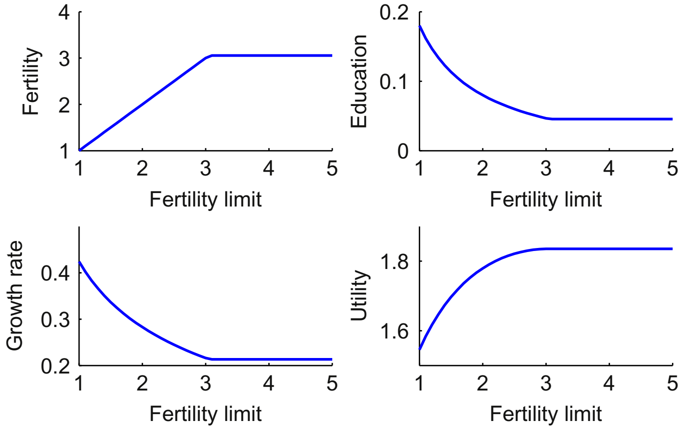


Fig. 21 Fertility restrictions.

obtain from children.[bd] In more elaborate settings, such negative effects can also arise from the differential effect of the fertility constraint on a heterogeneous population. Also, with a public social security system, lower fertility depresses future payouts, ie, the demographic dividend declines, a problem that is starting to become pressing in China right now.

These issues are analyzed in a small emerging literature. Liao (2013) analyzes how the one-child policy in China increased human capital and output. She simulates counterfactual experiments to analyze the effects of a relaxation of the policy. The main findings are that results differ across generations and skill groups. In particular, the initial old would benefit from a sudden unexpected relaxation of the policy, but future generations would be hurt. Moreover, such a policy would hurt unskilled people more than skilled people. Choukhmane et al. (2014) conduct a richer analysis using a life-cycle model and more detailed micro data. They argue that a large part of the rise in aggregate savings in China can be attributed to the one-child policy. The focus in Banerjee et al. (2014) is on the importance of general equilibrium effects when estimating how fertility restrictions (and their removal) would impact savings. These authors argue that appropriately taking general equilibrium effects into account reduces the size of such estimates. Coeurdacier et al. (2014) focus on the interaction between fertility policies and social security reform.[be] Since an expansion of social security lowers the incentives to have children (and thereby lowers the number of contributors to the system), the relaxation of the one child policy is likely to have smaller effects than typically anticipated. The authors find that this effect is quantitatively important for China.

[bd] The mechanism that lower fertility decreases utility is analyzed in Cordoba (2015), who finds that, during the 1970–2005 period, world growth in well-being was lower than the growth rate in per capita consumption precisely because fertility fell so dramatically during that period.

[be] Song et al. (2015) also analyze the consequences of low fertility for pension reform in China, albeit in a model with exogenous fertility.

3.4 Two-Parent Families: Decision Making

The vast majority of the literature on fertility and growth focuses on the interaction between parents and children in one-gender models. In other words, reproduction is asexual and differences between men and women in technology and preferences are abstracted from. We now expand our analysis by introducing two-gender families. In this version of our growth model, children have two parents: a mother and a father. For simplicity we return to exogenous fertility for now and assume that each couple has two children. Thus, families now consist of a husband, a wife, a son, and a daughter. Suppose men and women disagree about how much they care about their children's well-being.[bf] As in Section 2.4, suppose that the couple solves a Pareto problem with fixed bargaining weights, where $\lambda_f$is the bargaining weight of the woman, and λ_m is the weight of the man. Then the objective function is:

$$\lambda_f[u(c)+\delta_f u(y')]+(1-\lambda_f)[u(c)+\delta_m u(y')].$$

To keep the rest of the model comparable to the previous section, we assume that all consumption in marriage is public and the total time endowment (of the couple) is still one. We also make no distinction between sons and daughters in the parent's objective function. We will relax these assumptions further below. Assuming log utility, the objective function can be written as:

$$\max \lambda[\log(c)+\delta_f\theta \log(e)]+(1-\lambda)[\log(c)+\delta_m\theta \log(e)].$$

Equilibrium education now is

$$e^* = \frac{\tilde{\delta}\theta}{1+\tilde{\delta}\theta},$$

where $\tilde{\delta} \equiv \lambda_f \delta_f + (1-\lambda_f)\delta_m$. Thus, the equilibrium growth rate is:

$$\frac{H'}{H} = \left(B\frac{\tilde{\delta}\theta}{1+\tilde{\delta}\theta}\right)^{\theta}. \tag{14}$$

A comparison of Eqs. (12) and (14) shows not only that gender preference gaps matter for the growth rate, but also how such preferences make their way into decisions within the family. Specifically, assuming mothers care more about children than fathers do ($\delta_f > \delta_m$), the economy grows faster, the larger the bargaining power of women. Doepke and

[bf] There could be many reasons for such a disagreement, ranging from biological/evolutionary arguments to cultural factors. See Alger and Cox (2013) for a survey.

Tertilt (2009) explore the endogenous evolution of women's rights based on such a mechanism (details will be discussed in Section 4). However, whether female empowerment enhances growth depends on the details of the bargaining process within the household. Doepke and Tertilt (2014) use a noncooperative model to show that what looks like gender differences in preferences may ultimately be due to specialization in tasks within the household. Based on this mechanism, Doepke and Tertilt (2014) show that monetary transfers to women may reduce growth, even if women are more likely to spend transfers on children. The reason is that the equilibrium is characterized by a division of labor in which women are in charge of time-intensive tasks such as education, while men provide money-intensive goods and hence are in charge of savings and physical capital accumulation. In such a world, exogenous transfers to women (financed by a tax on men) increase human capital accumulation but reduce physical capital accumulation. Depending on the production function, such a reallocation may increase or decrease growth. Specifically, when returns to physical capital relative to human capital are high, then such a policy would lower growth. To assess whether this is an issue in reality, more empirical research is needed. The current literature on the effects of transfers to women largely focuses on child expenditures, but there is little work analyzing effects on savings and investment.

3.5 Two-Parent Families: Technology

Empirical research (eg, Del Boca et al., 2014) has shown that mothers and fathers are both important factors in the human capital formation process of their children. In most families, both mothers and fathers spend a significant amount of time with children (Schoonbroodt, 2016). Further, men and women may not be perfect substitutes in market production.[bg] To address these issues, we now extend our view of the family to include fathers and mothers explicitly in the human capital formation process and also men and women as entering separately into production. To isolate the role of women in technology (vs their role as decision makers), we assume again that all consumption in families is public and that men and women have the same preferences regarding their children. In other words, we ignore here the additional complication that arises if fathers and mothers disagree (which we analyzed in Section 3.4). We also focus on the education decision (rather than fertility choice); however, it would be straightforward to include both margins in the same model.

[bg] Large and persistent gender wage differentials exist (see Blau and Kahn, 2000 for a survey). There is an extensive empirical literature trying to analyze their causes. We do not take a stand here on what the ultimate cause is, but rather explore the implications of men and women being imperfect substitutes in production. Whether the gap is due to different innate skills, different preferences, or cultural factors leading to differences in skill acquisition is largely irrelevant for our analysis.

In contrast to the previous versions of the model, men and women enter differently into technology. The consumption good is produced with a Cobb-Douglas production function using both male and female efficiency units of time as inputs,

$$c = A(\ell_f H_f)^{\alpha}(H_m)^{1-\alpha},$$

where $\alpha \in (0,1)$. For simplicity, we assume that only women raise children, while men work full time. The female time constraint is $\ell_f + e_f + e_m \leq 1$, where e_f is the time invested in educating daughters, and e_m is time devoted to the education of sons. Full income is defined as the production function evaluated at $\ell_f = 1$ and is therefore given by:

$$y = AH_f^{\alpha}H_m^{1-\alpha}.$$

Each couple has two children: a daughter and a son. Both mothers and fathers are essential for their children's human capital accumulation:

$$H_f' = (Be_f)^{\theta}H_f^{\beta}H_m^{1-\beta}, \tag{15}$$

$$H_m' = (Be_m)^{\theta}H_f^{\beta}H_m^{1-\beta}, \tag{16}$$

with $\beta \in (0,1)$. In summary, there are two gender differences in this setup: the relative importance of women vs men in transmitting own human capital to children (β) and the relative importance of women vs men in production (α).[bh]

Assuming log utility, the objective function can be written as:

$$\max \log(c) + \delta[\alpha\theta \log(e_f) + (1-\alpha)\theta \log(e_m)].$$

The equilibrium allocation is:

$$\ell_f^* = \frac{\alpha}{\alpha + (1-\alpha)\delta\theta + \alpha\delta\theta},$$

$$e_m^* = \frac{(1-\alpha)\delta\theta}{\alpha + (1-\alpha)\delta\theta + \alpha\delta\theta},$$

$$e_f^* = \frac{\alpha\delta\theta}{\alpha + (1-\alpha)\delta\theta + \alpha\delta\theta}.$$

The equilibrium ratio of female to male human capital is given by:

$$\frac{H_f}{H_m} = \left(\frac{e_f}{e_m}\right)^{\theta} = \left(\frac{\alpha}{1-\alpha}\right)^{\theta}.$$

Note that the asymmetry between mothers and fathers in the human capital production function captured by β does not appear in this expression. This is not a fundamental

[bh] A third asymmetry is that we have assumed that only women can spend time educating children. But this asymmetry is made for tractability and is not essential for the qualitative results.

result, but rather a feature of our warm-glow altruism. In an altruistic model, parents would take into account that educating their children will turn the children themselves into better parents, and hence enable them to provide grandchildren with more education. In such a formulation, the relative importance of fathers vs mothers in child development will also enter the relative human capital of men and women in equilibrium.

This model features a gender education gap and accordingly a gender wage gap.[bi] Specifically, the wage ratio per unit of time is $\frac{w_f}{w_m} = \frac{\alpha}{1-\alpha}$. The more productive women are in production (higher α), the smaller is the gender education gap. Higher female wage increase the opportunity cost of time and hence make children more costly. In a variant of the model with endogenous fertility, this logic would lead to fertility decline in response to rising female productivity. This mechanism is analyzed by Galor and Weil (1996), who explore how this channel contributed to the demographic transition.

In a fully altruistic model, parents would further take into account that their sons and daughters will be working different hours in the market (because of the child-bearing obligations of mothers) and accordingly invest less in daughters.[bj] This amplification channel is explored by Echevarria and Merlo (1999). Lagerlöf (2003) further explores the effect of the marriage market in this context and stresses the importance of multiple equilibria. If all families invest more into sons, then daughters on average expect high spousal income, which lowers the incentive for each individual family to educate daughters. However, complete gender equality is also an equilibrium in his model.

Plugging the ratio of human capital back into the human capital production function, we get the following equilibrium growth rate (for both male and female human capital, and hence also output and consumption):

$$\frac{H'}{H} = B^{\theta}(e_m)^{(1-\beta)\theta}(e_f)^{\theta\beta} = \left\{\frac{B\delta\theta}{\alpha+\delta\theta}(1-\alpha)^{1-\beta}\alpha^{\beta}\right\}^{\theta}. \tag{17}$$

Eq. (17) shows that the growth rate depends on many features of the family. As before, the more parents care about their children, the higher the growth rate. What is new is that gender differences in technology also matter for growth. This is true for both the role women play in production (as captured by α) and the relative importance of fathers and mothers in human capital transmission (captured by β). Moreover, the two dimensions of technology interact. For example, in a world where men and women enter symmetrically into production ($\alpha = 0.5$), the relative importance of mothers and fathers in human capital transmission becomes irrelevant. On the other hand, α always enters,

[bi] Strictly speaking there are no wages in our formulation with home production. However, the model can be reinterpreted as one with market production and wages given by marginal products.

[bj] Our warm-glow altruism does not capture this channel, because parents care about the full income of their children and do not take into account the time daughters will spend on child-bearing.

even in a world where mothers and fathers are equally important in human capital transmission ($\beta = 0.5$). Closer inspection of (17) shows that the growth rate is hump-shaped in α. Thus, whether an increase in α increases or decreases the growth depends on the starting point. Starting from a low role of women in production, an increase in α will lead to a reduction in the gender education gap, an increase in relative female wages, an increase in female labor force participation, and an acceleration of economic growth. This mechanism may well have been historically relevant: recall that Fig. 20 displays a strong positive relationship between GDP per capita and the role of women in paid labor. Similarly, recall that Table 3 showed a negative correlation between the gender education gap and development.

Since World War II, all developed countries went through a period of increasing female labor force participation and declining gender wage gaps. How women's role for production evolved over longer historical time periods is less clear. Humphries and Weisdorf (2015) construct measures of relative male and female wages in England dating back to 1270 and find large swings over the centuries. They also try to measure the wages of married and single women separately, using the distinction between casual work (more relevant for married women) and annual contracts (mostly used for unmarried women). Using their data and accepting their interpretation, we find that the relative wages of married vs single women over time have sometimes moved in the opposite directions (Fig. 22). There is also evidence suggesting that in the long run, the relationship between development and female market work is not always monotonic. Specifically, based on cross-country data, Goldin (1995) argues that female labor supply is U-shaped in development.[bk] A similar point is made by Costa (2000), who argues

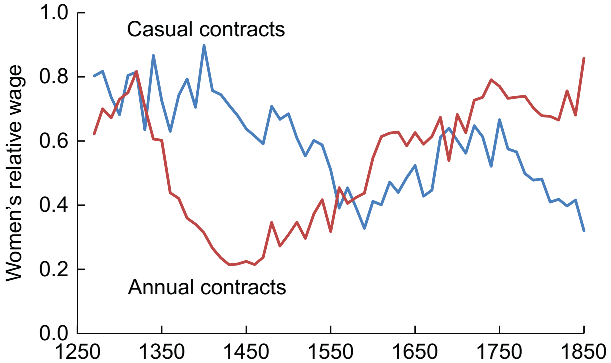


Fig. 22 Historical wage gap in England. *Humphries, J., Weisdorf, J., 2015. The wages of women in England, 1260–1850. J. Econ. Hist. 75 (2), 405–447 (Table A1).*

[bk] See also Olivetti (2014) for evidence of a U-shape in time series data of 16 developed countries (including the United States) and Mammen and Paxson (2000) for evidence from India and Thailand.

that female labor force participation is N-shaped if one goes back far enough in time. Establishing such historical facts is difficult not only due to lack of reliable data but also because of the lack of a sharp distinction between market and home production in agricultural economies.[bl]

A further complication arises when market production is made up of different tasks. If individuals differ in their ability to perform different tasks, then the allocation of talent to activities becomes important. Norms about gender roles (or other barriers) can then be an obstacle to the optimal allocation of talent to tasks. Hsieh et al. (2013) analyze the importance of this channel in the United States. They find that an improved allocation of talent across genders (and also ethnic groups) accounts for 15–20% of US growth during the 1960–2008 period. Lee (2015) explores the importance of misallocation of female talent for cross-country income differences. The paper finds that entry barriers for women in the nonagricultural sector play a large role for the observed low agricultural productivity in poor countries.

3.6 Two-Parent Families: Endogenous Bargaining

In Section 3.4, we have seen that who makes decisions in the household matters for growth. Hence, an important question is what determines bargaining power in marriage.[bm] Here we are interested in what changes bargaining weights across generations, which is distinct from the analysis of endogenous bargaining over time for a given couple (which we considered in Section 2). Initial bargaining power should be determined at time of marriage, which we do not model here. It is often assumed that relative educational attainments matter in the marriage market and hence for bargaining power. Relative education between men and women may itself be endogenous as we have seen in Section 3.5. In this section, we connect these two forces. To do so, we impose that the bargaining weight is a function of the gender education gap, which is itself chosen in the family. This assumption allows us to analyze the feedback from a gender education gap to bargaining power in the family.[bn]

We use a model that combines the setup with a gender preference gap in Section 3.4 with gender differences in technology as explored in Section 3.5. First, consider such a

[bl] For example, Goldin (1995) includes unpaid farm and family firm workers, while our Fig. 20 includes only paid workers.

[bm] There is a sizeable literature estimating models of household decision making. Key for identification is typically the existence of so-called distribution factors that affect bargaining weights but are exogenous to the bargaining process (see, for example, Blundell et al., 2005).

[bn] Basu (2006) also explores the implications of endogenous bargaining power, albeit in a different context. We are interested in how bargaining power changes across generations, while Basu (2006) analyzes the dynamic implications for a given couple. By adjusting labor supply, and thus income, spouses may affect their bargaining power in the household.

setup with exogenous bargaining power. Combining the features of the two models, the couple solves the following maximization problem:

$$\max_{c,\, e_f,\, e_m} u(c) + \widetilde{\delta}\left\{\alpha\theta \log(e_f) + (1-\alpha)\theta \log(e_m)\right\}$$

subject to:

$$1 = \ell_f + e_m + e_f,$$
$$c = A(\ell_f H_f)^{\alpha} H_m^{1-\alpha},$$

where $\widetilde{\delta} \equiv \lambda_f \delta_f + (1-\lambda_f)\delta_m$. As before, human capital evolves according to (15) and (16). This is the same problem as in Section 3.5, but with a modified δ. Thus, the equilibrium growth rate is:

$$1 + g^{exog} = \left\{\frac{B\widetilde{\delta}\theta}{\alpha + \widetilde{\delta}\theta}(1-\alpha)^{1-\beta}\alpha^{\beta}\right\}^{\theta}.$$

Now we can explore how endogenous bargaining differs from exogenous bargaining in this setup by assuming that λ is a function of relative education. A simple functional form assumption that captures this dependence and at the same time guarantees a bargaining weight between zero and one is $\lambda(e_f, e_m) = \frac{e_f}{e_f + e_m}$. Recall that relative education is a function of the relative importance of female labor in the market: $\frac{e_f}{e_f + e_m} = \alpha$. Thus, we can replace λ_f by α and write the growth rate as[bo]:

$$1 + g^{end} = \left\{\frac{B[\alpha\delta_f + (1-\alpha)\delta_m]\theta}{\alpha + [\alpha\delta_f + (1-\alpha)\delta_m]\theta}(1-\alpha)^{1-\beta}\alpha^{\beta}\right\}^{\theta}. \tag{18}$$

Proposition 4 *Assume* $\delta_f > \delta_m$. *If* $\lambda_f < \alpha$, *then the growth rate is higher in the endogenous bargaining model, while* $\lambda_f > \alpha$ *implies a higher growth rate in the exogenous bargaining model.* This result relates women's role in technology to women's role in decision making. Specifically when women's power in decision making is low relative to their importance for production, then endogenizing the link from education to bargaining power increases the growth rate. The opposite is true when women have a lot of bargaining power relative to their importance in production.

[bo] Note that with our warm-glow altruism, parents do not take into account that when increasing their daughter's education, they also increase the daughter's bargaining weight. de la Croix and Vander Donckt (2010) analyze a model with altruism where parents explicitly consider the impact of education choices on their children's future bargaining power.

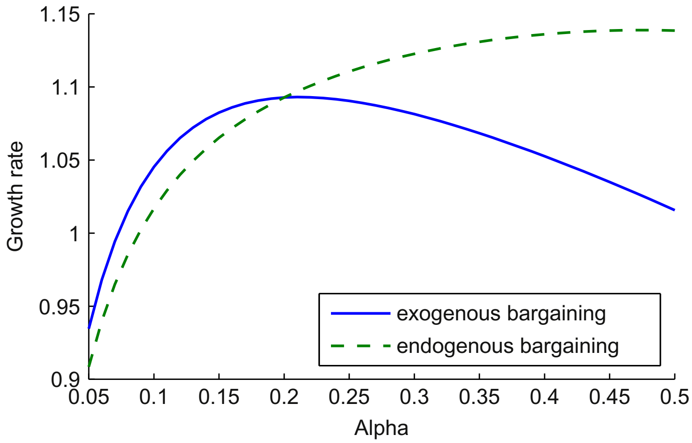


Fig. 23 Growth rate as a function of α, exogenous vs endogenous bargaining.

This result is illustrated in Fig. 23 with a numerical example.[bp] As was discussed in Section 3.5, the growth rate of the exogenous bargaining model is hump-shaped in α. This is not necessarily true in the endogenous bargaining model. In the example, growth monotonically increases in α. With fixed bargaining weights, an increase in women's role in production can lower growth because the resulting rise in female labor force participation decreases education time with children and thereby slows down human capital accumulation. This effect is mitigated in the endogenous bargaining model, where the resulting increase in bargaining power pushes toward more education (given that in the model women care more about children's education than men do). This example shows that the details of decision making in the family matter for growth and that asymmetries between men and women in decision making interact with asymmetries in technology.

3.7 Son Preferences

Many cultures are characterized by a preference for sons. This preference typically has effects on fertility behavior, where families that have only daughters are more likely to have another child (eg, Anukriti, 2014). Recently, sex-selective abortion has also been a concern (Ebenstein, 2010). Son preferences also manifest themselves in boys being treated better than girls. For example, Jayachandran and Kuziemko (2011) document gender differences in breast-feeding rates and Tarozzi and Mahajan (2007) document better nutritional status for boys in India. Further, such a preference is more pronounced in poorer countries (see Table 3).

[bp] The parameters in the example are: $\beta = 0.7, \theta = 0.5, B = 10, \delta_f = 0.5, \delta_m = 0.2, \lambda = 0.2$.

We now investigate the growth consequences of such a son preference in an extension of our model.[bq] First, consider an economy with physical capital in which parents leave bequests to sons and daughters. As before, consumption in marriage is public, fertility is exogenous, and each couple has one son and one daughter. Also as before, parents care about their children in a warm-glow fashion. In this case, parents derive utility from the bequest they give to their children. Output is produced using a linear technology in capital, ie, output is given by $y = AK$, where A is a parameter. All sons and daughters will be married. Without heterogeneity, it is irrelevant who marries whom. The capital of any given couple is made up of the sum of the bequests they each got, ie, $k = b_s + b_d$, where s denotes sons and d daughters.

Preferences are given by:

$$u(c) + \delta_s u(b_s) + \delta_d u(b_d),$$

where $\delta_s > \delta_d$ would capture a son preference. The budget constraint is $c + b_s + b_d \leq y$.

Assuming log utility, equilibrium bequests are

$$b_s = \frac{\delta_s}{1 + \delta_s + \delta_d} y,$$
$$b_d = \frac{\delta_d}{1 + \delta_s + \delta_d} y.$$

The equilibrium growth rate of income is:

$$\frac{y'}{y} = \frac{A(\delta_s + \delta_d)}{1 + \delta_s + \delta_d}.$$

The key result here is that the son preference is irrelevant for the growth rate. The only thing that matters is how much parents care on average about their children, ie, only the sum $\delta_s + \delta_d$ appears.

The finding changes if human capital accumulation is considered, as long as there are decreasing returns to educating a given person. In contrast to physical capital (where ownership does not matter for growth), it is plausible that total knowledge in an economy will be larger if knowledge is shared by more people. We now show how a son preference will interact with such decreasing returns in individual human capital.

The technologies for producing output and human capital are the same as in Section 3.5. Parents care only about their own children and hence they do not take into account that educating their daughter/son will also benefit the future son-in-law/daughter-in-law. Rather, they anticipate that their son-in-law will be endowed with the average male human capital in the economy, which we denote by $\bar{H}'_m$, and daughters-in-law are

[bq] Hazan and Zoabi (2015b) analyze endogenous son preferences in a related model with endogenous fertility.

anticipated to have human capital $\bar{H}'_f$. The optimization problem of a couple endowed with human capital (H_f, H_m) is thus given by:

$$\max_{e_f,\, e_m,\, \ell_f} u(c) + \delta_d u(y'_d) + \delta_s u(y'_s)$$

subject to:

$$\begin{aligned} c &= A(\ell_f H_f)^{\alpha} H_m^{1-\alpha}, \\ 1 &\geq \ell_f + e_f + e_m, \\ y'_d &= A(H'_f)^{\alpha}(\bar{H}'_m)^{1-\alpha}, \\ y'_s &= A(\bar{H}'_f)^{\alpha}(H'_m)^{1-\alpha}, \\ H'_f &= (Be_f)^{\theta} H_f^{\beta} H_m^{1-\beta}, \\ H'_m &= (Be_m)^{\theta} H_f^{\beta} H_m^{1-\beta}, \end{aligned}$$

where $\bar{H}'_f$ and $\bar{H}'_m$ are taken as given.

Assuming log utility, the maximization problem reduces to

$$\max_{\ell_f,\, e_f,\, e_m} \alpha \log(\ell_f) + \delta_d \alpha\theta \log(e_f) + \delta_s(1-\alpha)\theta \log(e_m)$$

subject to:

$$\ell_f + e_f + e_m \leq 1.$$

The resulting optimal education choices are

$$\begin{aligned} e_m^* &= \frac{\delta_s(1-\alpha)\theta}{\alpha + \delta_s(1-\alpha)\theta + \alpha\delta_d\theta}, \\ e_f^* &= \frac{\delta_d \alpha\theta}{\alpha + \delta_s(1-\alpha)\theta + \alpha\delta_d\theta}. \end{aligned}$$

As before, human capital, income, and consumption all grow at the same rate on the balanced growth path. The equilibrium growth rate is:

$$\left\{ \frac{B\theta}{\alpha + [\delta_d\alpha + \delta_s(1-\alpha)]\theta} (\delta_s[1-\alpha])^{1-\beta} (\delta_d \alpha)^{\beta} \right\}^{\theta}.$$

This expression shows how the effect of a son preference on the growth rate depends on the technology for goods production and human capital accumulation. First, consider the symmetric case where men and women are equally important in production (by setting $\beta = \alpha = 0.5$). Fix the total weight parents put on children: $\delta_s + \delta_d = 1$. In this case, the growth rate is maximized at $\delta_s = \delta_d = 0.5$. In other words, a son preference lowers growth. This is in contrast to the economy with only physical capital, where a son preference is irrelevant. Hence, a son preference is only growth-reducing when knowledge is the engine of growth. But even in a knowledge economy a son preference is not always disadvantageous. If men have the comparative advantage in knowledge production ($\beta < 0.5$), the growth-maximizing weight on children will display a son preference, the strength of which depends on the extent of men's comparative advantage.

On the other hand, in a world where men have a comparative advantage in goods production ($\alpha < 0.5$), but we have $\beta = 0.5$, a slight daughter preference enhances growth. The reason is that human capital is the engine of growth, implying that educating sons and daughters equally is the growth-maximizing strategy. Parents, on the other hand, do not maximize the growth rate, but rather output in the next period, where sons have the comparative advantage in production. Thus, parents overinvest in sons (compared to growth-maximizing solution). A son preference amplifies this problem.

Empirical evidence also links son preferences to the increasingly asymmetric sex ratios in some countries. In China, for example, in 2005 over 120 boys were born for each 100 girls (Wei and Zhang, 2011). Such asymmetries may have important aggregate consequences, which are largely unexplored in the literature. A notable exception is Wei and Zhang (2011), who find that rising sex ratios are an important determinant of the high Chinese savings rate. Du and Wei (2010) take this idea a step further and show in a calibrated model that this channel explains more than 50% of the current account surplus in China.

3.8 Polygyny

The role model for the family considered in most of this chapter is the Western nuclear family. The dominance of the nuclear family consisting of a husband, a wife, and the couple's own children is a relatively recent phenomenon, and even today typical families in some parts of the world do not follow this norm. Historically, the extended family (with multiple generations living together) was more prevalent than it is today.[br] Moreover, many families today no longer include married couples, as single parents are on the rise and many individuals no longer live in families at all (see Figs. 3 and 4 in Section 2.2).

Another important type of family structure is polygamy. In many parts of Africa men marrying multiple wives (polygyny) is common to the present day.[bs] Does such a family structure matter for macroeconomic outcomes? Tertilt (2005) suggests it does. The paper builds a model of polygynous families in which men buy brides and sell daughters to future husbands. The family structure reduces output (relative to enforced monogamy) through two channels. The market for daughters turns women into a valuable asset. This has two implications. First, the revenues from selling daughters become a useful way of financing old age, which depresses savings and thus physical capital. Second, it increases fertility as men want many daughters. This results in higher population growth rates,

[br] Although, because of shorter life spans, perhaps not as prevalent as one might think. See Ruggles (1994) for an extensive historical account of changing household structures in the United States over the last 150 years.

[bs] Polyandry (women having multiple husbands) is extremely rare, but a few societies exist as well.

which depresses capital per person and thus GDP per capita. The paper uses a calibrated general equilibrium model to show that this effect is quantitatively important, and shows that the mechanism can account for a large part of the observed differences between polygynous and monogamous countries shown in Table 4.

Polygyny matters for growth through its effect on brideprices. Thus, the marriage market is essential for the mechanism. It is not the case that an individual polygynous couple would save less than a monogamous couple living in the same country. Rather, if a large fraction of households is polygynous, the equilibrium price of women is high, which changes incentives for all families. In other words, polygyny lowers output precisely because of the general equilibrium effects in the marriage market. We thus turn to the importance of marriage markets for growth in the next section.

A few papers attempt to understand why polygyny exists in some cultures and not in others. Gould et al. (2008) and Lagerlöf (2005) relate the disappearance of polygyny to economic development. Heterogeneity plays a key role in both papers. Gould, Moav, and Simhon argue that the increasing skill premium has led men to want fewer, higher quality children. To educate their children, they accordingly demand higher quality wives, but fewer of them, which naturally leads to fewer wives per men. Lagerlöf relates the disappearance of polygyny to the decline in male inequality over time. Primitive societies are arguably more unequal, which allows wealthy men to marry more wives and have more children. Over time, this dilutes their wealth, making societies more equal, which eventually leads to a more equal distribution of wives across men. In both papers, the decline in polygyny goes hand in hand with fertility decline and economic growth. Both papers explain the decline in polygyny prevalence, but are silent on the introduction of formal restrictions.

Two recent papers analyze the political economy of the introduction of monogamy. Lagerlöf (2010) proposes a theory related to inequality of wives across men. When polygyny is allowed, the elites have many wives, while poor men have none. This may lead to revolutions and thus creates an incentive for the elites to impose a formal ban on polygyny. de la Croix and Mariani (2015) provide a comprehensive political economy analysis of the switch from polygyny to monogamy and then to serial monogamy. The theory is based on the voting behavior of the entire population (including women), rather than the incentives of the elites. The transition between regimes is endogenously generated by human capital accumulation that changes the coalitions that stand to gain from a change in the marriage regime.

3.9 The Marriage Market

While there is a substantial literature on marriage choices within family economics, incorporating a marriage market into macroeconomic models is no trivial undertaking. One approach was proposed by Tertilt (2005), who models a competitive market for brides featuring an equilibrium brideprice that clears the market. However, such a formulation works only if there is no heterogeneity; if potential spouses vary in "quality," it matters who marries whom.

A number of recent contributions analyze marriage formation with heterogeneous agents within macro models. This allows the analysis of questions such as the impact of changes in the assortativeness of mating on income inequality. An early example is Fernández et al. (2005).[bt] The paper investigates the relationship between inequality, assortative mating, human capital accumulation, and per capita GDP. Mating is modeled through a search model with random matching. The model also features an intergenerational transmission mechanism, because parental income is used as collateral that children need when investing in education. One main finding is that such a model can generate multiple steady states that differ in wage inequality. Across steady states, marital sorting and wage inequality are positively related, while marital sorting and GDP per capita are negatively related.

Eika et al. (2014) document empirically the importance of assortative mating for income inequality in the United States. While assortative mating is found to be an important determinant of inequality, the study finds that changes in inequality cannot be attributed to changes in sorting patterns. Greenwood et al. (2016a) analyze such a link in a structural quantitative model.

Beyond these few contributions, the importance of marriage for growth is largely unexplored. In part, this may be due to the computational complexity of models that feature sorting with heterogeneous agents. However, with recent advances in computational power allowing increasingly complex models to be analyzed, we expect this to be an active research area in the near future.

4. THE FAMILY AND THE POLITICAL ECONOMY OF INSTITUTIONAL CHANGE

Long-run economic development is characterized not just by economic transformations but also by a set of striking regularities in terms of political change. During the development process, almost all of today's rich countries went through a series of similar policy reforms: for instance, democracy spread, public education systems were built, and public pension systems were introduced. The only exception to this pattern are countries that are rich primarily because of endowments with natural resources such as oil. Among countries who owe their wealth to the productivity of their citizens, these political transformations are a universal characteristic of the development process.

The tight link between economic and political transformations raises the question of how the causality runs between the two realms. Does economic growth trigger political

[bt] Fernández and Rogerson (2001), Choo and Siow (2006), and Greenwood et al. (2014, 2016a) also analyze the relationship between marital sorting and income inequality, but do not consider broader macroeconomic implications.

change, or is political change a precondition for growth? Can today's poor countries, many of which have implemented only a subset of the political reforms that characterize rich countries, foster faster economic development by adopting rich-country political institutions and reforms?

In this section, we argue that in answering such questions the family once again plays a central role. Many of the political reforms that go along with development are directly about the family (such as the introduction of child labor laws and the expansion of women's rights). In other cases (such as education and pension reforms), the political changes concern areas that originally were organized within families but in which, over time, the state played an increasing role. We provide a brief overview of the facts of political change during the development process. We then discuss some of the political economy literature analyzing the causes and consequences of political change, arguing that in many cases changes in family life were driving reform. We illustrate the role of the family by zooming in on two specific reforms—the expansion of women's rights and the introduction of child labor laws.

4.1 Political Economy Facts

The main political transformations that go along with the development process are the introduction of democracy, public and compulsory schooling, and child labor regulation; the gradual expansion of women's rights; and more generally the creation of large welfare states that raise a significant fraction of GDP in tax revenue to provide welfare benefits and old-age pensions. Before the onset of modern economic growth (say, in 1750), no country in the world had any of these institutions. Most poor countries today have some but not all of these features.

There is considerable variation across countries in the timing of reforms. For some countries, the first transformation was the introduction of democracy, starting with the founding of the United States in 1776 and then followed by a series of franchise extensions in Britain. Other countries adopted other reforms first and achieved democracy later. Some European countries democratized after World War I, and others had to wait until after the fall of the Iron Curtain in the early 1990s. In some countries (such as South Korea and Taiwan), democracy was introduced only after most other political reforms had been implemented and after rapid economic growth had been achieved.

Initially, democracy generally meant that men, but not women, obtained the right to vote and run for office. In the United States, the first state to give women the right to vote was Wyoming in 1869, and most other states had followed by World War I.[bu] At the federal level, universal suffrage was introduced with the Nineteenth Amendment in

[bu] See Doepke et al. (2012) for a detailed timeline of the introduction of women's rights in the United States.

1920. In many European countries women were able to vote after World War I, but once again there is a lot of variation across countries. For example, in Switzerland women received the right to vote in federal elections only in 1972, and the last canton to allow women to vote was Appenzell Innerrhoden in 1990.[bv]

Compared to the spread of political rights, the timing of education reforms is more uniform across countries. In the United States, Canada, and the industrializing Western European countries, public and compulsory education was widely introduced in the late 19th and early 20th centuries. In many cases, these reforms went along with significant restrictions of child labor.

The first country to introduce a public pension system was Germany in 1891. Mandatory health and accident insurance for workers were introduced around the same time. Most other European countries, Canada, and the United States had followed these steps before the middle of the 20th century. The first unemployment benefit scheme was introduced in the United Kingdom with the National Insurance Act 1911. In the midst of the Great Depression, the US Congress passed the Social Security Act, which contained provisions for old age insurance, welfare, and unemployment insurance. Most European countries and Canada introduced similar provisions during the first half of the 20th century.

The timing of political reforms that affected families most directly (in particular the regulation of child labor, the public provision of education, and the spread of women's rights) is closely associated with a major transformation of families themselves. As discussed in Section 3, as countries transition from a preindustrial society to modern growth, they universally undergo a demographic transition from high to low fertility. In North America and Western Europe, the main phase of fertility decline took place between the middle of the 19th century and World War I. Access to primary education became near-universal during the same period. Given that formal schooling moved children from the family home (where many had been working from a young age) to schools, the rise of mass education implied a transformation of family life on its own.

4.2 The Family as a Driver of Political Change

To understand the political economy of reforms, one needs to understand who the winners and losers of a reform are. Political reforms happen if there is a constituency that stands to gain from the reform, and if this constituency has sufficient political power to implement the desired policy. The trigger for a reform can either be a change in how a policy affects specific groups, or an increase in the political power of a group that

[bv] In fact, the last canton to voluntarily introduce the right to vote for women was Appenzell Aussenrhoden in 1989. In Appenzell Innerrhoden women's suffrage was mandated by a Supreme Court decision in 1990.

stands to gain from a reform. One might expect that democratization, which increased the political power of broad parts of the population at the expense of established elites, should be a major engine for political change. While there are examples of democratization triggering reform, the introduction of the major reforms associated with economic development described above is not closely correlated with expansions in political rights. We therefore focus on mechanisms that change who gains and who loses from reforms, and take as given that the relevant groups have sufficient political power to be heard.[bw]

We argue that for most of the major political reforms associated with economic development, the reorganization of families is a key reason for why political incentives changed. Technological and structural change affects fertility choices, education choices, and the division of labor in the family, all of which determine how people are affected by reforms. For example, reforms such as mandatory schooling laws and public pensions move responsibilities from the family to the public sphere and affect the relationship between parents and children. How people feel about such changes will depend in part on how many children they have, on whether they plan to educate their children, and on whether they anticipate living with their children in old age. Other reforms—such as the expansion of women's rights—affect the interaction between spouses. How people are affected by such reforms depend in part on the division of labor in the household and on women's labor force participation, both of which vary with development.

Consider the introduction of public schooling systems. Before public schooling, most children were working with their parents from a young age. Hence, the spread of public and compulsory education implied a major change of parent-child relations. Galor and Moav (2006) provide a theory that explains the public provision of education as a consequence of the rising importance of human capital in the economy. They consider a model economy populated by capitalists and workers. The model features heterogeneity in wealth, and initially only capitalists are accumulating capital through bequests to their children. However, the model features complementary between physical and human capital, and as the stock of physical capital rises, over time the capitalists stand to gain from higher education among the workers. Ultimately, both workers and capitalists support a tax on capitalists to support public education. The accumulation of physical and human capital within families is central to this mechanism. The public provision of schooling was often followed by mandatory schooling laws. Such laws affect the family even more directly by forcing parents to send their children to school. A closely related policy is a child labor ban, which we will analyze in Section 4.4.

In the case of schooling and child labor bans, who is a winner and who is a loser from reform depends on people's factor endowments (physical capital and human capital) and

[bw] Key contributions examining the causes of expansions of political rights include Acemoglu and Robinson (2000) and Lizzeri and Persico (2004).

also on fertility. Thus, potential conflicts arise between capitalists and skilled workers on the one hand, and unskilled workers with large families and no desire to educate their children on the other hand. For other types of reforms, gender and marital status are the dividing lines. This point is emphasized in Edlund and Pande (2002), who analyze the importance of women as voters. The paper shows that the political gender gap in the United States—women are more likely to vote Democrat than men—is a relatively recent phenomenon. Up until the mid-1960s, women voted more conservative than men on average. The paper argues that the change in political preferences (which in turn may have impacted other reforms) was due to a specific change in the family, namely the increase in divorce. A large increase in divorce rates during the 1960s and 1970s (see Fig. 6) increased the fraction of relatively poor single women. These women tend to benefit from redistribution, which is typically favored by Democrats. The paper provides evidence in support of the hypothesis by showing that marriage tends to make a woman more Republican, while divorce tends to make her more Democrat.

There are also a few papers that emphasize the importance of women as policymakers. Chattopadhyay and Duflo (2004) use gender quotas in India to empirically analyze which public projects are implemented at the village level depending on the gender of the leader. While the paper is not specifically about reforms, it shows that the gender of the leader affects the types of public goods that are provided. A related point is made by Washington (2008) and Oswald and Powdthavee (2010), who show that the gender composition of children affects the voting behavior of (male) legislators in both the United States and the United Kingdom: having more daughters makes politicians take more liberal positions.

Another important reform is the introduction of public pension systems.[bx] Social security programs transfer resources from young and middle-aged workers to the elderly. Without public systems, such transfers typically happen within the family, with altruistic children voluntarily taking care of elderly parents. Because of the dramatic fertility decline during the 19th century (see Fig. 1), more people ended up without children caring for them during old age, increasing the risk of poverty. This fact probably played an important role in the introduction of public pension systems. At the same time, the existence of such systems further decreases the incentive to have children, which leads to a two-way interaction between the structure of the family and political reforms.

Finally, a large class of reforms affected the legal position of women. These include reforms affecting ownership rights of women (such as the Married Women's Property Act of 1870 in England), reforms affecting child custody laws, the introduction of suffrage for women, and laws banning labor market discrimination and removing occupational

[bx] There is a large literature on social security systems (see, for example, Cooley and Soares, 1996; Boldrin and Montes, 2005; Caucutt et al., 2013).

restrictions (such as allowing women to become judges and soldiers). Reforming the legal position of women also impacts the position of women in the household, eg, by changing their outside options. And conversely, changes in family structure (such as the decline in fertility and the increase in female labor force participation) affected the gains from such reforms. We will discuss the political economy of women's economic rights (such as married women's property rights) in Section 4.3. Other types of women's rights, such as suffrage or labor rights, imply different political economy trade-offs. While there is some empirical work on these other rights, there is a lack of work that formally analyzes the political economy of other types of rights for women.[by] We believe that this is an important issue to be addressed by future research.

4.3 Voting for Women's Rights

Throughout the course of development, all industrialized countries implemented reforms that changed the legal position of women. Doepke and Tertilt (2009) propose a mechanism that provides a causal link between women's rights and economic growth. The mechanism is based on women's role in nurturing children. In contrast, Geddes and Lueck (2002) argue that the initial expansion of women's rights was related to women's role in the labor market. Given that the main phase of expanding women's economic rights was in the 19th century, a time when female labor force participation was low, we argue that a mechanism related to a women's role in the family is more plausible.

We now illustrate the basic mechanism of Doepke and Tertilt (2009) in a simplified framework. The setup is similar to that in Section 3.4 with a modified utility function. We now assume that consumption is a private good, which allows for a stronger conflict of interest between husbands and wives. We also introduce grandchildren and assume that people derive utility from the human capital of children and grandchildren. This assumption introduces a conflict across generations: men want their grandchildren to have as much human capital as possible, but it is the next generation that makes the decision. Since the next generation also cares about their own consumption, fathers will not invest as much in their children's education as desired by the grandfathers. We will now show how this conflict across generations may induce men to vote for female empowerment.

Let the utility function of spouse of gender g be

$$\log(c_g) + \delta_g \log(H') + \delta_g^G \log(H'),$$

where δ_g is the weight spouse g attaches to the human capital of own children, while δ_g^G is the weight on grandchildren. As in Section 3.4, we assume that $\delta_f > \delta_m$.[bz] Given the private goods assumption, the budget constraint is

[by] See Duflo (2012) and Doepke et al. (2012) for two surveys.

[bz] While it may seem natural to assume the same for grandchildren, $\delta_f^G > \delta_m^G$, this assumption is not needed for the analysis.

$$c_m + c_f = A\ell H,$$

where ℓ is total working time of the couple. Assuming that each spouse has a time endowment of 1, the family time constraint is

$$\ell + 2e \leq 2,$$

where e is education time for each of two children.

We now consider two political regimes. In the first one—*patriarchy*—only men make decisions. In the second regime—*empowerment*—men and women make decisions jointly, ie, they solve a collective bargaining problem with equal weights. To find the equilibrium allocation under patriarchy, one can solve the following maximization problem:

$$\max_{\ell, e} \log(c_g) + \delta_g \log(H') + \delta_g^G \log(H')$$

subject to:

$$\ell + 2e \leq 2,$$

$$H' = (Be)^\theta,$$

$$c_m + c_f = A\ell H,$$

$$c_m, c_f \geq 0.$$

Note that $H'' = (Be')^\theta$, where e' is determined by the next generation and is taken as given by the grandparent. Given the technology, the choice of education for own children e will not affect H'', ie, there is no interdependence between the choices of different generations. Further, since a man does not derive utility from his wife's consumption, women's consumption will be zero, and hence male consumption equals production.[ca] The equilibrium allocation under patriarchy is:

$$e^P = \frac{\delta_m \theta}{1 + \delta_m \theta},$$

$$\ell^P = \frac{2}{1 + \delta_m \theta},$$

$$c_m^P = \frac{2AH}{1 + \delta_m \theta}.$$

[ca] This counterfactual result can be easily modified by introducing altruism, as we do in Doepke and Tertilt (2009).

In contrast, under empowerment, couples solve a joint maximization problem with equal bargaining weights. The objective function then is

$$\frac{1}{2}\log(c_m) + \frac{1}{2}\log(c_f) + \tilde{\delta}\log(H') + \tilde{\delta}^G\log(H''),$$

where $\tilde{\delta} = \frac{\delta_f + \delta_m}{2}$ and $\tilde{\delta}^G = \frac{\delta_f^G + \delta_m^G}{2}$. Given the objective function, women and men consume equal amounts, $c_f^E = c_m^E$. The optimal education and labor choices are:

$$e^E = \frac{\tilde{\delta}\theta}{1 + \tilde{\delta}\theta},$$

$$\ell^E = \frac{2}{1 + \tilde{\delta}\theta}.$$

Consumption is equalized and depends on the initial human capital:

$$c_m^E = c_f^E = \frac{AH}{1 + \tilde{\delta}\theta}.$$

We are interested in understanding under what conditions men prefer to live in a patriarchal world and when they prefer empowering women. We focus on men's preferences because women's economic rights were expanded long before women gained the right to vote. Hence, the expansion of women's right can be viewed as a voluntary sharing of power by men. To understand men's political preferences, we compare the indirect utility function of a man in both regimes starting from the same initial human capital. Denote the indirect utility functions by U^E and U^P. Plugging in the equilibrium allocations and simplifying, we see that $U^E > U^P$ if and only if:

$$(\delta_m + \delta_m^G)\theta\log\left(\frac{\tilde{\delta}}{\delta_m}\frac{1 + \delta_m\theta}{1 + \tilde{\delta}\theta}\right) > \log\left(\frac{2(1 + \tilde{\delta}\theta)}{1 + \delta_m\theta}\right). \tag{19}$$

From a man's perspective, there is a trade-off. Patriarchy implies strictly higher own consumption, since resources do not need to be shared with one's wife. On the other hand, from the grandfather's perspective, the son will underinvest in the education of the grandchild. Empowering women will lead the future daughter in law to have more bargaining power, and, given that women care more about children than men do ($\delta_f > \delta_m$), this will increase the education of the grandchildren.

We will now show how this trade-off changes with development. Assume that the human capital technology improves over time, ie, θ increases. When the returns to education are zero, ie, $\theta = 0$, men strictly prefer to live under patriarchy (this follows from

Eq. 19). The intuition is that with $\theta = 0$, there is no reason to educate children. With zero education, from a man's perspective empowering women imposes a cost in terms of lost consumption, but does not bring any benefits. However, as θ increases, the concern about the grandchildren's education becomes increasingly important. The next proposition shows that as long as the concern about grandchildren is above a threshold, when θ becomes large enough, the grandchild effect dominates and hence men gain from switching to the empowerment regime.

Proposition 5 *If the weight δ_m^G men attach to grandchildren is above a threshold (given in the proof), there is a threshold $\overline{\theta}$ such that men prefer empowerment if $\theta > \overline{\theta}$.*

Fig. 24 illustrates the result with a numerical example.[cb] The equilibrium education choice e increases with θ in both regimes. Initially, for low levels of θ, men prefer to live under patriarchy. However, as θ increases, patriarchy becomes too costly for men. By introducing women's empowerment, men gain because of the positive effect on grandchildren.

The result is in line with what was observed during the 19th century in both the United States and England. Primary education expanded rapidly at the same time when male legislators passed laws to grant property and other economic rights to married women. Fertility rates also decreased quickly and economic growth increased. These features can be incorporated by adding fertility choice and assuming that parental human capital is an input in children's human capital. In Doepke and Tertilt (2009), we analyze such an augmented model in a fully dynamic context. The main result of the model is also in line with cross-country data. Fig. 25 shows that the position

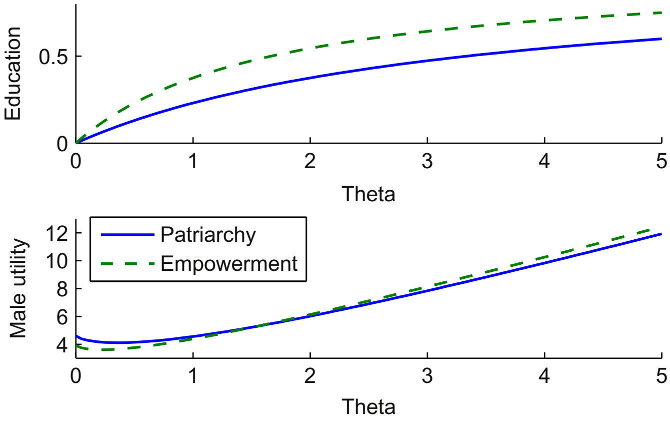


Fig. 24 Education and male utility as a function of θ, patriarchy vs empowerment.

[cb] The parameters used in the example are $\delta_m = 0.3, \delta_f = 0.9, \delta_m^G = 1.2, A = B = 5$. The initial level of human capital is set to $H_0 = 10$. The return to education θ varies between 0 and 5.

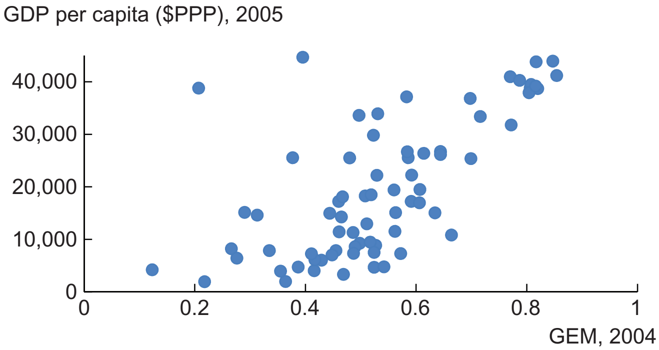


Fig. 25 Gender empowerment measure (GEM) and GDP per capita across countries. *GEM is an index constructed by the UN (Human Development Report, 2004), and GDP numbers are from the World Development Indicators.*

of women, as measured by the gender empowerment measure (GEM) constructed by the United Nations, is strongly positively correlated with GDP per capita. Assuming that returns to education differ systematically across countries, the model reproduces the same relationship.

A complementary theory is proposed by Fernández (2014). As in Doepke and Tertilt (2009), father's concern for their children is a central element. However, the key issue is not investment in education, but fathers preferring a more equal outcome between sons and daughters than what is produced under patriarchy. Economic growth widens disparities between sons and daughters in the patriarchy regime, which ultimately induces fathers to vote for empowerment. Fernández (2014) also provides empirical evidence based on the variation in extensions of women's economic rights across US states, showing that per capita wealth is positively associated with reform, whereas the association with fertility rates is negative (which is in line with the theories of both Doepke and Tertilt, 2009 and Fernández, 2014).

4.4 Voting for Children's Rights

Another near-universal policy reform associated with long-run development is the restriction of child labor. In preindustrial societies, child labor was the norm. In Western Europe and the United States, concern about child labor increased with industrialization, and ultimately industrializing countries introduced a variety of child labor restrictions such as minimum age laws and laws against working in hazardous occupations. A closely related policy reform that often coincided with child labor legislation is the introduction of compulsory schooling. This policy is usually the most effective constraint on child labor (in part because enforcement is straightforward). The close link between child labor and schooling is also part of the reason why child labor reforms matter for growth, as rising educational attainment is one engine of long-run development.

Whereas child labor bans are now in place in all industrial countries, in many developing countries child labor continues to be widespread. Child labor is especially common among poorer families who depend on the additional income. In these countries, public support for introducing restrictions is low.

What explains the passing of child labor reform in some countries, and persistent failure to do so in others? These questions are addressed in Doepke and Zilibotti (2005a), who present an analysis of the political economy of child labor legislation within a dynamic framework that endogenizes skill premia as well as fertility and education decisions.[cc] Here we use a simpler, static framework to highlight the main trade-offs. To understand the political support for and opposition to child labor laws, it is necessary to identify which groups stand to gain or lose from the introduction of regulation. Doepke and Zilibotti argue that the group that stands to gain most from banning child labor consists of unskilled adult workers. To the extent that these workers compete with children in the labor market, by banning child labor they can reduce competition and potentially raise their own wages.[cd] However, the situation is complicated by the fact that the same workers may also have working children themselves, so that the potential wage gains have to be traded off against the loss of child-labor income. A family's fertility and education choices therefore also matter.

To analyze these trade-offs more formally, consider an economy with N_S skilled and N_U unskilled workers. We start under the assumption that each worker has n children, but that only the children of the unskilled workers are working. This is consistent with the observation that child labor is generally more prevalent among poorer families, whereas richer, more highly educated families tend to send their children to school rather than to work. The production technology is:

$$Y = AX_S^{\alpha}X_U^{1-\alpha},$$

where X_S is skilled labor and X_U is unskilled labor. Each working child supplies λ units of unskilled labor, where $\lambda < 1$, reflecting that children are less productive than adult workers. If child labor is legal (the *laissez faire* policy), labor supply is given by:

$$X_S^{\text{laissez faire}} = N_S,$$

$$X_U^{\text{laissez faire}} = N_U + \lambda n N_U,$$

[cc] An analysis of the welfare implications of banning child labor is contained in Doepke and Krueger (2008).

[cd] The feedback from regulation to wages is also central to the seminal analysis of Basu and Van (1998), which focuses on the possibility of multiple equilibria.

and, under the assumption of competitive production, wages are given by:

$$w_S^{\text{laissez faire}} = A\alpha\left(\frac{(1+\lambda n)N_U}{N_S}\right)^{1-\alpha},$$

$$w_U^{\text{laissez faire}} = A(1-\alpha)\left(\frac{N_S}{(1+\lambda n)N_U}\right)^{\alpha}.$$

Workers seek to maximize their total income (ie, consumption). Adding adult and child-labor income, total family income for the two types of workers is given by:

$$I_S^{\text{laissez faire}} = w_S = A\alpha\left(\frac{(1+\lambda n)N_U}{N_S}\right)^{1-\alpha},$$

$$I_U^{\text{laissez faire}} = (1+\lambda n)w_U = (1+\lambda n)^{1-\alpha}A(1-\alpha)\left(\frac{N_S}{N_U}\right)^{\alpha}.$$

Let us now see who would gain or lose if child labor were to be banned. Under a child labor ban, no children are working, so that labor supply is simply $X_S^{\text{Ban}} = N_S$ and $X_U^{\text{Ban}} = N_U$, and wages are:

$$w_S^{\text{Ban}} = A\alpha\left(\frac{N_U}{N_S}\right)^{1-\alpha},$$

$$w_U^{\text{Ban}} = A(1-\alpha)\left(\frac{N_S}{N_U}\right)^{\alpha}.$$

The ratios of wages under the two policies are:

$$\frac{w_S^{\text{Ban}}}{w_S^{\text{laissez faire}}} = \left(\frac{1}{1+\lambda n}\right)^{1-\alpha} < 1,$$

$$\frac{w_U^{\text{Ban}}}{w_U^{\text{laissez faire}}} = (1+\lambda n)^{\alpha} > 1.$$

Thus, the skilled wage falls and the unskilled wage increases. This happens because child labor is a substitute for unskilled but a complement for skilled adult labor. The result suggests that unskilled workers may be in favor of banning child labor. However, this is no longer clear when we look at what happens to total family income:

$$I_S^{\text{Ban}} = w_S^{\text{Ban}} = A\alpha\left(\frac{N_U}{N_S}\right)^{1-\alpha},$$

$$I_U^{\text{Ban}} = w_U^{\text{Ban}} = A(1-\alpha)\left(\frac{N_S}{N_U}\right)^{\alpha}.$$

The income ratios are:

$$\frac{I_S^{\text{Ban}}}{I_S^{\text{laissez faire}}} = \left(\frac{1}{1+\lambda n}\right)^{1-\alpha} < 1,$$

$$\frac{I_U^{\text{Ban}}}{I_U^{\text{laissez faire}}} = \left(\frac{1}{1+\lambda n}\right)^{1-\alpha} < 1.$$

We see that, in fact, income falls for both groups, including the unskilled. The reason is that the unskilled workers' gain in terms of higher wages is more than offset by the loss of child labor income. Intuitively, the loss of child labor income is proportional to the total reduction in the supply of unskilled labor, whereas the increase in the unskilled wage is less than proportional to the decline in labor supply.

The analysis suggests that in a country where unskilled workers' children are working as well, public support for introducing child-labor restrictions should be low. The support for child labor restrictions should rise, however, if there is a group of unskilled workers whose children are not working (say, because they send their children to school). Assume that fraction s of unskilled workers send their children to school, while only fraction $(1 - s)$ has working children. The wages then become:

$$w_S^{\text{laissez faire}} = A\alpha\left(\frac{(1+\lambda(1-s)n)N_U}{N_S}\right)^{1-\alpha},$$

$$w_U^{\text{laissez faire}} = A(1-\alpha)\left(\frac{N_S}{(1+\lambda(1-s)n)N_U}\right)^{\alpha}$$

Income is now given by:

$$I_S^{\text{laissez faire}} = w_S = A\alpha\left(\frac{(1+\lambda(1-s)n)N_U}{N_S}\right)^{1-\alpha},$$

$$I_U^{\text{laissez faire}}(\text{working children}) = (1+\lambda n)w_U = (1+\lambda n)A(1-\alpha)\left(\frac{N_S}{(1+\lambda(1-s)n)N_U}\right)^{\alpha},$$

$$I_U^{\text{laissez faire}}(\text{children in school}) = w_U = A(1-\alpha)\left(\frac{N_S}{(1+\lambda(1-s)n)N_U}\right)^{\alpha}.$$

If child labor is now banned, incomes are:

$$I_S^{\text{Ban}} = w_S^{\text{Ban}} = A\alpha\left(\frac{N_U}{N_S}\right)^{1-\alpha},$$

$$I_U^{\text{Ban}}(\text{working children}) = I_U^{\text{Ban}}(\text{children in school}) = A(1-\alpha)\left(\frac{N_S}{N_U}\right)^{\alpha}.$$

Thus, for the unskilled workers with children in school, the introduction of a child labor ban unambiguously increases income. This result explains why child labor reform tends to happen in times when child labor is already declining for other reasons, such as an increased demand for human capital and a higher propensity among unskilled workers to send children to school. It is unskilled workers who do not depend on child labor themselves who should be the strongest advocates of reform.

Notice that the basic mechanism outlined so far is similar to our analysis of the political economy of women's rights in Section 4.3. First, technological change (not modeled explicitly here) increases the demand for human capital; next, the higher demand for human capital induces families to start educating their children; and finally, the families who now send their children to school become supporters of a child labor ban, triggering reform.

So far, we have focused on the case of a country in which child labor is initially legal. Our results show that as long as child labor is widespread among unskilled workers, support for introducing a child-labor ban will remain low. In cross-country data, we observe that differences in child-labor regulations are highly persistent over time, which suggests the existence of a status-quo bias. To examine whether such a bias can arise in our model, let us now consider the opposite situation of a country where a child labor ban is already in place. Are there any reasons why people might be more supportive of banning child labor if a child labor ban is already in place? As we will see, a status-quo bias can indeed arise in our theory, but only if fertility decisions are endogenous and depend on the current political regime.

We would like to find conditions under which the electorate would be willing to abandon an already existing child-labor ban. Consider first the case where fertility is independent of the policy, ie, every household continues to have n children as before. In this case, the trade-off that arises from abandoning an existing ban is exactly the reverse of the trade-off following from introducing a ban described above. In particular, if all unskilled households would actually send their children to work once the ban is abandoned, they would stand to gain from introducing child labor and abandoning the ban. In other words, the preferred policy is independent of the current policy, and a status-quo bias does not arise.

The situation is different, however, if the number of children depends on the current state of the law. It is a common observation that parents face a quantity–quality trade-off in their decisions on children: Parents who invest a lot in their children in terms of education tend to have fewer children than parents who send their children to work. We would therefore expect that once a child labor ban is in place (which effectively makes children more expensive), fertility would be lower. For concreteness, assume that fraction o of unskilled workers have already chosen their number of children under the assumption that the child-labor ban will stay in place, and that their fertility rate is $n^{\text{Ban}} < n$. The remaining families choose their family size later; in particular, if the ban is abandoned, they will optimally choose the larger fertility size n to maximize child labor income. What are now the relevant trade-offs? As above, in the presence of a ban, workers' incomes are $I_S^{\text{Ban}} = A\alpha(N_U/N_S)^{1-\alpha}$ and $I_U^{\text{Ban}} = A(1-\alpha)(N_S/N_U)^{\alpha}$, respectively. If the ban is now abandoned, income is:

$$I_S^{\text{laissez faire}} = A\alpha\left(\frac{(1+\lambda(on^{\text{Ban}}+(1-o)n))N_U}{N_S}\right)^{1-\alpha}$$

for the skilled,

$$I_U^{\text{laissez faire}}(\text{old}) = (1+\lambda n^{\text{Ban}})A(1-\alpha)\left(\frac{N_S}{(1+\lambda(on^{\text{Ban}}+(1-o)n))N_U}\right)^{\alpha}$$

for the "old" unskilled with small families, and:

$$I_U^{\text{laissez faire}}(\text{young}) = (1+\lambda n)A(1-\alpha)\left(\frac{N_S}{(1+\lambda(on^{\text{Ban}}+(1-o)n))N_U}\right)^{\alpha}$$

for the "young" unskilled with larger families. Comparing incomes, we can see that the old unskilled can now lose from the introduction of child labor. Their income ratio is:

$$\frac{I_U^{\text{laissez faire}}(\text{old})}{I_U^{\text{Ban}}(\text{old})} = \frac{1+\lambda n^{\text{Ban}}}{(1+\lambda(on^{\text{Ban}}+(1-o)n))^{\alpha}},$$

which is smaller than one if n^{Ban} is sufficiently small relative to n. These families made their low fertility choice under the assumption that child labor would not be an option. Given that they cannot change fertility ex-post, they have little to gain from making their own children work, but lose from the lower wages due to other families' children entering the labor force.

This mechanism induces policy persistence: Once a ban is in place, families start to make decisions that in the future increase political support for maintaining the ban. This mechanism can explain why differences in child labor and its regulations can be highly persistent across countries. In particular, the theory predicts that some countries can get locked into steady state equilibria featuring high fertility, high incidence of child labor, and little political support for the introduction of child labor regulation. In contrast, other countries with otherwise identical economic fundamentals have low fertility, no child labor, and widespread support for the ban of child labor.

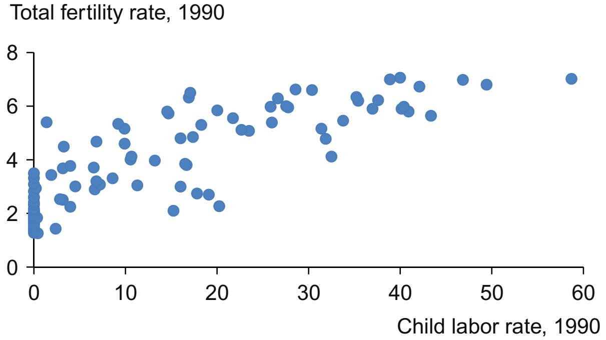


Fig. 26 The child labor rate (percentage of children aged 10–14 economically active) and total fertility rate across countries. *World Bank Development Indicators*.

Consistent with these predictions, we observe large cross-country differences in child labor rates, even among today's developing countries that are at similar levels of income per capita. The theory also predicts a positive correlation between fertility and child labor rates, even after controlling for other variables that might affect child labor or fertility. As Fig. 26 shows, there is a strong positive relationship between fertility rates and child labor rates across countries in contemporary data. Doepke and Zilibotti (2005a) examine the prediction more formally using an international panel of 125 countries from 1960 to 1990. They regress child labor rates on fertility rates, controlling for time dummies, GDP per capita, the Gini coefficient, and the share of agriculture in employment (arguably an independent factor affecting child labor) and find a positive and highly significant coefficient on the fertility rate, implying that a one standard deviation increase in fertility is associated with an increase in the child labor rate of 2.5 percentage points. The results are robust to the inclusion of country fixed effects.

The preceding analysis shows that the key feature of the political economy of child-labor regulation is that the group that most stands to gain from banning child labor (unskilled workers) is often simultaneously economically invested in child labor (because their own children are working). This observation leads to an explanation of why child labor was banned only after an increasing share of parents sent their children to school instead of work, and why differences in child labor and child-labor regulation across countries can be highly persistent over time. The analysis can also be used to help in designing policies that facilitate the passing of child labor regulations in developing countries today. Doepke and Zilibotti (2009, 2010) examine interventions such as international labor standards and trade restrictions aimed at reducing child labor from this perspective and argue that such well-intentioned policies can backfire and reduce the likelihood of comprehensive action of child labor within developing countries.

5. CONCLUSION

In this chapter, we have argued that accounting for the family should be an integral part of macroeconomics. The family is where many of the key decisions that are relevant for macroeconomics are made. Since families have been changing, with fewer marriages, more single households, lower fertility, and higher female labor supply, the answers to standard macroeconomic questions concerning, say, how labor supply and savings react to the business cycle have likely changed, too. Family structure also differs across countries. Developing countries are characterized by higher fertility, more traditional gender roles, often a son preference, and sometimes polygyny. These differences matter for the decisions that families make, and hence for the size and age structure of the population, for the accumulation of human and physical capital, and ultimately for the rate of economic growth.

The family matters not just for its role in household-level decisions but also through its effect on the evolution of institutions. Long-run economic development is characterized by a strikingly universal process of political change. Almost all of today's rich countries went through a series of similar reforms: democracy spread, public education systems were built, women and children gained rights, and public pension systems and the welfare state were introduced. We argue that many of these reforms transfer responsibility from the household to the public sphere, and that the ultimate triggers behind the reforms were often related to changes in the family.

There are additional ways in which the family matters for macroeconomics which we did not cover in this chapter. For example, the issues we discussed here are largely positive in nature. We touched only briefly on normative questions in a few places, for example, the discussion of the one child policy. We purposely did not talk about efficiency in this context, since this is not straightforward to do. The regular notion of Pareto efficiency is not defined in models where population size is endogenous, which includes all models with endogenous fertility. To evaluate policies that may affect fertility—such as education policies, child labor laws, policies banning abortion, or subsidies for single mothers—new concepts are required. Golosov et al. (2007) propose two new notions—$\mathcal{A}$- and $\mathcal{P}$-efficiency—and show how they can be used in standard fertility models. Schoonbroodt and Tertilt (2014) use the concepts to explore under what conditions fertility choice may be inefficiently low and hence pronatalist policies may be desired.

There is also a burgeoning literature on the role of the family for the transmission of preferences, cultural values, and attitudes, which can also feed back into macroeconomic outcomes. Theoretical models of the transmission of preferences and values in the family are developed by Bisin and Verdier (2001) and Doepke and Zilibotti (2005b, 2008). Empirical evidence for the intergenerational transmission of risk attitudes is provided by Dohmen et al. (2012). In Fernández et al. (2004), men's preferences for working vs stay-at-home wives are formed in childhood by the work behavior of their mothers. This leads to a dynamic process affecting female labor supply over time. Cultural transmission may also occur in society more generally. For example, in Fogli and Veldkamp (2011) and Fernández (2013), women learn from others about the costs of working. Both papers argue that a reduction in the perceived cost of working through this learning process is key to understanding the increase in female labor supply. The cultural transmission of fertility and female labor supply decisions is established empirically using data from second-generation immigrants to the United States by Fernández and Fogli (2006). Alesina and Giuliano (2010, 2014) argue that the strength of family ties varies across countries, and that these differences matter for cultural attitudes and macroeconomic outcomes. Alesina et al. (2013) take a historical perspective and trace unequal gender norms back to plough agriculture (and ultimately to soil type).[ce] Doepke and Zilibotti (2015) expand theories of preference transmission in the family to account for different parenting styles and link changes in parenting to macroeconomic trends such as increasing demand for human capital and increasing occupational differentiation in society.

Another important research area focuses on the importance of the family for understanding inequality. For example, de Nardi (2004) emphasizes the importance of bequest motives for the wealth distribution. Scholz and Seshadri (2009) build on this insight by investigating more generally the importance of children and fertility choice for the US wealth distribution. The interaction between parents and children is also analyzed for insights into the causes of intergenerational persistence of earnings.[cf] For example, parental inputs may amplify persistence if high-skill parents spend more resources and time on their children than low-skill parents. Other authors have emphasized the role of differences between women and men (and their interactions as couples) for understanding the distribution of earnings (and changes in earnings inequality over time). For example, Heathcote et al. (2010b) explicitly include male and female labor supply in their analysis of the US rising wage inequality. Other authors take this a step further and analyze how sorting and changes in sorting pattern have impacted inequality.[cg] Recent research also makes an explicit distinction between individual and household inequality.[ch] True consumption inequality may be lower than what is measured based on individual income data if the family plays a role in providing insurance (Blundell et al., 2008). Conversely, if family members do not provide full insurance

[ce] This hypothesis was first put forth by Boserup (1970), but had not been tested empirically until recently.

[cf] See, for example, Restuccia and Urrutia (2004), Lee and Seshadri (2015), and Yum (2015).

[cg] See, for example, Fernández and Rogerson (2001), Fernández et al. (2005), Choo and Siow (2006), and Greenwood et al. (2014, 2016a).

[ch] See Heathcote et al. (2010a).

to each other, true consumption inequality may be higher than what is measured based on household expenditure data (Lise and Seitz, 2011). Further, the mapping between individual and household inequality may change over time if the structure of the family is changing.

In our view, the intersection of family economics and macroeconomics offers many promising avenues for future research. Throughout this chapter, we have pointed out a number of particular questions that are in need of answers, and which could be addressed with the data, models, and methods available today. There is also a need to push out the frontier of theoretical modeling; in particular, we see a strong potential for intellectual arbitrage by applying methods of dynamic modeling that are common in macroeconomics to better understand the dynamics of household bargaining under commitment and private information frictions. Finally, there are promising applied topics that have barely been explored yet. For example, an important topic in recent macroeconomics concerns house price dynamics (see the chapter "Housing and macroeconomics" by Piazzesi and Schneider). Changes in family structure—such as the rise in single households—have a direct impact on housing demand. Further, singles are more eager to live in cities (where they can meet other singles) compared to families, who place higher value on space. Hence, changes in family formation and family structure should matter for the housing market. We hope that this and other research topics will be picked up by more researchers as family economics continues to become an integral part of macroeconomics.

APPENDICES

A Proofs for Propositions

Proof of Proposition 1 *As y_f approaches zero, the density $f(\eta_g) = F'(\eta_g)$ approaches zero, so that the elasticity of labor supply approaches zero also. In contrast, for married women, the fact that $w_m > 0$ guarantees that the elasticity of labor supply is bounded away from zero.* □

Proof of Proposition 2 *The first part follows from the fact that aggregate labor supply elasticity for single households equals one, whereas for married households, it is strictly smaller than one.*

For the second part, for any $w_f > 0$, we have $\hat{\eta}_u > \hat{\eta}_e$, which implies that the elasticity is smaller than one. As w_f converges to zero, $\hat{\eta}_e$ and $\hat{\eta}_u$ both converge to zero. Since $F(0) = 0$ and F is continuous, we then have that $F(\hat{\eta}_e)$ and $F(\hat{\eta}_u)$ both converge to zero, which implies that the elasticity of labor supply converges to one. Conversely, as w_f converges to infinity, $\hat{\eta}_e$ and $\hat{\eta}_u$ both converge to infinity, implying that $F(\hat{\eta}_e)$ and $F(\hat{\eta}_u)$ both converge to one and once again resulting in an elasticity of one. □

Proof of Proposition 3 *If $\widetilde{\lambda}_f \leq \lambda_f$ and $\widetilde{\lambda}_m \leq \lambda_m$, neither participation constraint (9) and (10) is binding. Hence, it is optimal to stay married, $D = 0$, and the consumption allocation follows from maximizing (7) subject to the budget constraint (8). If $\widetilde{\lambda}_f > \lambda_f$ and $\widetilde{\lambda}_f + \widetilde{\lambda}_m \leq 1$, the wife's* *participation constraint is binding. Staying married* ($D = 0$) *continues to be optimal, however, because it is possible to increase the wife's consumption share to make her indifferent between marriage and divorce, with the husband continuing to be better off married. The wife's consumption can then be solved from solving for* c_f *in her participation constraint (9) (imposed as an equality) while setting* $D = 0$. *The husband's consumption then follows from the budget constraint (8). The case where the husband's participation constraint is binding is parallel. Finally, when there is no allocation of ex-post bargaining power that keeps both spouses at least as well off married compared to being divorced, divorce* ($D = 1$) *is the optimal choice, and consumption follows from the individual budget constraints in the divorced state.* □

Proof of Proposition 4 *The ratio of the growth factors is*

$$\frac{1+g^{end}}{1+g^{exog}} = \left\{ \left(\frac{\alpha\delta_f + (1-\alpha)\delta_m}{\lambda_f\delta_f + (1-\lambda_f)\delta_m} \right) \left(\frac{\alpha + [\lambda_f\delta_f + (1-\lambda_f)\delta_m]\theta}{\alpha + [\alpha\delta_f + (1-\alpha)\delta_m]\theta} \right) \right\}^{\theta}.$$

Thus the result follows trivially, given the assumption $\delta_f > \delta_m$ *and* $\theta < 1$. □

Proof of Proposition 5 *Take the limit as* $\theta \to \infty$ *on both sides of Eq. (19) separately. The limit of the left-hand side can be written as:*

$$\lim_{\theta\to\infty} (\delta_m + \delta_m^G) \lim_{\theta\to\infty} \frac{\log\left(\frac{\tilde{\delta}}{\delta_m} \frac{1+\delta_m\theta}{1+\tilde{\delta}\theta} \right)}{\frac{1}{\theta}}.$$

Note that both numerator and denominator converge to zero. Applying L'Hopital's Rule, canceling terms and rearranging, the limit can be written as:

$$(\delta_m + \delta_m^G) \lim_{\theta\to\infty} \left(\frac{(\tilde{\delta} - \delta_m)}{\left(\frac{1}{\theta} + \tilde{\delta}\right)\left(\frac{1}{\theta} + \delta_m\right)} \right).$$

From this expression, we can see that the limit exits and is equal to:

$$(\delta_m + \delta_m^G) \left(\frac{\tilde{\delta} - \delta_m}{\tilde{\delta}\delta_m} \right).$$

The limit of the right hand side of (19) is $\log\left(\frac{2\tilde{\delta}}{\delta_m}\right)$. *Thus, in the limit* $U^E > U^P$ *if and only if* $(\delta_m + \delta_m^G)\left(\frac{\tilde{\delta} - \delta_m}{\tilde{\delta}\delta_m}\right) > \log\left(\frac{2\tilde{\delta}}{\delta_m}\right)$. *Using the definition of* $\tilde{\delta}$ *and rearranging, this can be expressed as:*

$$\delta_m^G > \log\left(\frac{\delta_f + \delta_m}{\delta_m}\right)\left(\frac{(\delta_f + \delta_m)\delta_m}{\delta_f - \delta_m}\right) - \delta_m.$$

Hence, as long as δ_m^G *is large enough, the equation is satisfied.* □

B Data Definitions and Sources

The data used in Table 3 are from two different editions of the OECD Gender, Institutions and Development Data Base (GID 2006 and 2014), the World Development Indicators (WDI 2003, 2005, and WDI 2014) and the United Nations Human Development Report 2004. Here we give the definition of each variable and its source.

GDP per capita: GDP data were used from two different years. The variables from GID 2014 and WDI 2014 were correlated with GDP p.c. from the WDI 2014. The variables from WDI 2003, UN 2004, and GID 2006 were correlated with GDP p.c. from the WDI 2005.

Share of agriculture: Measured as the value-added share of agriculture in GDP. Data were used from two different years. The variables from GID 2014 and WDI 2014 were correlated with percent agriculture from the WDI 2014. The variables from WDI 2003, UN 2004, and GID 2006 were correlated with percent agriculture from the WDI 2005.

Total fertility rate: Source: GID 2006.

Child mortality rate: Under-five mortality rate. Source: WDI 2014.

Average years of schooling: Source: WDI 2003.

Boy/girl sex ratio at birth: Measured as boys born per girl. Source: GID 2006.

Son preference in education: Percentage of people agreeing that university is more important for boys than for girls. GID 2014.

Inheritance discrimination against daughters: Whether daughters have the same inheritance rights as sons. Reported in three categories between 0 ("equal") and 1 ("unequal"). Source: GID 2014.

Female literacy relative to male: Female literacy as percentage of male literacy. Source: GID 2006.

Percent females in paid labor force: Percentage of women among wage and salaried workers. Source: GID 2006.

Unpaid care work by women: Female to male ratio of time devoted to unpaid care work. Source: GID 2014.

Year first woman in parliament: Source: Human Development Report 2004.

Women's access to land: Whether women and men have equal and secure access to land use, control and ownership. Categorical (three categories = 0, 0.5, 1), where 1 ("full") and 0 ("impossible"). Source: GID 2014.

Gender empowerment measure: Measures inequality between men's and women's opportunities, combining measures of inequality in political participation and decision making, in economic participation and decision making, and in power over economic resources. The level is between 1 ("full equality") and 0 ("no equality"). Source: UN 2004.

Early marriage: Share of female population between ages 15 and 19 ever married. GID 2014.

Agreement with wife beating: Percentage of women who agree that a husband/partner is justified in beating his wife/partner under certain circumstances. Source: GID 2014.

Inheritance discrimination against widows: Whether a widow has the same inheritance rights as a widower. Reported in three categories (0, 0.5, 1), where 0 means equal rights. Source: GID 2014.

Laws on domestic violence: Whether the legal framework offers women legal protection from domestic violence. Reported in five categories = 0, 0.25, 0.5, 0.75, 1, where 1 means no protection and 0 full protection. Source: GID 2014.

ACKNOWLEDGMENTS

We thank Stefania Albanesi, Raquel Fernández, Moshe Hazan, Claudia Olivetti, Víctor Ríos-Rull, John Taylor, Harald Uhlig, and Alessandra Voena for comments that helped to greatly improve the manuscript. Titan Alon, Florian Exler, Xue Zhang, Tim Hildebrand, Katarina Kuske, and Clara Schäper provided excellent research assistance. Financial support from the National Science Foundation (Grant SES-1260961) and the European Research Council (Grant SH1-313719) is gratefully acknowledged

REFERENCES

Acemoglu, D., Robinson, J.A., 2000. Why did the west extend the Franchise? Democracy, inequality and growth in historical perspective. Q. J. Econ. 115 (4), 1167–1199.

Adda, J., Dustmann, C., Stevens, K., 2016. The career costs of children. J. Polit. Econ., forthcoming.

Aguiar, M., Hurst, E., 2007. Measuring trends in leisure: the allocation of time over five decades. Q. J. Econ 122 (3), 969–1006 (Table II).

Aiyagari, S.R., 1994. Uninsured idiosyncratic risk and aggregate saving. Q. J. Econ. 109 (3), 659–684.

Aizer, A., Cunha, F., 2012. The production of child human capital: endowments, investments and fertility. NBER Working Paper No. 18429.

Albanesi, S., 2014. Jobless recoveries and gender biased technological change. Unpublished Manuscript, Federal Reserve Bank of New York.

Albanesi, S., Olivetti, C., 2009. Home production, market production, and the gender wage gap: Incentives and expectations. Rev. Econ. Dyn. 12 (1), 80–107.

Albanesi, S., Olivetti, C., 2014. Maternal health and the baby boom. Quant.Econ. 5 (2), 225–269.

Albanesi, S., Olivetti, C., 2016. Gender roles and medical progress. J. Polit. Econ. 124, 650–695.

Albanesi, S., Şahin, A., 2013. The gender unemployment gap. Federal Reserve Bank of New York Staff Report 613.

Alderman, H., Chiappori, P.A., Haddad, L., Hoddinott, J., Kanbur, R., 1995. Unitary versus collective models of the household: is it time to shift the burden of proof? World Bank Research Observer 10 (1), 1–19. ISSN 0257-3032, 1564-6971. http://www.jstor.org/stable/3986564.

Alesina, A., Giuliano, P., 2010. The power of the family. J. Econ. Growth 15 (2), 93–125.

Alesina, A., Giuliano, P., 2014. Family ties. In: Aghion, P., Durlauf, S.N. (Eds.), Handbook of Economic Growth, vol. 2A. North Holland, Amsterdam, Netherlands, pp. 177–215.

Alesina, A., Giuliano, P., Nunn, N., 2013. On the origins of gender roles: women and the plough. Q. J. Econ. 128 (2), 469–530.

Alger, I., Cox, D., 2013. The evolution of altruistic preferences: Mothers versus fathers. Rev. Econ. Household 11 (3), 421–446.

Ambler, K., 2015. Don't tell on me: experimental evidence of asymmetric information in transnational households. J. Dev. Econ. 113, 52–69.

Anderson, S., Baland, J.M., 2002. The economics of roscas and intrahousehold resource allocation. Q. J. Econ. 117 (3), 963–995. ISSN 0033-5533. http://www.jstor.org/stable/4132493.

Anukriti, S., 2014. Financial incentives and the fertility-sex ratio trade-off. Unpublished Manuscript, Boston College.

Ashraf, N., 2009. Spousal control and intra-household decision making: an experimental study in the Philippines. Am. Econ. Rev. 99 (4), 1245–1277. http://dx.doi.org/10.1257/aer.99.4.1245. http://www.ingentaconnect.com/content/aea/aer/2009/00000099/00000004/art00008.

Ashraf, N., Field, E., Lee, J., 2014. Household bargaining and excess fertility: an experimental study in Zambia. Am. Econ. Rev. 104 (7), 2210–2237.

Atkeson, A., Lucas Jr., R.E., 1992. On efficient distribution with private information. Rev. Econe Stud. 59 (3), 427–453.

Attanasio, O.P., Browning, M., 1995. Consumption over the life cycle and over the business cycle. Am. Econ. Rev. 85 (5), 1118–1137. ISSN 0002-8282. http://www.jstor.org/stable/2950978.

Attanasio, O., Lechene, V., 2002. Tests of income pooling in household decisions. Rev. Econ. Dyn. 5 (4), 720–748. http://dx.doi.org/10.1006/redy.2002.0191. http://www.sciencedirect.com/science/article/B6WWT-473VNCJ-2/2/cf9affb1fba57fc0d782dbe93c7753db. ISSN 1094-2025.

Attanasio, O., Lechene, V., 2014. Efficient responses to targeted cash transfers. J. Polit. Econ. 122 (1), 178–222.

Attanasio, O., Low, H., Sánchez-Marcos, V., 2005. Female labor supply as insurance against idiosyncratic risk. J. Eur. Econ. Assoc. 3 (2/3), 755–764. http://www.jstor.org/stable/40005017. ISSN 1542-4766.

Attanasio, O., Low, H., Sánchez-Marcos, V., 2008. Explaining changes in female labor supply in a life-cycle model. Am. Econ. Rev. 98 (4), 1517–1552.

Attanasio, O., Levell, P., Low, H., Sánchez-Marcos, V., 2015. Aggregating elasticities: intensive and extensive margins of female labour supply. NBER Working Paper No. 21315.

Bachmann, R., 2012. Understanding the jobless recoveries after 1991 and 2001. Unpublished Manuscript, University of Notre Dame.

Banerjee, A., Meng, X., Porzio, T., Qian, N., 2014. Aggregate fertility and household savings: a general equilibrium analysis with micro data. NBER Working Paper No. 20050.

Barro, R.J., Becker, G.S., 1989. Fertility choice in a model of economic growth. Econometrica 57 (2), 481–501.

Basu, K., 2006. Gender and say: a model of household behaviour with endogenously determined balance of power. Econ. J. 116 (511), 558–580.

Basu, K., Van, P.H., 1998. The economics of child labor. Am. Econ. Rev. 88 (3), 412–427.

Baudin, T., de la Croix, D., Gobbi, P.E., 2015. Fertility and childlessness in the United States. Am. Econ. Rev. 105 (6), 1852–1882. http://dx.doi.org/10.1257/aer.20120926.

Becker, G.S., 1988. Family economics and macro behavior. Am. Econ. Rev. 78 (1), 1–13.

Becker, G.S., Barro, R.J., 1988. A reformulation of the economic theory of fertility. Q. J. Econ. 103 (1), 1–25.

Becker, G.S., Murphy, K.M., Tamura, R., 1990. Human capital, fertility, and economic growth. J. Polit. Econ. 98 (5), 12–37.

Benhabib, J., Rogerson, R., Wright, R., 1991. Homework in macroeconomics: household production and aggregate fluctuations. J. Polit. Econ. 99 (6), 1166–1187.

Bick, A., 2016. The quantitative role of child care for female labor force participation and fertility. J. Eur. Econ. Assoc., 13 (3), 639–668.

Bick, A., Fuchs-Schündeln, N., 2014. Taxation and labor supply of married couples across countries: a macroeconomic analysis. Unpublished Manuscript, Arizona State University.

Bisin, A., Verdier, T., 2001. The economics of cultural transmission and the dynamics of preferences. J. Econ. Theor. 97 (2), 298–319.

Blau, F., Kahn, L., 2000. Gender differences in pay. J. Econ. Perspect. 14 (4), 75–99.

Blau, F., Kahn, L., 2007. Changes in the labor supply behavior of married women: 1980–2000. J. Labor Econ. 25 (3), 393–438.

Bleakley, H., Lange, F., 2009. Chronic disease burden and the interaction of education, fertility, and growth. Rev. Econ. Stat. 91 (1), 52–65.

Blundell, R., MaCurdy, T., 1999. Labor supply: a review of alternative approaches. In: Ashenfelter, O., Card, D. (Eds.), Chapter 27 of Handbook of Labor Economics, vol. 3. Elsevier, Amsterdam, Netherlands.

Blundell, R., Chiappori, P.A., Meghir, C., 2005. Collective labor supply with children. J. Polit. Econ. 113 (6), 1277–1306.

Blundell, R., Pistaferri, L., Preston, I., 2008. Consumption inequality and partial insurance. Am. Econ. Rev. 98 (5), 1887–1921.

Blundell, R., Pistaferri, L., Saporta-Eksten, I., 2016. Consumption inequality and family labor supply. Am. Econ. Rev. 106 (2), 387–435.

Boldrin, M., Montes, A., 2005. The intergenerational state, education and pensions. Rev. Econ. Stud. 72 (3), 651–664.

Boserup, E., 1970. Women's Role in Economic Development. George Allen and Unwin Ltd, London.

Buera, F.J., Kaboski, J.P., Zhao, M.Q., 2013. The rise of services: the role of skills, scale, and female labor supply. NBER Working Paper No. 19372. http://www.nber.org/papers/w19372.

Castilla, C., Walker, T., 2013. Is ignorance bliss? The effect of asymmetric information between spouses on intra-household allocations. Am. Econ. Rev. 103 (3), 263–268.

Caucutt, E., Lochner, L., 2012. Early and late human capital investments, borrowing constraints, and the family. Unpublished Manuscript, University of Western Ontario.

Caucutt, E., Cooley, T.F., Guner, N., 2013. The farm, the city, and the emergence of social security. J. Econ. Growth 18 (1), 1–32.

Cervellati, M., Sunde, U., 2005. Human capital formation, life expectancy, and the process of development. Am. Econ. Rev. 95 (5), 1653–1672.

Cesarini, D., Lindqvist, E., Notowidigdo, M.J., Östling, R., 2015. The effect of wealth on individual and household labor supply: evidence from Swedish lotteries. Unpublished Manuscript, Northwestern University.

Chade, H., Ventura, G., 2005. Income taxation and marital decisions. Rev. Econ. Dyn. 8 (3), 565–599. ISSN 1094-2025. http://dx.doi.org/10.1016/j.red.2005.01.008. http://www.sciencedirect.com/science/article/pii/S109420250500013X.

Chattopadhyay, R., Duflo, E., MIT, 2004. Women as policy makers: evidence from a randomized policy experiment in India. Econometrica 72 (5), 1409–1443. ISSN 0012-9682. http://search.ebscohost.com/login.aspx?direct=true&db=ecn&AN=0749671&site=ehost-live.

Chetty, R., Guren, A., Manoli, D., Weber, A., 2011. Are micro and macro labor supply elasticities consistent? A review of evidence on the intensive and extensive margins. Am. Econ. Rev. 101 (3), 471–475.

Chetty, R., Guren, A., Manoli, D., Weber, A., 2012. Does indivisible labor explain the difference between micro and macro elasticities? A meta-analysis of extensive margin elasticities. NBER Macroecon. Annu. 27, 1–56.

Chiappori, P.A., 1988. Rational household labor supply. Econometrica 56 (1), 63–90. ISSN 0012-9682. http://www.jstor.org/stable/1911842.

Chiappori, P.A., 1992. Collective labor supply and welfare. J. Polit. Econ.y 100 (3), 437–467. ISSN 0022-3808. http://www.jstor.org/stable/2138727.

Chiappori, P.A., Iyigun, M., Weiss, Y., 2009. Investment in schooling and the marriage market. Am. Econ. Rev. 99 (5), 1689–1713. http://dx.doi.org/10.1257/aer.99.5.1689.

Cho, J.O., Rogerson, R., 1988. Family labor supply and aggregate fluctuations. J. Monetary Econ. 21 (2–3), 233–245. ISSN 0304-3932. http://dx.doi.org/10.1016/0304-3932(88)90031-1.

Choo, E., Siow, A., 2006. Who marries whom and why. J. Polit. Econ. 114 (1), 175–201.

Choukhmane, T., Coeurdacier, N., Jin, K., 2014. The one child policy and household savings. Unpublished Manuscript, London School of Economics.

Clarke, S.C., 1995. Advance report of final marriage statistics, 1989 and 1990. Monthly Vital Stat. Rep. 43 (12), 3–5.

Coeurdacier, N., Guibaud, S., Jin, K., 2014. Fertility policies and social security reforms in China. IMF Econ. Rev. 62 (3), 371–408.

Cooley, T.F., Prescott, E.C., 1995. Economic growth and business cycles. In: Cooley, T.F. (Ed.), Frontiers of Business Cycle Research. Princeton University Press, Princeton.

Cooley, T.F., Soares, R.R., 1996. Will social security survive the baby boom? Carnegie-Rochester Conference Series Publ. Policy 45, 89–121.

Cordoba, J.C., 2015. Children, dynastic altruism and the wealth of nations. Rev. Econ. Dyn. 18 (4), 774–791.

Cordoba, J., Ripoll, M., 2014. The elasticity of intergenerational substitution, parental altruism, and fertility choice. Unpublished Manuscript, University of Pittsburgh.

Costa, D.L., 2000. From mill town to board room: the rise of women's paid labor. J. Econ. Perspect. 14 (4), 101–122.

Cubeddu, L., Ríos-Rull, J.V., 2003. Families as shocks. J. Eur. Econ. Assoc. 1 (2–3), 671–682.

Cunha, F., Heckmann, J., 2007. The technology of skill formation. Am. Econ. Rev. 97 (2), 31–47.

Da Rocha, J.M., Fuster, L., 2006. Why are fertility rates and female employment ratios positively correlated across OECD countries? Int. Econ. Rev. 47 (4), 1187–1222. ISSN 1468-2354. http://dx.doi.org/10.1111/j.1468-2354.2006.00410.x.

de la Croix, D., Doepke, M., 2003. Inequality and growth: why differential fertility matters. Am. Econ. Rev. 93 (4), 1091–1113.

de la Croix, D., Doepke, M., 2004. Public versus private education when differential fertility matters. J. Dev. Econ. 73 (2), 607–629.

de la Croix, D., Doepke, M., 2009. To segregate or to integrate: education politics and democracy. Rev. Econ. Stud. 76 (2), 597–628.

de la Croix, D., Mariani, F., 2015. From polygyny to serial monogamy: a unified theory of marriage institutions. Rev. Econ. Stud. 82 (2), 565–607.

de la Croix, D., Vander Donckt, M., 2010. Would empowering women initiate the demographic transition in least developed countries? J. Hum. Capital 4 (2), 85–129.

de la Croix, D., Doepke, M., Mokyr, J., 2016. Clans, guilds, and markets: apprenticeship institutions and growth in the pre-industrial economy. Unpublished Manuscript, Northwestern University.

de Laat, J., 2014. Household allocations and endogenous information: the case of split migrants in Kenya. J. Dev. Econ. 106, 108–117.

de Nardi, M., 2004. Wealth inequality and intergenerational links. Rev. Econ. Stud. 71 (3), 743–768.

Del Boca, D., Flinn, C., Wiswall, M., 2014. Household choices and child development. Rev. Econ. Stud. 81 (1), 137–185.

Doepke, M., 2004. Accounting for fertility decline during the transition to growth. J. Econ. Growth 9 (3), 347–383.

Doepke, M., Kindermann, F., 2015. Bargaining over babies: theory, evidence, and policy implications. Unpublished Manuscript, Northwestern University.

Doepke, M., Krueger, D., 2008. Origins and consequences of child labor restrictions: a macroeconomic perspective. In: Rupert, P. (Ed.), Frontiers of Family Economics. Emerald Press, Bingley, UK (England).

Doepke, M., Tertilt, M., 2009. Women's liberation: what's in it for men? Q. J. Econ. 124 (4), 1541–1591.

Doepke, M., Tertilt, M., 2014. Does female empowerment promote economic development? NBER Working Paper No. 19888.

Doepke, M., Tertilt, M., 2015. Asymmetric information in couples. Unpublished Manuscript, Northwestern University.

Doepke, M., Zilibotti, F., 2005a. The macroeconomics of child labor regulation. Am. Econ. Rev. 95 (5), 1492–1524.

Doepke, M., Zilibotti, F., 2005b. Social class and the spirit of capitalism. J. Eur. Econ. Assoc. 3 (2–3), 516–524.

Doepke, M., Zilibotti, F., 2008. Occupational choice and the spirit of capitalism. Q. J. Econ. 123 (2), 747–793.

Doepke, M., Zilibotti, F., 2009. International labor standards and the political economy of child labor regulation. J. Eur. Econ. Assoc. 7 (2–3), 508–518.

Doepke, M., Zilibotti, F., 2010. Do international labor standards contribute to the persistence of the child labor problem? J. Econ. Growth 15 (1), 1–31.

Doepke, M., Zilibotti, F., 2015. Parenting with style: altruism and paternalism in intergenerational preference transmission. Unpublished Manuscript, Northwestern University.

Doepke, M., Tertilt, M., Voena, A., 2012. The economics and politics of women's rights. Annu. Rev. Econ. 4, 339–372.

Doepke, M., Hazan, M., Maoz, Y.D., 2015. The Baby Boom and World War II: a macroeconomic analysis. Rev. Econ. Stud. 82 (3), 1031–1073.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., 2012. The intergenerational transmission of risk and trust attitudes. Rev. Econ. Stud. 79 (2), 645–677.

Du, Q., Wei, S.-J., 2010. A sexually unbalanced model of current account imbalances. NBER Working Paper No. 10498.

Duflo, E., 2012. Women empowerment and economic development. J. Econ. Liter. 50 (4), 1051–1079.

Duflo, E., Udry, C., 2004. Intrahousehold resource allocation in Cote d'ivoire: social norms, separate accounts and consumption choices. NBER Working Paper No. 10498. http://www.nber.org/papers/w10498.

Dyrda, S., Kaplan, G., Ríos-Rull, J.V., 2016. Business cycles and household formation. Unpublished Manuscript, University of Pennsylvania.

Ebenstein, A., 2010. The 'missing girls' of China and the unintended consequences of the one child policy. J. Hum. Resour. 45 (1), 87–115.

Echevarria, C., Merlo, A., 1999. Gender differences in education in a dynamic household bargaining model. Int. Econ. Rev. 40 (2), 265–286.

Eckstein, Z., Lifshitz, O., 2011. Dynamic female labor supply. Econometrica 79 (6), 1675–1726. ISSN 0012-9682, 1468-0262. http://www.jstor.org/stable/41336534.

Eckstein, Z., Lifshitz, O., 2015. Household interaction and the labor supply of married women. Int. Econ. Rev. 56 (2), 427–455. ISSN 0012-9682, 1468-0262. http://www.jstor.org/stable/41336534.

Edlund, L., Pande, R., Columbia, U., 2002. Why have women become left-wing? the political gender gap and the decline in marriage. Q. J. Econ. 117 (3), 917–961. ISSN 0033-5533. http://search.ebscohost.com/login.aspx?direct=true&db=ecn&AN=0620220&site=ehost-live.

Eika, L., Mogstad, M., Zafar, B., 2014. Educational assortative mating and household income inequality. NBER Working Paper No. 20271.

Erosa, A., Fuster, L., Restuccia, D., 2010. A general equilibrium analysis of parental leave policies. Rev. Econ. Dyn. 13 (4), 742–758.

Fernández, R., 2013. Cultural change as learning: the evolution of female labor force participation over a century. Am. Econ. Rev. 103 (1), 472–500.

Fernández, R., 2014. Women's rights and development. J. Econ. Growth 19 (1), 37–80.

Fernández, R., Fogli, A., 2006. Fertility: the role of culture and family experience. J. Eur. Econ. Assoc. 4 (2–3), 552–561.

Fernández, R., Rogerson, R., 2001. Sorting and long-run inequality. Q. J. Econ. 116 (4), 1305–1341.

Fernández, R., Wong, J.C., 2014a. Unilateral divorce, the decreasing gender gap, and married women's labor force participation. Am. Econ. Rev. 104 (5), 342–347. http://dx.doi.org/10.1257/aer.104.5.342.

Fernández, R., Wong, J.C., 2014b. Divorce risk, wages and working wives: a quantitative life-cycle analysis of female labour force participation. Econ. J. 124 (576), 319–358. ISSN 1468-0297. http://dx.doi.org/10.1111/ecoj.12136.

Fernández, R., Wong, J.C., 2014c. Free to leave? A welfare analysis of divorce regimes. Unpublished Manuscript, NYU. doi:10.1111/ecoj.12136.

Fernández, R., Fogli, A., Olivetti, C., 2004. Mothers and sons: preference formation and female labor force dynamics. Q. J. Econ. 119 (4), 1249–1299.

Fernández, R., Guner, N., Knowles, J., 2005. Love and money: a theoretical and empirical analysis of household sorting and inequality. Q. J. Econ. 120 (1), 273–344.

Fernández-Villaverde, J., Krueger, D., 2007. Consumption over the life cycle: facts from consumer expenditure survey data. Rev. Econ. Stat. 89 (3), 552–565. ISSN 0034-6535, 1530-9142. http://www.jstor.org/stable/40043048.

Fernández-Villaverde, J., Krueger, D., 2011. Consumption and saving over the life cycle: how important are consumer durables? Macroecon. Dyn. 15, 725–770. http://dx.doi.org/10.1017/S1365100510000180. ISSN 1469-8056.

Fogli, A., Veldkamp, L., 2011. Nature or nurture? Learning and the geography of female labor force participation. Econometrica 79 (4), 1103–1138.

Galí, J., Gambetti, L., 2009. On the sources of the great moderation. Am. Econ. J. Macroecon. 1 (1), 26–57.

Galor, O., Moav, O., 2006. Das Human-Kapital: a theory of the demise of the class structure. Rev. Econ. Stud. 73 (1), 85–117.

Galor, O., Weil, D.N., 1996. The gender gap, fertility, and growth. Am. Econ. Rev. 86 (3), 374–387.

Galor, O., Weil, D.N., 2000. Population, technology, and growth: from malthusian stagnation to the demographic transition and beyond. Am. Econ. Rev. 90 (4), 806–828.

Geddes, R., Lueck, D., 2002. The gains from self-ownership and the expansion of women's rights. Am. Econ. Rev. 92 (4), 1079–1092.

Goldin, C., 1990. Understanding the Gender Gap: An Economic History of American Women. Oxford University Press, Oxford.

Goldin, C., 1995. The u-shaped female labor force function in economic development and economic history. In: Schultz, T.P. (Ed.), Investment in Women's Human Capital and Economic Development. University of Chicago Press, Chicago, IL, USA, pp. 61–90.

Goldin, C., Olivetti, C., 2013. Shocking labor supply: a reassessment of the role of world war ii on women's labor supply. Am. Econ. Rev. 103 (3), 257–262. http://dx.doi.org/10.1257/aer.103.3.257.

Goldstein, M., Udry, C., 2008. The profits of power: land rights and agricultural investment in ghana. J. Polit. Econ. 116 (6), 981–1022.

Golosov, M., Tsyvinski, A., Werning, I., 2006. New dynamic public finance: a user's guide. In: Acemoglu, D., Rogoff, K., Woodford, M. (Eds.), NBER Macroeconomic Annual, vol. 21. MIT Press, Cambridge, MA, USA, pp. 317–388.

Golosov, M., Jones, L.E., Tertilt, M., 2007. Efficiency with endogenous population growth. Econometrica 75 (4), 1039–1071.

Gould, E., Moav, O., Simhon, A., 2008. The mystery of monogamy. Am. Econ. Rev. 98 (1), 333–357.

Greenwood, J., Hercowitz, Z., 1991. The allocation of capital and time over the business cycle. J. Polit. Econ. 99 (4), 1188–1214.

Greenwood, J., Seshadri, A., 2002. The U.S. demographic transition. Am. Econ. Rev. 92 (2), 153–159.

Greenwood, J., Rogerson, R., Wright, R., 1995. Household production in real business cycle theory. In: Cooley, T.F. (Ed.), Frontiers of Real Business Cycle Theory. Princeton University Press, Princeton.

Greenwood, J., Seshadri, A., Vandenbroucke, G., 2005a. The baby boom and baby bust. Am. Econ. Rev. 95 (1), 183–207.

Greenwood, J., Seshadri, A., Yorukoglu, M., 2005b. Engines of liberation. Rev. Econ. Stud. 72 (1), 109–133.

Greenwood, J., Guner, N., Kocharkov, G., Santos, C., 2014. Marry your like: assortative mating and income inequality. Am. Econ. Rev. 104 (5), 348–353.

Greenwood, J., Guner, N., Kocharkov, G., Santos, C., 2016a. Technology and the changing family: a unified model of marriage, divorce, educational attainment and married female labor-force participation. Am. Econ. J. Macroecon. 8 (1), 1–41.

Greenwood, J., Guner, N., Vandenbroucke, G., 2016b. Family economics writ large. J. Econ. Liter., forthcoming.

Groshen, E.L., Potter, S., 2003. Has structural change contributed to a jobless recovery? Curr. Issues Econ. Finan. Fed. Reserv. Bank N. Y. 9 (8), 1–7.

Guidolin, M., Jeunesse, E.A.L., 2007. The decline in the U.S. personal saving rate: is it real or is it a puzzle? Fed. Reserv. Bank St. Louis Rev. 89 (6), 491–514.

Guler, B., Guvenen, F., Violante, G., 2012. Joint-search theory: new opportunities and new frictions. J. Monetary Econ. 59 (4), 352–369.

Guner, N., Kaygusuz, R., Ventura, G., 2012a. Taxation and household labour supply. Rev. Econ. Stud. 79, 1113–1149.

Guner, N., Kaygusuz, R., Ventura, G., 2012b. Taxing women: a macroeconomic analysis. J. Monetary Econ. 59 (1), 111–128. http://dx.doi.org/10.1016/j.jmoneco.2011.10.004. http://www.sciencedirect.com/science/article/pii/S0304393211001036ISSN 0304-3932.

Guner, N., Kaygusuz, R., Ventura, G., 2014. Childcare subsidies and household labor supply. Unpublished Manuscript, University of Arizona.

Guvenen, F., Rendall, M., 2015. Women's emancipation through education: a macroeconomic analysis. Rev. Econ. Dyn. 18 (4), 931–956.

Guvenen, F., Kaplan, G., Song, J., 2014. The glass ceiling and the paper floor: gender differences among top earners, 1981–2012. NBER Working Paper No. 20560.

Hansen, G.D., Prescott, E.C., 2002. Malthus to solow. Am. Econ. Rev. 92 (4), 1205–1217.

Hazan, M., Zoabi, H., 2015a. Do highly educated women choose smaller families? Econ. J. 125 (587), 1191–1226.

Hazan, M., Zoabi, H., 2015b. Sons or daughters? Sex preferences and the reversal of the gender educational gap. J. Demograph. Econ. 81 (2), 179–201.

Hazan, M., Zoabi, H., 2006. Does longevity cause growth? A theoretical critique. J. Econ. Growth 11 (4), 363–376. http://www.jstor.org/stable/40216110. ISSN 1381-4338, 1573-7020.

Heathcote, J., Storesletten, K., Violante, G.L., 2009. Quantitative macroeconomics with heterogeneous households. Ann. Rev. Econ. 1 (1), 319–354. http://dx.doi.org/10.1146/annurev.economics.050708.142922.

Heathcote, J., Perri, F., Violante, G.L., 2010a. Unequal we stand: an empirical analysis of economic inequality in the United States, 1967–2006. Rev. Econ. Dyn. 13 (1), 15–51.

Heathcote, J., Storesletten, K., Violante, G.L., 2010b. The macroeconomic implications of rising wage inequality in the united states. J. Polit. Econ. 118 (4), 681–722.

Heckman, J.J., 2008. Schools, skills, and synapses. Econ. Inquiry 46 (3), 289–324.

Heim, B.T., 2007. The incredible shrinking elasticities: married female labor supply, 1978–2002. J. Hum. Resour. 42 (4), 881–918. http://www.jstor.org/stable/40057333. 0022-166X.

Hoel, J.B., 2015. Heterogeneous households: a within-subject test of asymmetric information between spouses in Kenya. J. Econ. Behav. Org. 118, 123–135.

Hong, J.H., Rios-Rull, J.V., 2012. Life insurance and household consumption. Am. Econ. Rev. 102 (7), 3701–3730.

Hsieh, C.T., Hurst, E., Jones, C.I., Klenow, P.J., 2013. The allocation of talent and U.S. economic growth. Unpublished Manuscript, Stanford University.

Humphries, J., Weisdorf, J., 2015. The wages of women in England, 1260–1850. J. Econ. Hist. 75 (2), 405–447.

Hyslop, D.R., 2001. Rising U.S. earnings inequality and family labor supply: the covariance structure of intrafamily earnings. Am. Econ. Rev. 91 (4), 755–777. http://dx.doi.org/10.1257/aer.91.4.755.

Iyigun, M., Walsh, R.P., 2007a. Building the family nest: premarital investments, marriage markets, and spousal allocations. Rev. Econ. Stud. 74 (2), 507–535.

Iyigun, M., Walsh, R.P., 2007b. Endogenous gender power, household labor supply, and the quantity-quality tradeoff. J. Dev. Econ. 82 (1), 138–155.

Jaimovich, N., Siu, H.E., 2009. The young, the old, and the restless: demographic and business cycle volatility. Am. Econ. Rev. 99 (3), 804–826.

Jaimovich, N., Siu, H.E., 2014. The trend is the cycle: job polarization and jobless recoveries. Unpublished Manuscript, USC.

Jaimovich, N., Pruitt, S., Siu, H.E., 2013. The demand for youth: explaining age differences in the volatility of hours. Am. Econ. Rev. 103 (7), 3022–3044.

Jayachandran, S., Kuziemko, I., 2011. Why do mothers breastfeed girls less than boys: evidence and implications for child health in India. Q. J. Econ. 126 (3), 1485–1538.

Jones, L.E., Schoonbroodt, A., 2015. Baby busts and baby booms: the fertility response to shocks in dynastic models. Unpublished Manuscript, University of Minnesota.

Jones, L.E., Tertilt, M., 2008. An economic history of the relationship between occupation and fertility—U.S. 1826–1960. In: Rupert, P. (Ed.), Frontiers of Family Economics, vol. 1. Emerald Group Publishing Limited, Bingley, UK.

Jones, L.E., Manuelli, R.E., McGrattan, E.R., 2015. Why are married women working so much? J. Demograph. Econ. 81 (1), 75–114.

Juhn, C., Potter, S., 2007. Is there still an added-worker effect? Federal Reserve Bank of New York Staff Report 310.

Kaplan, G., 2012. Moving back home: insurance against labor market risk. J. Polit. Econ. 120 (3), 446–512.

Karaivanov, A., Townsend, R.M., 2014. Dynamic financial constraints: distinguishing mechanism design from exogenously incomplete regimes. Econometrica 82 (3), 887–959. http://dx.doi.org/10.3982/ECTA9126. ISSN 1468-0262.

Keane, M., Rogerson, R., 2012. Micro and macro labor supply elasticities: a reassessment of conventional wisdom. J. Econ. Liter. 50 (2), 464–476.

Kinnan, C., 2014, December. Distinguishing barriers to insurance in Thai villages. Unpublished Manuscript, Northwestern University.

Knowles, J.A., 2013. Why are married men working so much? An aggregate analysis of intra-household bargaining and labour supply. Rev. Econ. Stud. 80 (3), 1055–1085. http://dx.doi.org/10.1093/restud/rds043. http://restud.oxfordjournals.org/content/80/3/1055.abstract.

Konrad, K.A., Lommerud, K.E., 1995. Family policy with non-cooperative families. Scand. J. Econ. 97 (4), 581–601.

Lagerlöf, N.P., 2003. Gender equality and long-run growth. J. Econ. Growth 8 (4), 403–426.

Lagerlöf, N.P., 2005. Sex, equality, and growth. Can. J. Econ. 38 (3), 807–831.

Lagerlöf, N.P., 2010. Pacifying monogamy. J. Econ. Growth 15 (3), 235–262.

Lee, M., 2015. Allocation of female talent and cross-country productivity differences. Unpublished Manuscript, University of Chicago.

Lee, T., Seshadri, A., 2015. On the intergenerational transmission of economic status. Unpublished Manuscript, University of Mannheim.

Liao, P.J., 2013. The one-child policy: a macroeconomic analysis. J. Dev. Econ. 101, 49–62.

Lise, J., Seitz, S., 2011. Consumption inequality and intra-household allocations. Rev. Econ. Stud. 78, 328–355.

Lise, J., Yamada, K., 2015. Household sharing and commitment: evidence from panel data on individual expenditures and time use. Unpublished Manuscript, University College London.

Lizzeri, A., Persico, N., 2004. Why did the elites extend the suffrage? Democracy and the scope of government, with an application to Britain's 'age of reform'. Q. J. Econ. 119 (2), 705–763.

Love, D.A., 2010. The effects of marital status and children on savings and portfolio choice. Rev. Finan. Stud. 23 (1), 385–432.

Lucas, R.E., 1988. On the mechanics of economic development. J. Monetary Econ. 22 (1), 3–42.

Lundberg, S., 1985. The added worker effect. J. Labor Econ. 3 (1), 11–37. http://www.jstor.org/stable/2535048. ISSN 0734-306X, 1537-5307.

Lundberg, S., Pollak, R.A., 1993. Separate spheres bargaining and the marriage market. J. Polit. Econ. 101 (6), 988–1010. http://www.jstor.org/stable/2138569. ISSN 0022-3808.

Lundberg, S., Pollak, R.A., 1994. Noncooperative bargaining models of marriage. Am. Econ. Rev. 84 (2), 132–137.http://www.jstor.org/stable/2138558ISSN 0895-3309.

Mammen, K., Paxson, C., 2000. Women's work and economic development. J. Econ. Perspect. 14 (4), 141–164.

Mankart, J., Oikonomou, R., 2015. Household search and the aggregate labor market. Unpublished Manuscript, UC Louvain.

Manser, M., Brown, M., 1980. Marriage and household decision-making: a bargaining analysis. Int. Econ. Rev. 21 (1), 31–44. http://www.jstor.org/stable/2526238. ISSN 0020-6598.

Manuelli, R., Seshadri, A., 2009. Explaining international fertility differences. Q. J. Econ. 124 (2), 771–807.

Mazzocco, M., 2007. Household intertemporal behaviour: A collective characterization and a test of commitment. Rev. Econ. Stud. 74 (3), 857–895. http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=25378400&site=ehost-live. ISSN 0034-6527.

Mazzocco, M., 2008. Individual rather than household euler equations: identification and estimation of individual preferences using household data. Unpublished Manuscript, UCLA.

Mazzocco, M., Ruiz, C., Yamaguchi, S., 2013. Labor supply, wealth dynamics, and marriage decisions. Unpublished Manuscript, UCLA.

McElroy, M.B., Horney, M.J., 1981. Nash-bargained household decisions: toward a generalization of the theory of demand. Int. Econ. Rev. 22 (2), 333–349. http://www.jstor.org/stable/2526280. ISSN 0020-6598.

Mennuni, A., 2015. Labour force composition and aggregate fluctuations. Unpublished Manuscript, University of Southampton.

Miles, D., 1999. Modelling the impact of demographic change upon the economy. Econ. J. 109 (452), 1–36. http://www.jstor.org/stable/2565892. ISSN 0013-0133, 1468-0297.

Miller, A.R., 2011. The effects of motherhood timing on career path. J. Popul. Econ. 24 (3), 1071–1100.

Ngai, L.R., Petrongolo, B., 2014. Gender gaps and the rise of the service economy. IZA Discussion Paper No. 8134.

Olivetti, C., 2014. The female labor force and long-run development: the American experience in comparative perspective. In: Boustan, L.P., Frydman, C., Margo, R.A. (Eds.), Human Capital in History: The American Record. University of Chicago Press, Chicago, IL, USA.

Olivetti, C., Petrongolo, B., 2016. The evolution of gender gaps in industrialized countries. Annu. Rev. Econ., 8, forthcoming.

Ortigueira, S., Siassi, N., 2013. How important is intra-household risk sharing for savings and labor supply? J. Monetary Econ. 60 (6), 650–666.

Oswald, A.J., Powdthavee, N., 2010. Daughters and left-wing voting. Rev. Econ. Stat. 92 (2), 213–227.

Pistaferri, L., 2003. Anticipated and unanticipated wage changes, wage risk, and intertemporal labor supply. J. Labor Econ. 21 (3), 729–754.

Rendall, M., 2010. Brain versus brawn: the realization of womens comparative advantage. Unpublished Manuscript, University of Zurich.

Rendall, M., 2015. Female market work, tax regimes, and the rise of the service sector. Unpublished Manuscript, University of Zurich.

Restuccia, D., Urrutia, C., 2004. Intergenerational persistence of earnings: the role of early and college education. Am. Econ. Rev. 94 (5), 1354–1378.

Ríos-Rull, J.V., 1993. Working in the market, working at home, and the acquisition of skills: a general-equilibrium approach. Am. Econ. Rev. 83 (4), 893–907.

Ríos-Rull, J.V., 1996. Life-cycle economies and aggregate fluctuations. Rev. Econ. Stud. 63 (3), 465–489. http://dx.doi.org/10.2307/2297891.http://restud.oxfordjournals.org/content/63/3/465.abstract.

Ríos-Rull, J.V., 2001. Population changes and capital accumulation: the aging of the baby boom. B.E. J. Macroecon. Adv. Macroecon. 1 (1), Article 7. http://dx.doi.org/10.2307/2297891. http://restud.oxfordjournals.org/content/63/3/465.abstract.

Robinson, J., 2012. Limited insurance within the household: evidence from a field experiment in Kenya. Am. Econ. J. Appl. Econ. 4 (4), 140–164.

Rosenzweig, M.R., Wolpin, K.I., 1980. Testing the quantity-quality fertility model: the use of twins as a natural experiment. Econometrica 48 (1), 227–240.

Ruggles, S., 1994. The transformation of American family structure. Am. Hist. Rev. 99 (1), 103–128.

Salcedo, A., Schoellmann, T., Tertilt, M., 2012. Families as roommates: changes in U.S. household size from 1850 to 2000. Quant. Econ. 3 (1), 133–175.

Schaner, S.G., 2015. Do opposites detract? intrahousehold preference heterogeneity and inefficient strategic savings. Am. Econ. J. Appl. Econ., 7 (2), 135–174.

Scholz, J.K., Seshadri, A., 2009. Children and household wealth. Unpublished Manuscript, University of Wisconsin.

Schoonbroodt, A., 2016. Parental child care during and outside of typical work hours. Rev. Econ. Househ. forthcoming.

Schoonbroodt, A., Tertilt, M., 2014. Property rights and efficiency in OLG models with endogenous fertility. Journal of Economic Theory 150, 551–582.

Shore, S.H., 2010. For better, for worse: intrahousehold risk-sharing over the business cycle. Rev. Econ. Stat. 92 (3), 536–548.

Shore, S.H., 2015. The co-movement of couples incomes. Rev. Econ. Household 13 (3), 569–588.

Soares, R., 2005. Mortality reductions, educational attainment, and fertility choice. Am. Econ. Rev. 95 (3), 580–601.

Song, Z., Storesletten, K., Wang, Y., Zilibotti, F., 2015. Sharing high growth across generations: pensions and demographic transition in China. Am. Econ. J. Macroecon. 7 (2), 1–39.

Tarozzi, A., Mahajan, A., 2007. Child nutrition in India in the nineties. Econ. Dev. Cult. Change 55 (3), 441–486.

Tertilt, M., 2005. Polygyny, fertility, and savings. J. Polit. Econ. 113 (6), 1341–1371.

Townsend, R., 2010. Financial structure and economic welfare: applied general equilibrium development economics. Annu. Rev. Econ. 2 (1), 507–546. http://dx.doi.org/10.1146/annurev.economics.102308.124427.

Udry, C., 1996. Gender, agricultural production, and the theory of the household. J. Polit. Econ. 104 (5), 1010–1046.

Voena, A., 2015. Yours, mine, and ours: do divorce laws affect the intertemporal behavior of married couples? Am. Econ. Rev. 105 (8), 2295–2332.

Vogl, T., 2016. Differential fertility, human capital, and development. Rev. Econ. Stud. 83 (1), 365–401.

Washington, E., 2008. Female socialization: how daughters affect their legislator fathers. Am. Econ. Rev. 98 (1), 311–332.

Wei, S.J., Zhang, X., 2011. The competitive saving motive: evidence from rising sex ratios and savings rates in China. J. Polit. Econ. 119 (3), 511–564.

Yum, M., 2015. Parental time investment and intergenerational mobility. Unpublished Manuscript, University of Mannheim.

CHAPTER 24

Environmental Macroeconomics

J. Hassler[*,†,‡], P. Krusell[*,†,‡,§], A.A. Smith, Jr.[§,¶]
[*] Institute for International Economic Studies (IIES), Stockholm University, Stockholm, Sweden
[†] University of Gothenburg, Gothenburg, Sweden
[‡] CEPR, London, United Kingdom
[§] NBER, Cambridge, MA, United States
[¶] Yale University, New Haven, CT, United States

Contents

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.04.007

Abstract

We discuss climate change and resource scarcity from the perspective of macroeconomic modeling and quantitative evaluation. Our focus is on climate change: we build a very simple “integrated assessment model,” ie, a model that integrates the global economy and the climate in a unified framework. Such a model has three key modules: the climate, the carbon cycle, and the economy. We provide a description of how to build tractable and yet realistic modules of the climate and the carbon cycle. The baseline economic model, then, is static but has a macroeconomic structure, ie, it has the standard features of modern macroeconomic analysis. Thus, it is quantitatively specified and can be calibrated to obtain an approximate social cost of carbon. The static model is then used to illustrate a number of points that have been made in the broad literature on climate change. Our chapter begins, however, with a short discussion of resource scarcity—also from the perspective of standard macroeconomic modeling—offering a dynamic framework of analysis and stating the key challenges. Our last section combines resource scarcity and the integrated assessment setup within a fully dynamic general equilibrium model with uncertainty. That model delivers positive and normative quantitative implications and can be viewed as a platform for macroeconomic analysis of climate change and sustainability issues more broadly.

Keywords

Climate system, Climate change, Carbon cycle, Damages, Growth, Discounting, Externality, Pigou tax

JEL Classification Code

H23, O4, Q01, Q3, Q4, Q54

1. INTRODUCTION

In this chapter we discuss climate change and resource scarcity from the perspective of macroeconomic modeling and quantitative evaluation. Our focus is to build toward an "integrated assessment model," (IAM) ie, a model that integrates the global economy and the climate in a unified framework. The chapter is not meant to be a survey of the rather broad field defined by interconnections between climate and economics. Rather, it has a sharp focus on the use of microeconomics-based macroeconomic models in this area, parameterized to match historical data and used for positive and normative work. Our understanding of the literature is that this approach, which is now standard macroeconomic in analyses (rather broadly defined), has not been dominant in the literature focused on developing IAMs, let alone anywhere else in the climate literature. We consider it a very promising approach also for climate-economy work, however, having contributed to it recently; in fact, the treatment we offer here is naturally built up around some of our own models and substantive contributions. Although there is a risk that this fact will be interpreted as undue marketing of our own work, it is rather that our climate-economy work from the very beginning made an effort precisely to formulate the IAM, and all the issues that can be discussed with an IAM, in terms of a standard macroeconomic settings and in such a way that calibration and model evaluation could be conducted with standard methods. Ex-post, then, one can say that our work grew out of an effort to write something akin to a climate-economy handbook for macroeconomists, even though the kind offer to write an actual such a chapter arrived much later. At this point, with this work, we are simply hopeful that macroeconomists with modern training will find our exposition useful as a quick introduction to a host of issues and perhaps also as inspiration for doing research on climate change and sustainability. We do find the area of great importance and, at the same time, rather undeveloped in many ways.

One exception to our claim that IAMs are not microeconomics-based macroeconomic models is Nordhaus's work, which started in the late 1970s and which led to the industry standards DICE and RIce: dynamic integrated models of climate and the economy, DICE depicting a one-region world and RICE a multiregion world. However, these models remain the nearest thing to the kind of setting we have in mind, and even the DICE and RICE models are closer to pure planning problems. That is, they do not fully specify market structures and, hence, do not allow a full analysis of typical policies such as a carbon tax or a quota system. Most of the models in the literature—to the extent they are fully specified models—are simply planning problems, so a question such as "What happens if we pursue a suboptimal policy?" cannot be addressed. This came as a surprise to us when we began to study the literature. Our subsequent research and the present chapter thus simply reflect this view: some more focus on the approach used in modern macroeconomics is a useful one.

So as a means of abstract introduction, consider a growth economy inhabited by a representative agent with utility function $\sum_{t=0}^{\infty}\beta^t u(C_t, S_t)$ with a resource constraint $C_t + K_{t+1} = (1-\delta)K_t + F(K_t,\ E_t,\ S_t)$ and with a law of motion $S_{t+1} = H(S_t,\ E_t)$. The new variables, relative to a standard macroeconomic setting, are S and E. S, a stock, represents something that is affects utility directly and/or affects production, whereas E, a flow, represents an activity that influences the stock. To a social planner, this would be nothing but an augmented growth model, with (interrelated) Euler equations both for K and S. In fact, standard models of human capital accumulation map into this setup, with H increasing in both arguments and F increasing in S but decreasing in E.[a] However, here we are interested in issues relating to environmental management—from a macroeconomic perspective—and then the same setup can be thought of, at least in abstract, with different labels: we could identify S with, say, clean air or biodiversity, and E with an activity that raises output but lowers the stock S. Our main interest will be in the connections between the economy and the climate. Then, S_t can be thought of as the climate at t, or a key variable that influences it, namely, the stock of carbon in the atmosphere; and E_t would be emissions of carbon dioxide caused by the use of fossil fuel in production. The carbon stock S then hurts both utility (perhaps because a warmer climate makes people suffer more in various ways) and output. Thus, $u_2 < 0$, $F_2 > 0$, $F_3 < 0$, $H_1 > 0$, and $H_2 > 0$. The setting still does not appear fully adequate for looking at the climate issue, because there ought to be another stock: that of the available amounts of fossil fuel (oil, coal, and natural gas), which are depletable resources in finite supply. Indeed, many of our settings below do include such stocks, but as we will argue even the setting without an additional stock is quite useful for analyzing the climate issue.

[a] See, eg, Lucas, 1988.

Furthermore, one would also think that technology, and technological change of different sorts, must play a role, and indeed we agree. Technology can enhance the production possibilities in a neutral manner but also amount to specific forms of innovation aimed at developing nonfossil energy sources or more generally saving on fossil-based energy. We will discuss these issues in the chapter too, including endogenous technology, but the exposition covers a lot of ground and therefore only devotes limited attention to technology endogeneity.

Now so far the abstract setting just described simply describes preferences and technology. So how would markets handle the evolution of the two stocks K and S? The key approach here is that it is reasonable to assume, in the climate case, that the evolution of S is simply a byproduct of economic activity: an externality. Thus, tracing out the difference between an optimal path for K and S and a laissez-faire market path becomes important, as does thinking about what policies could be used to move the market outcome toward the optimum as well as what intermediate cases would imply. Thus, the modern macroeconomist approach would be to (i) define a dynamic competitive equilibrium with policy (say, a unit tax on E), with firms, consumers, and markets clearly spelled out, then (ii) look for insights about optimal policy both qualitatively and quantitatively (based on, say, calibration), and perhaps (iii) characterize outcomes for the future for different (optimal and suboptimal) policy scenarios. This is the overall approach we will follow here.

We proceed in three steps. In the first step, contained in Section 2, we discuss a setting with resource scarcity alone—such as an economy with a limited amount of oil. How will markets then price the resource, and how will it be used up over time? Thus, in this section we touch on the broader area of "sustainability," whereby the question is how the economy manages a set of depletable resources. It appears to be a common view in the public debate that markets do not carry this task out properly, and our view is that it really is an open question whether they do or not; indeed, we find this issue intriguing in itself, quite aside from any interest in the specific area of climate change. The basic market mechanisms we go through involve the Hotelling rule for pricing and then, coupled with a representative agent with preferences defined over time as in our abstract setting above and a specific demand for the resource (say, from its use in production), a dynamic path for resource use. As a preliminary exploration into whether our market-based analysis works, one can compare the models implications for prices and quantities and we briefly do. As a rough summary, it is far from clear that Hotelling-based pricing can explain our past data for depletable resources (like fossil fuel or metals). Similarly, it is challenging to account for the historical patterns of resource use, though here the predictions of the theory are arguably less sharp. Taken together, this suggests that it is not obvious that at least our benchmark theories of markets match the data, so it seems fruitful to at least consider alternatives. In Section 2 we also look at the case of fossil fuel in more detail and, in this context, look at (endogenous) technical change: we look at how markets could potentially react to resource scarcity by saving on the scarce resource instead of saving on other inputs. Thus, we apply the notion of "directed technical change" in this context and propose it as an interesting avenue for conducting further macroeconomic research within the area of sustainability more broadly. Finally, Section 2 should be viewed as a delivering a building block for the IAMs to be discussed later in the chapter, in particular that in Section 5.

In Section 4, we take our second step and develop a very simple, static integrated assessment model of climate change and the global economy. Despite its being simple and stylized, this baseline model does have a macroeconomic structure, ie, it makes assumptions that are standard in modern macroeconomic analysis. Many of its key parameters are therefore straightforwardly calibrated to observables and thus, with the additional calibration necessary to introduce climate into the model, it can be used to obtain an approximate social cost of carbon. The static model is then used to illustrate a number of points that have been made in broad literature on climate change. None of these applications do full justice to the literature, of course, since our main purpose is to introduce the macroeconomic analyst to it. At the same time, we do offer a setting that is quantitatively oriented and one can imagine embedding each application in a fully dynamic and calibrated model; in fact, as far as we are aware, only a (minority) subset of these applications exist as full quantitative studies in the literature.

In our last section, Section 5, which is also the third and final step of the chapter, we describe a fully dynamic, stochastic IAM setting. With it, we show how to derive a robust formula for the (optimal) marginal cost of carbon and, hence, the appropriate Pigou tax. We show how to assign parameter values and compute the size of the optimal tax. The model can also be used as a complete setting for predicting the climate in the future—along with the paths for consumption, output, etc.—for different policy paths. We conclude that although the optimal-tax formula is quite robust, the positive side of the model involves rather strong sensitivity to some parameters, such as those involving different sources for energy generation and, of course, the total sizes of the stocks of fossil fuels.

Before transiting from discussing sustainability in Section 2 to climate modeling in Section 4, we offer a rather comprehensive introduction to the natural-science aspects of climate change. Section 3 is important for explaining what we perceive as the basic and (among expert natural scientists) broadly agreed upon mechanisms behind global warming: how the climate is influenced by the carbon concentration in the atmosphere (the climate model) and how the carbon concentration evolves over time as a function of the time path for emissions (the model of the carbon cycle). This presentation thus offers two "modules" that are crucial elements in IAMs. These modules are extremely simplified versions of what actual climate models and carbon-cycle models in use look like. However, they are, we argue, decent approximations of up-to-date models. The reason why simplifications are necessary is that our economic models have forward-looking agents and it is well known that such models are much more difficult to analyze, given any complexity in the laws of motions of stocks given flows: they involve finding dynamic fixed points, unlike any natural-science model where particles behave mechanically.[b]

Finally, although it should be clear already, let us reiterate that this chapter fails to discuss many environmental issues that are of general as well as macroeconomic interest. For example, the section on sustainability does not discuss, either empirically or theoretically, the possible existence of a "pollution Kuznets curve": the notion that over the course of economic development, pollution (of some or all forms) first increases and then decreases.[c] That section also does not offer any theoretical discussion of other common-pool problems than that associated with our climate (such as overfishing or pollution). The sections on IAMs, moreover, does not contain a listing/discussion of the different such models in the literature; such a treatment would require a full survey in itself.

2. LIMITED NATURAL RESOURCES AND SUSTAINABILITY CONCERNS

Climate change is a leading example within environmental economics where global macroeconomic analysis is called for. It involves a global externality that arises from the release of carbon dioxide into the atmosphere. This release is a byproduct of our economies' burning of fossil fuel, and it increases the carbon dioxide concentration worldwide and thus causes warming not just where the emission occurs. In two ways, climate change makes contact with the broader area of *sustainability*: it involves two stocks that are important for humans and that are affected by human activity. The first stock is the carbon concentration in the atmosphere. It exerts an influence on the global climate; to the extent warming causes damages on net, it is a stock whose size negatively impacts human welfare. The second stock is that of fossil fuels, ie, coal, oil, and natural gas. These stocks are not harmful per se but thus can be to the extent they are burnt.

More generally, sustainability concerns can be thought of in terms of the existence of stocks in finite supply with two properties: (i) their size is affected by economic activity and (ii) they influence human welfare.[d] Obvious stocks are natural resources in finite supply, and these are often traded in markets. Other stocks are "commons," such as air quality, the atmosphere, oceans, ecosystems, and biodiversity. Furthermore, recently, the term "planetary boundaries" has appeared (Rockström et al., 2009). These boundaries represent other limits that may be exceeded with sufficient economic growth (and therefore, according to the authors, growth should be limited). This specific *Nature* article lists

[b] The statement about the complexity of economic models does not rely on fully rational expectations, which we do assume here, but at least on some amount of forward-looking because any forward-looking will involve a dynamic fixed-point problem.

[c] See, eg, Grossman and Krueger, 1991 and Stokey, 1998.

[d] Relatedly, but less relevant from the perspective taken in this section, there is theoretical work on sustainability, defining, based on a utility-function representation, what the term means: roughly, an allocation is sustainable if the indirect utility function of generation t is not be below that of generation $t-k$.

nine boundaries, among them climate change; the remaining items are (i) stratospheric ozone depletion, (ii) loss of biosphere integrity (biodiversity loss and extinctions), (iii) chemical pollution and the release of novel entities, (iv) ocean acidification, (v) freshwater consumption and the global hydrological cycle, (vi) land system change, (vii) nitrogen and phosphorus flows to the biosphere and oceans, and (viii) atmospheric aerosol loading. Thus, these are other examples of commons.

Aside from in the work on climate change, the macroeconomic literature has had relatively little to say on the effects and management of global stocks. The Club of Rome (that started in the late 1960s) was concerned with population growth and a lack of food and energy. The oil crisis in the 1970s prompted a discussion about the finiteness of oil (see, eg, the 1974 *Review of Economic Studies* issue on this topic), but new discoveries and a rather large fall in the oil price in the 1980s appeared to have eliminated the concern about oil among macroeconomists. Similarly, technology advances in agriculture seemed to make limited food supply less of an issue. Nordhaus (1973, 1974) discussed a limited number of metals in finite supply, along with their prices, and concluded that the available stocks were so large at that point that there was no cause for alarm in the near to medium-run future. Thus, the concerns of these decades did not have a long-lasting impact on macroeconomics. Perhaps relatedly, so-called green accounting, where the idea is to measure the relevant stocks and count their increases or decreases as part of an extended notion of national economic product, was proposed but has been implemented and used in relatively few countries.[e] Limited resources and sustainability are typically not even mentioned in introductory or intermediate undergraduate textbooks in macroeconomics, let alone in PhD texts. In PhD texts specifically on growth, there is also very little: Aghion and Howitt's (2008) growth book has a very short, theoretical chapter on the subject, Jones (2001) has a chapter in his growth book which mentions some data; Acemoglu's (2009) growth book has nothing.[f]

The purpose here is not to review the literature but to point to this broad area as one of at least potential relevance and as one where we think that more macroeconomic research could be productive. To this end, we will discuss the basic theory and its confrontation with data. This discussion will lay bare some challenges and illustrate the need for more work.

We will focus on finite resources that are traded in markets and hence abstract from commons, mainly because these have not been subject to much economic macroeconomic analysis (with the exception of the atmosphere and climate change, which we will

[e] For example, in the United States, the BEA started such an endeavor in the 1990s but it was discontinued.

[f] The area of *ecological economics* is arguably further removed from standard economic analysis and certainly from macroeconomics. It is concerned precisely with limited resources but appears, at least in some of its versions, to have close connections Marx's labor theory of value, but with "labor" replaced by "limited resources" more broadly and, in specific cases, "energy" or "fossil fuel".

discuss in detail later). Thus, our discussion begins with price formation and quantity determination in markets for finite resources and then moves on to briefly discuss endogenous technological change in the form of resource saving.

2.1 Prices and Quantities in Markets for Finite Resources

To begin with, let us consider the simplest of all cases: a resource e in finite supply R that is costless to extract and that has economic value. Let us suppose the economic value is given by an inverse demand function $p_t = D(e_t)$, which we assume is time-invariant and negatively sloped. In a macroeconomic context we can derive such a function assuming, say, that e is an input into production. Abstracting from capital formation, suppose $y_t = F(n_t, e_t) = An_t^{1-\nu}e_t^{\nu}$, with inelastic labor supply $n_t = 1$, that $c_t = y_t$, and that utility is $\sum_{t=0}^{\infty}\beta^t \log c_t$.[g] Let time be $t = 0, \ldots, T$, with T possibly infinite. Here, the demand function would be derived from the firm's input decision: $p_t = \nu A e_t^{\nu-1}$.

2.1.1 The Hotelling Result: The Price Equation in a Baseline Case

The key notion now is that the resource can be saved. We assume initially that extraction/use of the resource is costless. The decision to save is therefore a dynamic one: should the resource be sold today or in the future? For a comparison, an interest rate is needed, so let r_t denote the interest rate between $t-1$ and t. If the resource is sold in two consecutive periods, it would then have to be that on the margin, the owner of the resource is indifferent between selling at t and at $t+1$:

$$p_t = \frac{1}{1+r_{t+1}} p_{t+1}.$$

This is the *Hotelling equation*, presented in Hotelling (1931). The price of the finite resource, thus, grows at the real rate of interest. The equation can also be turned around, using the inverse demand function, to deliver predictions for how the quantity sold will develop; for now, however, let us focus on the price. Thus, we notice that an arbitrage condition delivers a sharp prediction for the dynamics of the price that is independent of the demand. For the *price dynamics*, the demand is only relevant to the extent it may be such that the resource is not demanded at all at some point in time. For the *price level(s)*, however, demand is of course key: one needs to solve the difference equation along with the inverse demand function and the constraint on the resource to arrive at a value for p_0 (and, consequently, all its subsequent values). Here, p_t would be denoted the *Hotelling rent* accruing to the owner: as it is costless to extract it, the price is a pure rent. Thus, to the extent the demand is higher, the price/rent path will be at a higher level. Similarly, if there is more of the resource, the price/rent path will be lower, since more will be used at each point in time.

[g] In all of this section, we use logarithmic utility. More general CRRA preferences would only slightly change the analysis and all the key insights remain the same in this more general case.

2.1.2 Prices and Quantities in Equilibrium: Using a Planning Problem

Let us consider the planning problem implicit in the earlier discussion and let us for simplicity assume that $T=\infty$. Thus the planner would maximize $\nu\sum_{t=0}^{T}\beta^t \log c_t$ subject to $c_t = Ae_t^{\nu}$ for all t and $\sum_{t=0}^{T} e_t = R$.[h] This delivers the condition $\nu\beta^t/e_t = \mu$, where μ is the multiplier on the resource constraint, and hence $e_{t+1} = \beta e_t$. Inserting this into the resource constraint, one obtains $e_0(1+\beta+\ldots) = e_0/(1-\beta) = R$. Hence, $e_0 = (1-\beta)R$ and the initial price of the resource in terms of consumption (which can be derived from the decentralization) will be $p_0 = A\nu((1-\beta)R)^{\nu-1}$. Furthermore, $p_t = A\nu((1-\beta)R)^{\nu-1}\beta^{(\nu-1)t}$; notice that the gross interest rate here is constant over time and equal to $\beta^{\nu-1}$.[i] We see that a more abundant resource translates into a lower price/rent. In particular, as R goes to infinity, the price approaches 0: marginal cost. Similarly, higher demand (eg, through a higher A or higher weight on future consumption, β, so that the resource is demanded in more periods and will thus not experience as much diminishing returns per period), delivers a higher price/rent. Consider also the extension where the demand parameter A is time varying. Then the extraction path is not affected at all, due to income and substitution effects canceling. The consumption interest rates will change, since the relative price between consumption and the resource must change. The equation for price dynamics applies just as before, however, so price growth is affected only to the extent the interest rate changes. The price level, of course, is also affected by overall demand shifts.

2.1.3 Extraction Costs

More generally, suppose that the marginal cost of extraction of the resource is c_t in period t, and let us for simplicity assume that these marginal costs are exogenous (more generally it would depend on the amount extracted and the total remaining amount of the resource). The Hotelling formula for price dynamics becomes

$$p_t - c_t = \frac{1}{1+r_{t+1}}(p_{t+1} - c_{t+1}).$$

Put differently, the Hotelling rent, which is now the marginal profit per unit, $p-c$, grows at the real rate of interest. This is thus the more general formula that applies. It is robust in a number of ways; eg, allowing endogenous extraction costs delivers the same formula and the consideration of uncertainty reproduces the formula in expectation).[j] The discussion of determinants of prices and quantities above thus still applies, though the

[h] For $\nu=1$ this is a standard cake-eating problem.

[i] The Euler equation of the consumer delivers $1+r_{t+1} = c_{t+1}/(c_t\beta) = e_{t+1}^{\nu}/(e_t^{\nu}\beta) = \beta^{\nu}/\beta = \beta^{\nu-1}$.

[j] The case where the natural resource is owned by a monopolist produces a more complicated formula, as one has to consider marginal revenue instead of price and as the interest rate possibly becomes endogenous. However, the case of monopoly does not appear so relevant, at least not today. In the case of oil, Saudi oil production is currently only about 10% of world production.

key object now becomes the marginal profit per unit. First, the general idea that more of the resource (higher R) lowers the price survives: more of the resource moves the price toward marginal cost, thus gradually eliminating the rent. Second, regarding the effects of costs, let us consider three key cases: one where marginal costs are constant (and positive), one where they are declining, and one where they are increasing. We assume, for simplicity, that there is a constant interest rate. A constant positive marginal cost thus implies that the price is rising at a somewhat lower rate initially than when extraction is costless, since early on the price is a smaller fraction of the rent (early on, there is more left of the resource). If the marginal cost of extraction rises over time—a case that would apply in the absence of technological change if the easy-to-extract sources are exploited first—the price will rise at a higher rate; and under the assumption of a falling marginal extraction cost, typically reflecting productivity improvements in extraction, prices rise more slowly. Quantity paths change accordingly, when we use an invariant demand function. With a faster price rise, quantities fall faster, and vice versa. In particular, when the future promises lower (higher) extraction costs, extraction is postponed (slowed down) and so falls less (more) rapidly.

2.2 Confronting Theory with Data

The Hotelling predictions are, in principle, straightforwardly compared with data. The ambition here is not to review all the empirical work evaluating the Hotelling equation for finite resources but merely to mention some stylized facts and make some general points.[k] As for prices, it is well known that (real) prices of metals fall at a modest rate over the "long run," measured as one hundred years or more; see, eg, Harvey et al. (2010). The prices of fossil fuels (oil, coal, and natural gas) have been stable, with a slight net increase over the last 40 or so years. The volatilities of all these time series are high, on the order of magnitude of those for typical stock-market indices.[l] When it comes to quantities, these time series have been increasing steadily, and with lower fluctuations than displayed by the corresponding prices. Are these observations broadly consistent with Hotelling's theory?

To answer this question, note that Hotelling's theory is mainly an arbitrage-based theory of prices and that quantity predictions involve more assumptions on supply and demand, such as those invoked in our planning problem above. To evaluate Hotelling's rule, we first need to have an idea of the path for extraction costs, as they figure prominently in the more general version of the theory. The situation is somewhat complicated by the fact that extraction occurs on multiple sites. For oil at least, it is also clear that the marginal costs differ greatly between active oil wells, for example with much lower costs in Saudi Arabia than in the North Sea. This in itself appears inefficient, as the less

[k] For excellent discussions, see, eg, Krautkraemer, 1998 and Cuddington and Nülle, 2014.

[l] There are also attempts to identify long-run cycles; see, eg, Erten and Ocampo, 2012.

expensive oil ought to be extracted first in order to minimize overall present-value costs. We know of no study that has good measurements of marginal extraction costs going far back in time. Suppose, however, that productivity growth in the mining/extraction sector was commensurate with that in the rest of our economies. Then it would be reasonable to assume that the *relative* cost of extracting natural resources—and that is the relevant price given that we are referring to evidence on real prices—does not have any sharp movements upward or downward. Hence, the Hotelling formula, given a known total depletable stock of the resource, would imply an increasing price series, at a rate of a few percent per year, with a slightly lower growth rate early on, as explained earlier. This is clearly not what we see. It is, alternatively, possible that extraction costs have developed unevenly. Pindyck (1978) argues, for the case of oil, that lower and lower extraction costs explained a stable price path initially but that later extraction costs stabilized (or even increased), hence pushing prices up. In retrospect, however, although prices rose again in 1979 they did not continue increasing after that and rather fell overall; today, the oil price is back at a real price that is not terribly far from the pre-1973 level.

An proposed explanation for the lack of price growth in the data is a gradual finding of new deposits (of oil, metals, and so on). As explained earlier, the theory does predict lower prices for higher total deposits of the resource. However, it would then have to be that markets systematically underpredicted the successes of new explorations, and over very long periods of time.

Relatedly, it is possible that markets expect technological change in the form of the appearance of close substitutes to the resource in question. Consider a very simple case with a costless-to-extract raw material as in the baseline Hotelling model but where next period a perfect substitute, in infinite supply and with a constant marginal cost $\bar{p}$, appears with some probability. Then the arbitrage equation reads $p_t = \frac{1}{1+r_{t+1}}(\pi_{t+1}p_{t+1} + (1-\pi_{t+1})\bar{p})$, where π_{t+1} is the probability of the perfect substitute appearing. Clearly, such uncertainty and potential price competition will influence price dynamics and will lead to richer predictions. However, we know of no systematic study evaluating a quantitative version of this kind of hypothesis and comparing it to data.

A different view of the prices of natural resources (and commodities more generally) is the Prebisch (1962) and Singer (1950) hypothesis: that commodities have lower demand elasticities, so that when income rises, prices fall. Their hypothesis, thus, is in contrast with Hotelling's rule, since scarcity is abstracted from. Clearly, if one formulated a model with the Prebisch–Singer assumption and scarcity, as discussed earlier, the Hotelling formula would survive, and any demand effects would merely affect the level of the price path and not its dynamics.

In sum, although many authors claim that richer versions of the Hotelling model take its predictions closer to data, it seems safe to say that there is no full resolution of the contrast between the model's prediction of rising prices/profits per unit (at the rate of interest) and the data showing a stable or declining real price of the typical resource.

Some would argue that markets are not fully rational, or not forward-looking enough: the power of the scarcity argument in Hotelling's seminal work is very powerful but relies crucially on forward-looking with a long horizon, to the extent there is a relatively large amount of the resource left in ground. It seems to us that this hypothesis deserves some attention, though it is a challenge even to formulate it.[m]

To evaluate quantities, as underlined earlier, a fuller theory needs to be specified. This leads to challenges as well, as we shall see. Here, we will simply look at an application, albeit a well-known one and one that is relevant to the climate context. In the context of this application, we will also discuss technological change as a means toward saving on a scarce resource.

2.3 An Application: Fossil Energy

On a broad level, when a resource is in scarce supply, a key question is its substitutability with alternative resources. In this section, we look at fossil energy and provide an outline of how one could go about looking at one aspect of scarcity in this market: the response of energy saving, ie, one of the ways in which markets can respond to a shortage. This analysis, like the rest of this chapter (that addresses climate change), is built on a quantitatively oriented macroeconomic model. It can also be regarded as one of the building blocks in the climate-economy model; indeed, the exhaustible-resource formulation in Section 5 coincides with the core formulation entertained here.

The starting point is the extension of basic growth theory to include energy; the standard reference is Dasgupta and Heal (1974), but noteworthy other contributions include those by Solow (1974) and Stiglitz (1974). One of the main concerns here was precisely sustainability, ie, whether production functions (or various sorts) would allow future generations to be as well off as current generations. The Cobb–Douglas function was found to be an in-between case here; with more substitutability between energy and the other inputs, sustainability was possible. This line of work did not much address technical change, neither quantitatively nor theoretically. Clearly, much of the literature on scarce resources was written shortly after the oil-price hikes in the 1970s and it was not until the late 1980s that the theoretical developments allowed technological change to be endogenized in market-based environments.

We build a similar framework to that in Dasgupta and Heal's work and formulate an aggregate production function with three inputs—capital, labor, and fossil energy—and we use it to account for postwar US data. This analysis follows Hassler et al. (2015) closely. We allow technical change in this production function in the form of capital/labor saving and energy saving and we consider three broad issues: (i) what substitution elasticity (between a capital-labor composite, on the one hand, and energy, on the other) fits the data best; (ii) measurement of the series for input saving and to what extent they appear to respond to price movements (ie, does energy-saving appear to respond to the price of fossil fuel?); and (iii) the model's predictions for future input saving and fossil-fuel

[m] See, eg, Spiro (2014).

dependence. The model focuses on energy demand, as derived from an aggregate production function, and all of the discussion can be carried out without modeling supply.

So consider an aggregate production function of the nested CES form

$$y = \left[(1-\nu)\left[Ak^{\alpha}l^{1-\alpha}\right]^{\frac{\varepsilon-1}{\varepsilon}} + \left[A^{e}e\right]^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\varepsilon}{\varepsilon-1}},$$

with the obvious notation.[n] Here, we see that $\varepsilon \in [0, \infty]$ expresses the substitutability between capital/labor and energy. A is the technology parameter describing capital/labor saving and A^{e} correspondingly describes energy saving. If there is perfect competition for inputs, firms set the marginal product of each input equal to its price, delivering—expressed in terms of shares—the equations

$$\frac{wl}{y} = (1-\alpha)(1-\gamma)\left[\frac{Ak^{\alpha}l^{1-\alpha}}{y}\right]^{\frac{\varepsilon-1}{\varepsilon}} \tag{1}$$

and

$$\frac{pe}{y} = \gamma\left[\frac{A^{e}e}{y}\right]^{\frac{\varepsilon-1}{\varepsilon}}. \tag{2}$$

2.3.1 Accounting for Input Saving Using US Data

Eqs. (1) and (2) can be rearranged and solved directly for the two technology trends A and A^{E}. This means that it is possible, as do Hassler et al., to use data on output and inputs and their prices to generate time paths for the input-saving technology series. This is parallel with Solow's growth-accounting exercise, only using a specific functional form. In particular, A^{e} can be examined over the postwar period, when the price of fossil fuel—oil in particular—has moved around significantly, as shown in Fig. 1.

The authors use this setting and these data to back out series for A^{e} and A, conditional on a value for ε. With the view that the A and A^{e} series are technology series mainly, one can then examine the extent to which the backed-out series for different ε look like technology series: are fairly smooth and mostly nondecreasing. It turns out that ε has to be close to zero for the A^{e} series to look like a technology series at all; if ε is higher than 0.2 or so, the implied up-and-down swings in A^{e} are too high to be plausible. On the other hand, for a range of ε values between 0 (implying that production is Leontief) and 0.1, the series is rather smooth and looks like it could be a technology series. Fig. 2 plots both the A and A^{e} series. We see that A^{e} grows very slowly until prices rise; then it starts

[n] This production function introduces a key elasticity, along with input-specific technology levels, in the most tightly parameterized way. Extensions beyond this functional class, eg, to the translog case, would be interesting not only for further generality but because it would introduce a number of additional technology shifters; see, eg, Berndt and Christensen, 1973.

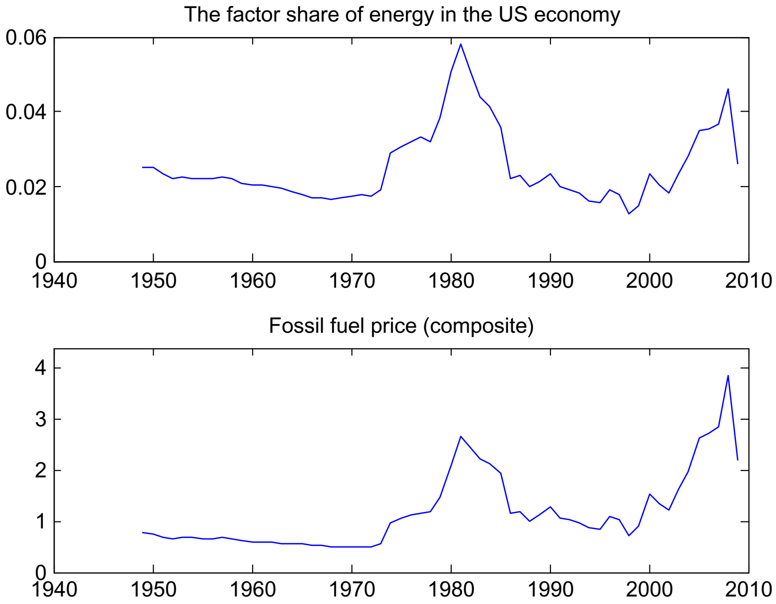


Fig. 1 Fossil energy share and its price.

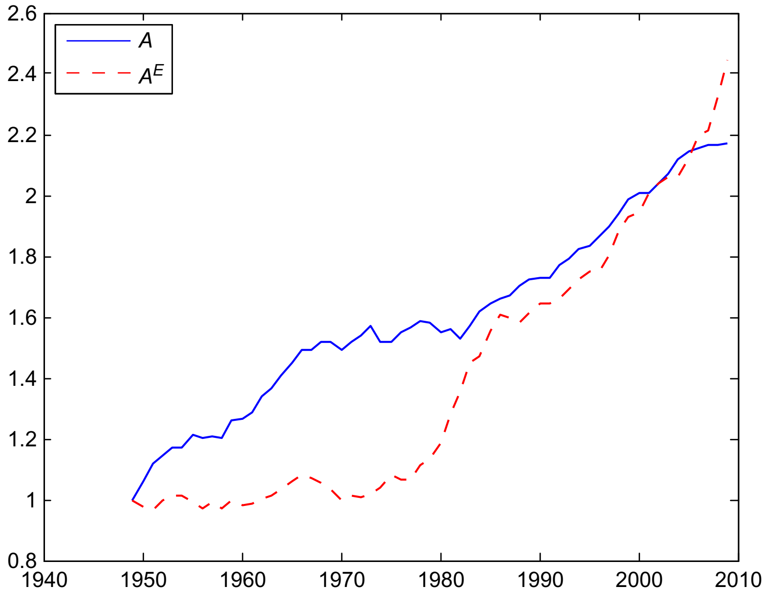


Fig. 2 Energy- and capital/labor-saving technologies compared.

growing significantly. Hence, the figure does suggest that the scarcity mechanism is operative in a quantitatively important way. It is also informative to look at how the two series compare. A it looks like TFP overall, but more importantly it does seem to covary negatively in the medium run with A^e, thus suggesting that the concept of *directed technical change* may be at play. In other words, when the oil price rose, the incentives to save on oil and improve oil efficiency went up, and to the extent these efforts compete for a scarce resource that could alternatively be used for saving on/improving the efficiency of capital and labor, as a result the latter efforts would have fallen.

Hassler et al. (2015) go on to suggest a formal model for this phenomenon and use it, with a calibration of the technology parameters in R&D based on the negative historical association between A and A^E, to also predict the future paths of technology and of energy dependence. We will briefly summarize this research later, but first it is necessary to formulate a quantitatively oriented dynamic macroeconomic model with energy demand and supply included explicitly.

2.3.2 A Positive Model of Energy Supply and Demand with a Finite Resource

Using the simple production function above and logarithmic preferences, it is straightforward to formulate a planner's problem, assuming that energy comes from a finite stock. We will first illustrate with a production function that is in the specified class and that is often used but that does not (as argued earlier) fit the macroeconomic data: the Cobb–Douglas case, where $F(Ak^\alpha, A^e e) = k^\alpha e^\nu$, where a constant labor supply (with a share $1-\alpha-\nu$) is implicit and we have normalized overall TFP including labor to 1. We also assume, to simplify matters, that (i) there is 100% depreciation of capital between periods (which fits a period of, say, 20 years or more) and that (ii) the extraction of energy is costless (which fits oil rather well, as its marginal cost is much lower than its price, at least for much of the available oil). For now, we abstract from technological change; we will revisit it later. Thus, the planner would maximize

$$\sum_{t=0}^{\infty} \log c_t$$

subject to

$$c_t + k_{t+1} = k_t^\alpha e_t^\nu$$

and $\sum_{t=0}^{\infty} e_t = R$, with R being the total available stock. It is straightforward to verify that we obtain a closed-form solution here: consumption is a constant fraction $1-\alpha\beta$ of output and $e_t = (1-\beta)\beta^t R$, ie, energy use falls at the rate of discount. As energy falls, so does capital, consumption, and output. In fact, this model asymptotically delivers balanced (negative) growth at a gross rate g satisfying (from the resource constraint) $g = g^\alpha \beta^\nu = \beta^{\overline{1-\alpha}}$ with ν over the bar: $g = g^\alpha\beta^\nu = \beta^{\frac{\nu}{1-\alpha}}$. Capital is not on the balanced path at all times, unless its initial value is in the proper relation to initial energy use.[o] This model of course also generates the Hotelling result: p_{t+1} must equal $p_t(1+r_{t+1})$, where $1+r$ is the marginal product of capital and $1+r$ hence the gross real interest rate. Notice, thus, that the interest rate will be constant on the balanced growth path but that it obeys transition dynamics. Hence, even though energy use falls at a constant rate at all times, the energy price will not grow at a constant rate at all times (unless the initial capital stock is at its balanced-growth level): it will grow either faster or slower. Consumption, along with output and capital, goes to zero here along a balanced growth path, but when there is sufficient growth in technology (which is easily added in the model), there will be positive balanced growth. The striking fall in energy use over time would of course be mitigated by an assumption that marginal extraction costs are positive and decreasing over time, as discussed earlier, but it is not obvious that such an assumption is warranted.

Fig. 3, which is borrowed from Hassler et al. (2015), shows that, in the data, energy (defined as a fossil composite) rises significantly over time. In contrast, as we have just shown, the simple Cobb–Douglas model predicts falling energy use, at a rate equaling the discount rate. Suppose instead one adopts the model Hassler et al. (2015) argue fit the data better, ie, a function that is near Leontief in k^α and e. Let us first assume that the technology coefficients A and A^e are constant over time. Then, there will be transition

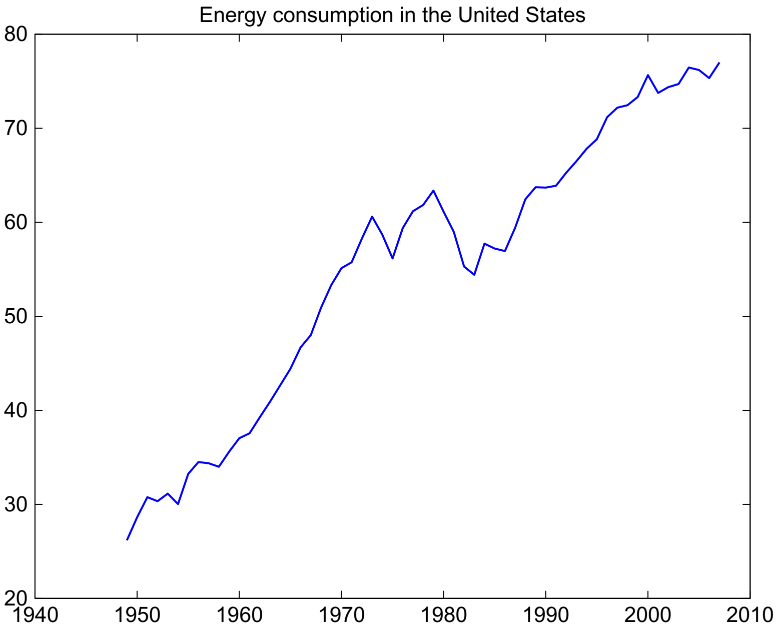


Fig. 3 US energy use.

[o] Initial capital then has to equal $(\alpha(R(1-\beta))^\nu)^{\frac{1}{1-\alpha}}\beta^{\frac{1-\alpha-\nu}{1-\alpha}}$.

dynamics in energy use, for Ak^α has to equal $A^e e$ at all points in time. Thus, the initial value of capital and R may not admit balanced growth in e at all times, given A and A^e. Intuitively, if Ak_0^α is too low, e will be held back initially and grow over time as capital catches up to its balanced path. Thus, it is possible to obtain an increasing path for energy use over a period of time. Eventually, of course, energy use has to fall. There is no exact balanced growth path in this case. Instead, the saving rate has to go to zero since any positive long-run saving rate would imply a positive capital stock.[p] Hence, the asymptotic economy will be like one without capital and in this sense behave like in a cake-eating problem: consumption and energy will fall at rate β. In sum, this model can deliver *peak oil*, ie, a path for oil use with a maximum later than at time 0. As already pointed out, increasing oil use can also be produced from other assumptions, such as a decreasing sequence of marginal extraction costs for oil; these explanations are complementary.

With exogenous technology growth in A and A^e it is possible that very different long-run extraction behavior results.[q] In particular, it appears that a balanced growth path with the property that $g_A g^\alpha = g_{A^e} g_e = g$ is at least feasible. Here, the first equality follows from the two arguments of the production function growing at the same rate—given that the production function F is homogeneous of degree one in the two arguments Ak^α and $A^e e$—and the second equality says that output and capital have to grow at that same rate. Clearly, if the planner chooses such asymptotic behavior, g_e can be solved for from the two equations to equal $g_A^{\frac{1}{1-\alpha}}/g_{A^e}$, a number that of course needs to be less than 1. Thus, in such a case, g_e will not generally equal β. A more general study of these cases is beyond the scope of the present chapter.

2.3.3 Endogenous Energy-Saving Technical Change

Given the backed-out series for A and A^e, which showed negative covariation in the medium run, let us consider the model of technology choice Hassler et al. (2015) propose. In it, there is an explicit tradeoff between raising A and raising A^e. Such a tradeoff arguably offers one of the economy's key behavioral responses to scarcity. That is, growth in A^e can be thought of as energy-saving technological change. In line with the authors' treatment, we consider a setup with directed technological change in the form of a planning problem, thus interpreting the outcome as one where the government has used policy optimally to internalize any spillovers in the research sector. It would be straightforward, along the lines of the endogenous-growth literature following Romer (1990), to consider market mechanisms based on variety expansion or quality improvements,

[p] If the saving rate asymptotically stayed above $_s > 0$, then $k_{t+1} \geq_s Ak_t^\alpha$. This would imply that capital would remain uniformly bounded below from zero. However, here, it does have to go to zero as its complement energy has to go to zero.

[q] An exception is the Cobb–Douglas case for which it is easy to show that the result above generalizes: e falls at rate β.

monopoly power, possibly with Schumpeterian elements, and an explicit market sector for R&D. Such an analysis would be interesting and would allow interesting policy questions to be analyzed. For example, is the market mechanism not allowing enough technical change in response to scarcity, and does the answer depend on whether there are also other market failures such as a climate externality? We leave these interesting questions for future research and merely focus here on efficient outcomes. The key mechanism we build in rests on the following simple structure: we introduce one resource, a measure one of "researchers." Researchers can direct their efforts to the advancement of A and A^e. We look at a very simple formulation:

$$A_{t+1} = A_t f(n_t) \quad \text{and} \quad A^e_{t+1} = A^e_t f_e(1 - n_t),$$

where $n_t \in [0,1]$ summarizes the R&D choice at time t and where f and f_e are both strictly increasing and strictly concave; these functions thus jointly demarcate the frontier for technologies at $t+1$ given their positions at t. Hence, at a point in time t, A_t and A^e_t are fixed. In the case of a Leontief technology, there would be absolutely no substitutability at all between capital and energy ex-post, ie, at time t when A_t and A^e_t have been chosen, but there is substitutability ex-ante, by varying n_s for $s < t$. With a less extreme production function there would be substitutability ex-post too but less so than ex-ante.[r] Relatedly, whereas the share of income in this economy that accrues to each of the inputs is endogenous and, typically, varies with the state of the economy, on a balanced growth path the share settles down. As we shall see, in fact, the share is determined in a relatively simple manner.

The analysis proceeds by adding these two equations to the above planning problem. Taking first-order conditions and focusing on a balanced-growth outcome, this model rather surprisingly delivers the result that the extraction rate must be equal to β, regardless of the values of all the other primitives.[s] This means, in turn, that two equations jointly determining the long-run growth rates of A and A_e can be derived. One captures the technology tradeoff and follows directly from the equations above stating that these growth rates, respectively, are $g_A = f(n)$ and $g_{A^e} = f^e(1-n)$. The other equation comes from the balanced-growth condition that $A_t k^\alpha_t = A^e_t e_t$, given that F is homogeneous of degree one; from this equality the growth rates of A and A^e are positively related. In fact, given that e_t falls at rate β, we obtain n from $g_A^{\frac{1}{1-\alpha}} = g_{A^e}\beta$.

[r] The Cobb–Douglas case is easy to analyze. It leads to an interior choice for n that is constant over time, regardless of initial conditions and hence looks like the case above where the two technology factors are exogenous.

[s] The proof is straightforward; for details, see Hassler et al. (2015). It is thus the endogeneity of the technology levels in the CES formulation that makes energy fall at rate β; when they grow exogenously, we saw that energy does not have to go to zero at rate β.

One can also show, quite surprisingly as well, that the long run share of energy s_e in output is determined by $(1-s_e)/s_e = -\partial \log g_A / \partial \log g_{A^e}$.[t] In steady state, this expression is a function of n only, and as we saw above it is determined straightforwardly knowing β, α, f, and f^e. How, then, can these primitives be calibrated? One way to proceed is to look at historical data to obtain information about the tradeoff relation between g_A and g_{A^e}. If this relation is approximately log-linear (ie, the net rates are related linearly), the observed slope is all that is needed, since it then gives $\partial \log g_A / \partial \log g_{A^e}$ directly. The postwar behaviors of A and A^e reported above imply a slope of -0.235 and hence a predicted long-run value of s_e of around 0.19, which is significantly above its current value, which is well below 0.1.

2.3.4 Takeaway from the Fossil-Energy Application

The fossil-energy application shows that standard macroeconomic modeling with the inclusion of an exhaustible resource can be used to derive predictions for the time paths for quantities and compare them to data. Moreover, the same kind of framework augmented with endogenous directed technical change can be used to look at optimal/market responses to scarcity. It even appears possible to use historical data reflecting past technological tradeoffs in input saving to make predictions for the future. The presentation here has been very stylized and many important real-world features have largely been abstracted from, such as the nature of extraction technologies over time and space. The focus has also been restricted to the long-run behaviors of the prices and quantities of the resources in limited supply, but there are other striking facts as well, such as the high volatilities in most of these markets. Natural resources in limited supply can become increasingly limiting for economic activity in the future and more macroeconomic research may need to be directed to these issues. Hopefully the analysis herein can give some insights into fruitful avenues for such research.

3. CLIMATE CHANGE: THE NATURAL-SCIENCE BACKGROUND

An economic model of climate change needs to describe three phenomena and their dynamic interactions. These are (i) economic activity; (ii) carbon circulation; and (iii) the climate. From a conceptual as well as a modeling point of view it is convenient to view the three phenomena as distinct sub-subsystems. We begin with a very brief description of the three subsystems and then focus this section on the two latter.

The economy consists of individuals that act as consumers, producers and perhaps as politicians. Their actions are drivers of the economy. In particular, the actions are determinants of emissions and other factors behind climate change. The actions are also

[t] The authors show that this result follows rather generally in the model: utility is allowed to be any power function and production any function with constant returns to scale.

responses to current and expected changes in the climate by adaptation. Specifically, when fossil fuel is burned, carbon dioxide (CO_2) is released and spreads very quickly in the atmosphere. The atmosphere is part of the carbon circulation subsystem where carbon is transported between different reservoirs; the atmosphere is thus one such reservoir. The biosphere (plants, and to a much smaller extent, animals including humans) and the soil are other reservoirs. The oceans constitute the largest carbon reservoir.

The climate is a system that determines the distribution of weather events over time and space and is, in particular, affected by the carbon dioxide concentration in the atmosphere. Due to its molecular structure, carbon dioxide more easily lets through short-wave radiation, like sun-light, than long-wave, infrared radiation. Relative to the energy outflow from earth, the inflow consists of more short-wave radiation. Therefore, an increase in the atmospheric CO_2 concentration affects the difference between energy inflow and outflow. This is the *greenhouse effect*.

It is straightforward to see that we need at minimum the three subsystems to construct a climate-economy model. The economy is needed to model emissions and economic effects of climate change. The carbon circulation model is needed to specify how emissions over time translate into a path of CO_2 concentration. Finally, the climate model is needed to specify the link between the atmospheric CO_2 concentration and the climate.

3.1 The Climate

3.1.1 The Energy Budget

We will now present the simplest possible climate model. As described earlier, the purpose of the climate model is to determine how the (path of) CO_2 concentration determines the (path of the) climate. A minimal description of the climate is the global mean atmospheric temperature near the surface. Thus, at minimum we need a relationship between the path of the CO_2 concentration and the global mean temperature. We start the discussion by describing the *energy budget* concept.

Suppose that the earth is in a radiative steady state where the incoming flow of short-wave radiation from the sun light is equal to the outgoing flow of largely infrared radiation.[u] The energy budget of the earth is then balanced, implying that the earth's heat content and the global mean temperature is constant.[v] Now consider a perturbation of this equilibrium that makes the net inflow positive by an amount F. Such an increase could be caused by an increase in the incoming flow and/or a reduction in the outgoing flow. Regardless of how this is achieved, the earth's energy budget is now in surplus

[u] We neglect the additional outflow due to the nuclear process in the interior of the earth, which is in the order of one to ten thousands in relative terms when compared to the incoming flux from the sun; see the Kam et al. (2011).

[v] We disregard the obvious fact that energy flows vary with latitude and over the year producing differences in temperatures over space and time. Since the outflow of energy is a nonlinear (convex) function of the temperature, the distribution of temperature affects the average outflow.

causing an accumulation of heat in the earth and thus a higher temperature. The speed at which the temperature increases is higher the larger is the difference between the inflow and outflow of energy, ie, the larger the surplus in the energy budget.

As the temperature rises, the outgoing energy flow increases since all else equal, a hotter object radiates more energy. Sometimes this simple mechanism is referred to as the 'Planck feedback'. As an approximation, let this increase be proportional to the increase in temperature over its initial value. Denoting the temperature perturbation relative to the initial steady state at time t by T_t and the proportionality factor between energy flows and temperature by κ, we can summarize these relations in the following equation:

$$\frac{dT_t}{dt} = \sigma(F - \kappa T_t). \tag{3}$$

The left-hand side of the equation is the speed of change of the temperature at time t. The term in parenthesis on the right-hand side is the net energy flow, ie, the difference in incoming and outgoing flows. The equation is labeled the energy budget and we note that it should be thought of as a flow budget with an analogy to how the difference between income and spending determines the speed of change of assets.

When the right-hand side of (3) is positive, the energy budget is in surplus, heat is accumulated, and the temperature increases. Vice versa, if the energy budget has a deficit, heat is lost, and the temperature falls. When discussing climate change, the variable F is typically called *forcing* and it is then defined as the change in the energy budget caused by human activities. The parameter σ is (inversely) related to the heat capacity of the system for which the energy budget is defined and determines how fast the temperature changes for a given imbalance of the energy budget.[w]

We can use Eq. (3) to find how much the temperature needs to rise before the system reaches a new steady state, ie, when the temperature has settled down to a constant. Such an equilibrium requires that the energy budget has become balanced, so that the term in parenthesis in (3) again has become zero. Let the steady-state temperature associated with a forcing F be denoted $T(F)$. At $T(F)$, the temperature is constant, which requires that the energy budget is balanced, ie, that $F - \kappa T(F) = 0$. Thus,

$$T(F) = \frac{F}{\kappa}. \tag{4}$$

Furthermore, the path of the temperature is given by

$$T_t = e^{-\sigma\kappa t}\left(T_0 - \frac{F}{\kappa}\right) + \frac{F}{\kappa}.$$

[w] The heat capacity of the atmosphere is much lower than that of the oceans, an issue we will return to below.

Measuring temperature in Kelvin (K), and F in Watt per square meter, the unit of κ is $\frac{W/m^2}{K}$.[x] If the earth were a blackbody without an atmosphere, we could calculate the exact value of κ from laws of physics. In fact, at the earth's current mean temperature $\frac{1}{\kappa}$ would be approximately 0.3, ie, an increase in forcing by 1 W/m^2 would lead to an increase in the global temperature of 0.3 K (an equal amount in degrees Celsius).[y] In reality, various feedback mechanisms make it difficult to assess the true value of κ. One of the important feedbacks is that a higher temperature increases the concentration of water vapor, which is also is a greenhouse gas; another is that the polar ice sheets melt, which decreases direct reflection of sun light and changes the cloud formation. We will return to this issue below but note that the value of κ is likely to be substantially smaller than the blackbody value of 0.3^{-1}, leading to a higher steady-state temperature for a given forcing.

Now consider how a given concentration of CO_2 determines F. This relationship can be well approximated by a logarithmic function. Thus, F, the change in the energy budget relative to preindustrial times, can be written as a logarithmic function of the increase in CO_2 concentration relative to the preindustrial level or, equivalently, as a logarithmic function of the amount of carbon in the atmosphere relative to the amount in preindustrial times. Let S_t and $\bar{S}$, respectively, denote the current and preindustrial amounts of carbon in the atmosphere. Then, forcing can be well approximated by the following equation.[z]

$$F_t = \frac{\eta}{\log 2} \log\left(\frac{S_t}{\bar{S}}\right). \tag{5}$$

The parameter η has a straightforward interpretation: if the amount of carbon in the atmosphere in period t has doubled relative to preindustrial times, forcing is η. If it quadruples, it is 2η, and so forth. An approximate value for η is 3.7, implying that a doubling of the amount of carbon in the atmosphere leads to a forcing of 3.7 watts per square meter on earth.[aa]

[x] Formally, a flow rate per area unit is denoted flux. However, since we deal with systems with constant areas, flows and fluxes are proportional and the terms are used interchangeably.

[y] See Schwartz et al. (2010) who report that if earth were a blackbody radiator with a temperature of $288K \approx 15°C$, an increase in the temperature of 1.1 K would increase the outflow by 3.7 W/m^2, implying $\kappa^{-1} = 1.1/3.7 \approx 0.3$.

[z] This relation was first demonstrated by the Swedish physicist and chemist and 1903 Nobel Prize winner in Chemistry, Svante Arrhenius. Therefore, the relation is often referred to as the *Arrhenius's Greenhouse Law*. See Arrhenius (1896).

[aa] See Schwartz et al. (2014). The value 3.7 is, however, not undisputed. Otto et al. (2013) use a value of 3.44 in their calculations.

We are now ready to present a relation between the long-run change in the earth's average temperature as a function of the carbon concentration in the atmosphere. Combining Eqs. (4) and (5) we obtain

$$T(F_t) = \frac{\eta}{\kappa}\frac{1}{\log 2}\log\left(\frac{S_t}{\bar{S}}\right). \tag{6}$$

As we can see, a doubling of the carbon concentration in the atmosphere leads to an increase in temperature given by $\frac{\eta}{\kappa}$. Using the Planck feedback, $\eta/\kappa \approx 1.1$°C. This is a modest sensitivity, and as already noted very likely too low an estimate of the overall sensitivity of the global climate due to the existence of positive feedbacks.

A straightforward way of including feedbacks in the energy budget is by adding a term to the energy budget. Suppose initially that feedbacks can be approximated by a linear term xT_t, where x captures the marginal impact on the energy budget due to feedbacks. The energy budget now becomes

$$\frac{dT_t}{dt} = \sigma(F + xT_t - \kappa T_t), \tag{7}$$

where we think of κ as solely determined by the Planck feedback. The steady-state temperature is now given by

$$T(F) = \frac{\eta}{\kappa - x}\frac{1}{\ln 2}\ \ln\left(\frac{S}{\bar{S}}\right). \tag{8}$$

Since the ratio $\eta/(\kappa - x)$ has such an important interpretation, it is often labeled the*Equilibrium Climate Sensitivity (ECS)* and we will use the notation λ for it.[ab] Some feedbacks are positive but not necessarily all of them; theoretically, we cannot rule out either $x < 0$ or $x \geq \kappa$. In the latter case, the dynamics would be explosive, which appears inconsistent with historical reactions to natural variations in the energy budget. Also $x < 0$ is difficult to reconcile with the observation that relatively small changes in forcing in the earth's history have had substantial impact on the climate. However, within these bands a large degree of uncertainty remains.

According to the IPCC, the ECS is "likely in the range 1.5–4.5°C," "extremely unlikely less than 1°C," and "very unlikely greater than 6° C."[ac] Another concept, taking some account of the shorter run dynamics, is the *Transient Climate Response (TCR)*. This is the defined as the increase in global mean temperature at the time the CO_2 concentration

[ab] Note that equilibrium here refers to the energy budget. For an economist, it might have been more natural to call λ the *steady-state climate sensitivity*.

[ac] See IPCC (2013, page 81 and Technical Summary). The report states that "likely" should be taken to mean a probability of 66–100%, "extremely unlikely" 0–5%, and "very unlikely" 0–10%.

has doubled following a 70-year period of annual increases of 1%.[ad] IPCC et al. (2013b, Box 12.1) states that the TCR is "likely in the range 1°C–2.5°C" and "extremely unlikely greater than 3°C."

3.1.2 Nonlinearities and Uncertainty

It is important to note that the fact that $\frac{1}{\kappa - x}$ is a nonlinear transformation of x has important consequences for how uncertainty about the strength of feedbacks translate into uncertainty about the equilibrium climate sensitivity.[ae] Suppose, for example, that the uncertainty about the strength in the feedback mechanism can be represented by a symmetric triangular density function with mode 2.1 and endpoints at 1.35 and 2.85. This is represented by the upper panel of Fig. 4. The mean, and most likely, value of x translates into a climate sensitivity of 3. However, the implied distribution of climate sensitivities is severely skewed to the right.[af] This is illustrated in the lower panel, where $\frac{\eta}{\kappa - x}$ is plotted with $\eta = 3.7$ and $\kappa = 0.3^{-1}$.

The models have so far assumed linearity. There are obvious arguments in favor of relaxing this linearity. Changes in the albedo due to shrinking ice sheets and abrupt weaking of the Gulf are possible examples.[ag] Such effects could simply be introduced by making x in (7) depend on temperature. This could for example, introduce dynamics with so-called *tipping points*. Suppose, for example, that

$$x = \begin{cases} 2.1 \text{ if } T < 3^{o}C \\ 2.72 \text{ else} \end{cases}$$

Using the same parameters as earlier, this leads to a discontinuity in the climate sensitivity. For CO_2, concentrations below $2 \times \bar{S}$ corresponding to a global mean temperature deviation of 3 degrees, the climate sensitivity is 3. Above that tipping point, the climate sensitivity is 6. The mapping between $\frac{S_t}{\bar{S}}$ and the global mean temperature using Eq. (6) is shown in Fig. 5.

[ad] This is about twice as fast as the current increases in the CO_2 concentration. Over the 5, 10, and 20 year-periods ending in 2014, the average increases in the CO_2 concentration have been 0.54%, 0.54%, and 0.48% per year, respectively. However, note that also other greenhouse gases, in particular methane, affect climate change. For data, see the Global Monitor Division of the Earth System Research Labroratory at the US Department of Commerce.

[ae] The presentation follows Roe and Baker (2007).

[af] The policy implications of the possibility of a very large climate sensitivity is discussed in Weitzman (2011).

[ag] Many state-of-the-art climate models feature regional tipping points; see Drijfhouta et al. (2015) for a list. Currently, there is, however, no consensus on the existence of specific global tipping points at particular threshold levels; see Lenton et al. (2008), Levitan (2013), and IPCC (2013, Technical Summary page 70).

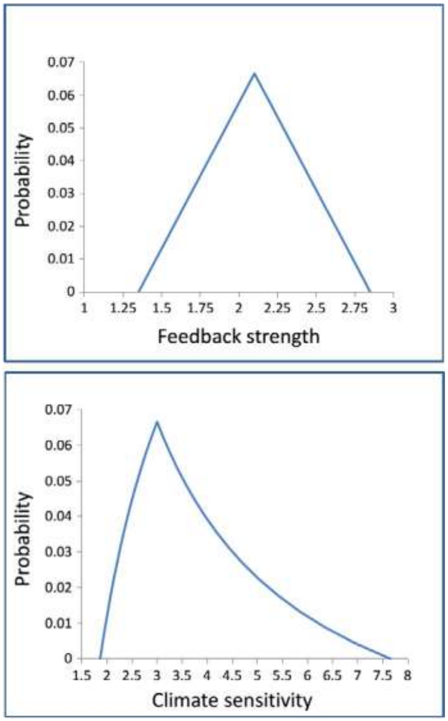


Fig. 4 Example of symmetric uncertainty of feedbacks producing right-skewed climate sensitivity.

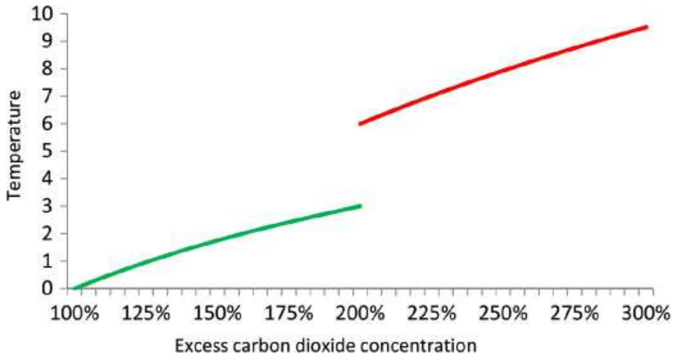


Fig. 5 Tipping point at 3 K due to stronger feedback.

It is also straightforward to introduce irreversibilities, for example by assuming that feedbacks are stronger (higher x) if a state variable like temperature or CO_2 concentration has ever been above some threshold value.

3.1.3 Ocean Drag

We have presented the simplest possible model of how the CO_2 concentration determines climate change. There are of course endless possibilities of extending this simplest framework. An example is to include another energy-budget equation. In Eqs. (3) and (7), we described *laws of motion* for the atmospheric temperature, which heats much faster than the oceans. During the adjustment to a steady state, there will be net energy flows between the ocean and the atmosphere. Let T_t and T_t^L, respectively, denote the atmospheric and ocean temperatures in period t, both measured as deviations from the initial (preindustrial) steady state. With two temperatures, we can define energy budgets separately for the atmosphere and for the oceans. Furthermore, allow for a variation in forcing over time and let F_t denote the forcing at time t. We then arrive at an extended version of Eq. (7) given by

$$\frac{dT_t}{dt} = \sigma_1\left(F_t + xT_t - \kappa T_t - \sigma_2\left(T_t - T_t^L\right)\right). \tag{9}$$

Comparing (9) to (7), we see that the term $\sigma_2\left(T_t - T_t^L\right)$ is added. This term represents a new flow in the energy budget (now defined specifically for the atmosphere), namely the net energy flow from the atmosphere to the ocean. To understand this term, note that if the ocean is cooler than the atmosphere, energy flows from the atmosphere to the ocean. This flow is captured in the energy budget by the term $-\sigma_2\left(T_t - T_t^L\right)$. If $T_t > T_t^L$, this flow has a negative impact on the atmosphere's energy budget and likewise on the rate of change in temperature in the atmosphere (the LHS). The cooler is the ocean relative to the atmosphere, the larger is the negative impact on the energy budget.

To complete this dynamic model, we need to specify how the ocean temperature evolves by using the energy budget of the ocean. If the temperature is higher in the atmosphere than in the oceans, energy will flow to the oceans, thus causing an increase in the ocean temperature. Expressing this as a linear equation delivers

$$\frac{dT_t^L}{dt} = \sigma_3\left(T_t - T_t^L\right). \tag{10}$$

Eqs. (9) and (10) together complete the specification of how the temperatures of the atmosphere and the oceans are affected by a change in forcing.

We can simulate the behavior of the system once we specify the parameters of the system (σ_1, σ_2, σ_3, and κ all positive) and feed in a sequence of forcing levels F_t. Nordhaus and Boyer (2000) use $\sigma_1 = 0.226$, $\sigma_2 = 0.44$, and $\sigma_3 = 0.02$ for a discrete-time version of (9) and (10) defined as the analogous difference equations with a 10-year step. In (6) we show the dynamic response of this model to a constant forcing of $1\,W/m^2$ for $(\kappa - x)^{-1} = 0.81$. The lower curve represents the ocean temperature T_t^L, which increases quite slowly. The middle curve is the atmospheric temperature, T_t, which increases more quickly (Fig. 6).

Clearly, the long-run increase in both temperatures is given by $\frac{1}{\kappa}$ times the increase in forcing, ie, by 0.81°C. Most of the adjustment to the long-run equilibrium is achieved after a few decades for the atmosphere but takes several hundred years for the ocean temperature. Without the dragging effect of the oceans, the temperature increases faster, as

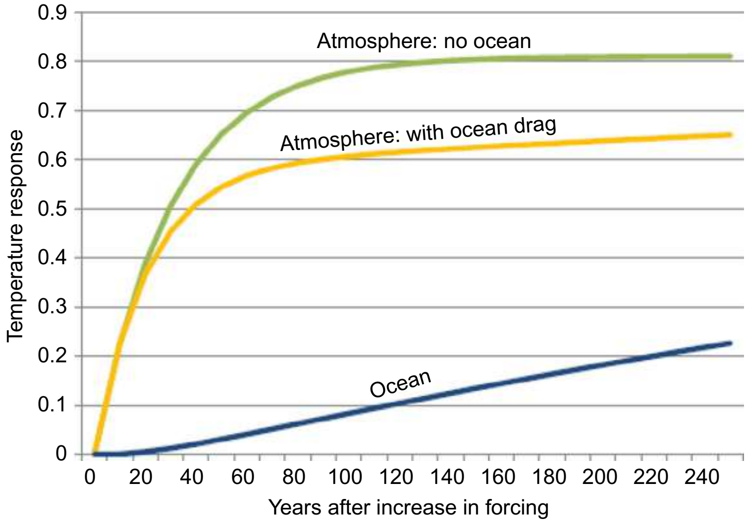


Fig. 6 Increase in atmospheric and ocean temperatures after a permanent forcing of $1W/m^2$.

shown by the top curve where we have set $\sigma_2 = 0$, which shuts down the effect of the slower warming of the ocean. However, we see that the time until half of the adjustment is achieved is not very different in the two cases.

3.1.4 Global Circulation Models

The climate models discussed so far are extremely limited in scope from the perspective of a climate scientist. In particular, they are based on the concept of an energy budget. Such models are by construction incapable of predicting the large disparity in climates over the world. For this, substantially more complex general circulation models (GCMs) need to be used. Such models are based on the fact that the energy flow to earth is unevenly spread over the globe both over time and space. This leads to movements in air and water that are the drivers of weather events and the climate. These models exist in various degrees of complexity, often with an extremely large number of state variables.[ah]

The complexity of general circulation models make them difficult to use in economics. In contrast to systems without human agents, such models do not contain any forward-looking agents. Thus, causality runs in one time direction only and the evolution of the system does not depend on expectations about the future. Therefore, solving such a complex climate model with a very large set of state variables may pose difficulties—in practice, because they are highly nonlinear and often feature chaotic behavior—but not the kind of difficulties economists face when solving their dynamic models.

[ah] See IPCC (2013, chapter 9) for a list and discussion of GCMs.

One way of modeling a heterogeneous world climate that does not require a combination of a very large state space and forward-looking behavior builds on *statistical downscaling*.[ai] The output of large-scale dynamic circulation models or historical data is then used to derive a *statistical* relation between aggregate and disaggregated variables. This is in contrast to the actual nonlinear high-dimensional models because they do not feature randomness; the model output only looks random due to the nonlinearities. The basic idea in statistical downscaling is thus to treat a small number of state variables as sufficient statistics for a more detailed description of the climate. This works well due in part to the fact that climate change is ultimately driven by a global phenomenon: the disruption of the energy balance due to the release of green house gases, where CO_2 plays the most prominent role.

Let $T_{i,\,t}$ denote a particular measure of the climate, eg, the yearly average temperature, in region i in period t. We can then estimate a model like

$$\begin{aligned} T_{i,t} &= \bar{T}_i + f(l_i, \psi_1) T_t + z_{i,t} \\ z_{i,t} &= \rho z_{i,t-1} + \nu_{i,t} \\ \mathrm{var}(\nu_{i,t}) &= g(l_i, \psi_2) \\ \mathrm{corr}\left(\nu_{i,t}, \nu_{j,t}\right) &= h\left(d\left(l_i, l_j\right), \psi_3\right). \end{aligned}$$

This very simple system, used for illustration mainly, explains downscaling conceptually. Here, $\bar{T}_i$ is the baseline temperature in region i. f, g, and h are specified functions parameterized by ψ_1, ψ_2, and ψ_3. $z_{i,t}$ is the prediction error and it is assumed to follow an AR(1) process. l_i is some observed characteristic of the region, eg, latitude, and $d\left(l_i, l_j\right)$ is a distance measure. Krusell and Smith (2014) estimate such a model on historical data. The upper panel in Fig. 7 shows the estimated function f with l_i denoting latitude. We see that an increase in the global mean temperature T_t has an effect on regional temperature levels that depends strongly on the latitude. The effect of a 1°C increase in the global temperature ranges from 0.25°C to 3.6°C. The lower panel in the figure shows the correlation pattern of prediction errors using d to measure Euclidian distance.

Now consider a dynamic economic model (where agents are forward-looking) with a small enough number of state variables that the model can be solved numerically. With one of these state variables playing the role of global temperature in the above equation system, one can imagine adding a large amount of heterogeneity without losing tractability, so long as the heterogeneous climate outcomes (eg, the realization of the local temperature distribution) do not feed back into global temperature. This is the approach featured in Krusell and Smith (2015), whose model can be viewed as otherwise building directly on the models (static and dynamic) presented in the sections later in this chapter.[aj]

[ai] See IPCC (2013, chapter 9) for a discussion of statistical downscaling.

[aj] Krusell and Smith (2015) actually allow some feedback, through economic variables, from the temperature distribution on global temperature but develop numerical methods that nevertheless allow the model to be solved.

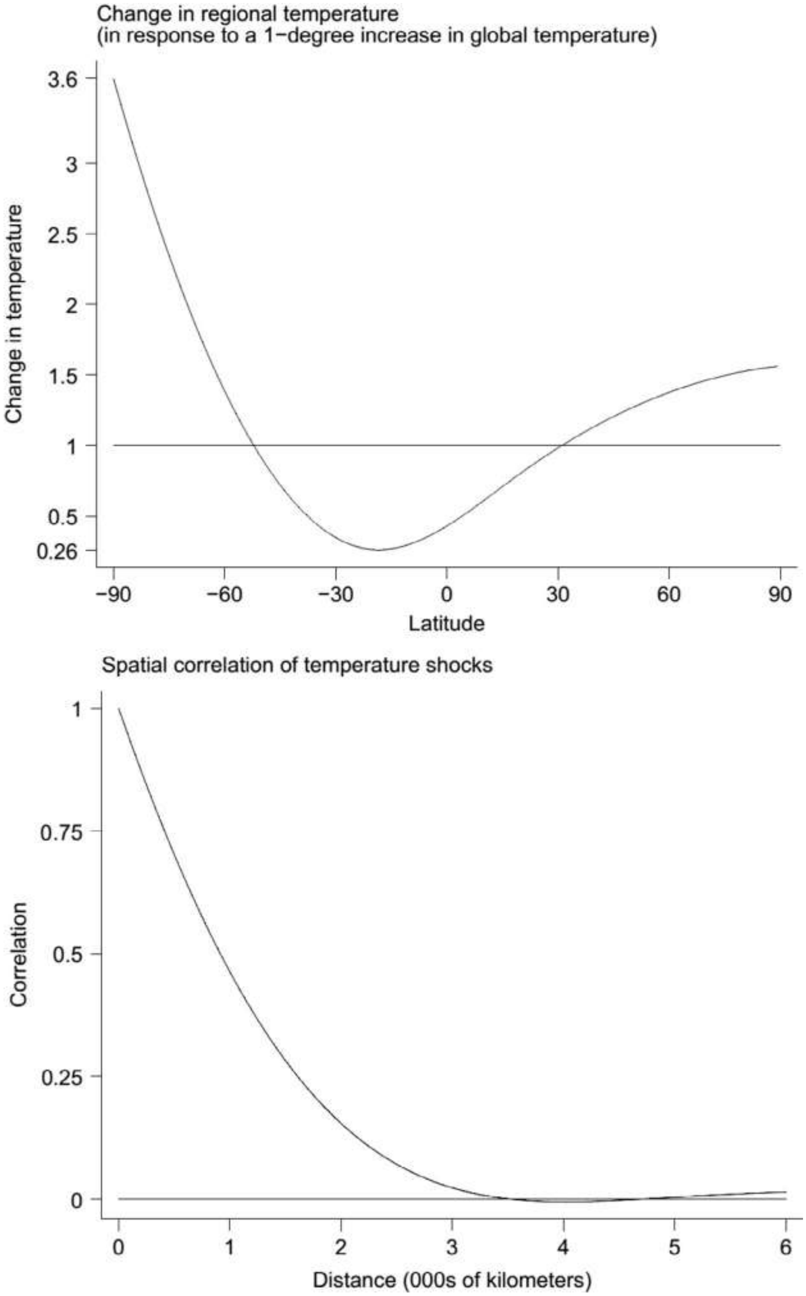


Fig. 7 Statistical downscaling: regional climate responses to global temperature.

3.2 Carbon Circulation

We now turn to carbon circulation (also called the carbon cycle). The purpose of the modeling here is to produce a mapping between emissions of CO_2 and the path of the CO_2 concentration in the atmosphere. The focus on CO_2 is due to the fact that while other gases emitted by human activities, in particular methane, are also important contributors to the greenhouse effect, CO_2 leaves the atmosphere much more slowly. The half-life of methane is on the order of 10 years, while as we will see, a sizeable share of emitted CO_2 remains in the atmosphere for thousands of years.[ak]

3.2.1 Carbon Sinks and Stores

The burning of fossil fuel leads to emissions of carbon dioxide into the atmosphere. The carbon then enters into a circulation system between different global reservoirs of carbon (carbon sinks) of which the atmosphere is one. In Fig. 8, the carbon reservoirs are represented by boxes. The number in black in each box indicates the size of the reservoir in GtC, ie, billions of tons of carbon. As we can see, the biggest reservoir by far is the intermediate/deep ocean, with more than 37,000 GtC. The vegetation and the atmosphere are of about the same size, around 600 GtC, although the uncertainty about the former is substantial. Soils represent a larger stock as does carbon embedded in the permafrost. Black arrows in the figure indicate preindustrial flows between the stocks measured in GtC per year. The flows between the atmosphere and the ocean were almost balanced, implying a constant atmospheric CO_2 concentration.

By transforming carbon dioxide into organic substances, vegetation in the earth's biosphere induces a flow of carbon from the atmosphere to the biosphere. This is the photosynthesis. The reverse process, respiration, is also taking place in plants' fungi, bacteria, and animals. This, together with oxidation, fires, and other physical processes in the soil, leads to the release of carbon in the form of CO_2 to the atmosphere. A similar process is taking place in the sea, where carbon is taken up by phytoplankton through photosynthesis and released back into the surface ocean. When phytoplankton sink into deeper layers they take carbon with them. A small fraction of the carbon that is sinking into the deep oceans is eventually buried in the sediments of the ocean floor, but most of the carbon remains in the circulation system between lower and higher ocean water. Between the atmosphere and the upper ocean, CO_2 is exchanged directly. Carbon dioxide reacts with water and forms dissolved inorganic carbon that is stored in the water. When the CO_2-rich surface water cools down in the winter, it falls to the deeper ocean and a similar exchange occurs in the other direction. From the figure, we also note that there are large flows of carbon between the upper layers of the ocean and the atmosphere via gas exchange. These flows are smaller than, but of the same order of magnitude as, the photosynthesis and respiration.

[ak] Prather et al. (2012) derive a half-life of methane of 9.1 years with a range of uncertainty of 0.9 years.

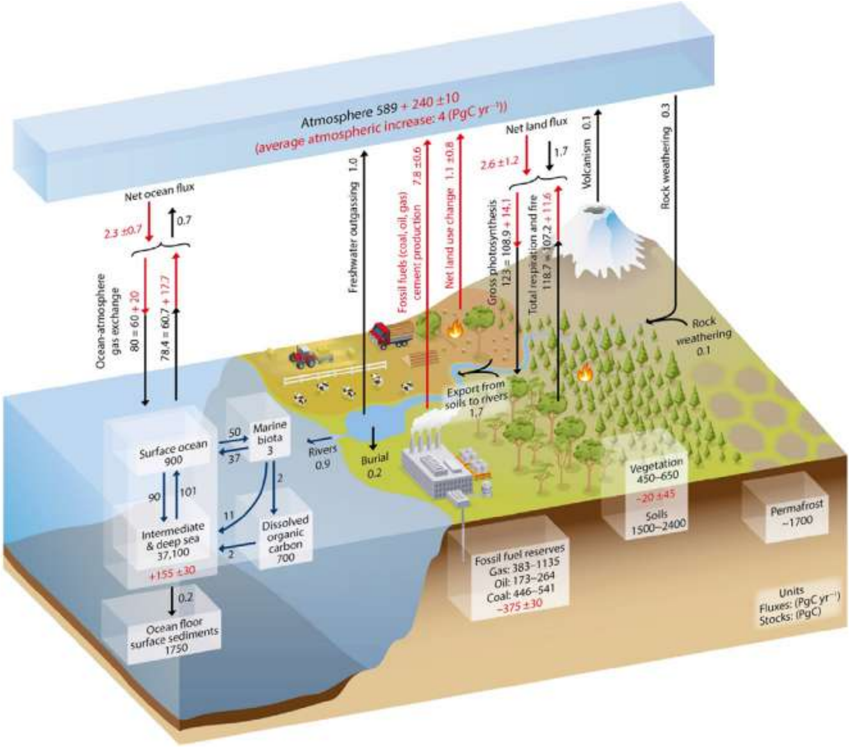

Atmosphere 589 + 240 ±10
(average atmospheric increase: 4 (PgC yr⁻¹))
Net ocean flux
2.3 ±0.7
0.7
Ocean-atmosphere gas exchange
80 = 60 + 20
78.4 = 60.7 + 17.7
Freshwater outgassing 1.0
Fossil fuels (coal, oil, gas) cement production 7.8 ±0.6
Net land use change 1.1 ±0.8
Net land flux
2.6 ±1.2
1.7
Gross photosynthesis 123 = 108.9 + 14.1
Total respiration and fire 118.7 = 107.2 + 11.6
Volcanism 0.1
Rock weathering 0.3
Rock weathering 0.1
Export from soils to rivers 1.7
Surface ocean 900
50
37
Marine biota 3
Rivers 0.9
Burial 0.2
90
101
2
11
Intermediate & deep sea 37,100
+155 ±30
Dissolved organic carbon 700
2
0.2
Ocean floor surface sediments 1750
Fossil fuel reserves
Gas: 383–1135
Oil: 173–264
Coal: 446–541
−375 ±30
Vegetation 450–650
−20 ±45
Soils 1500–2400
Permafrost ~1700
Units
Fluxes: (PgC yr⁻¹)
Stocks: (PgC)

Fig. 8 Global carbon cycle. Stocks in GtC and flows GtC/year. *IPCC, Stocker, T.F., Qin, D., Plattner, G.K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M., 2013. Climate Change 2013: the Physical Science Basis. Cambridge University Press, Cam- Q16 bridge, UK, fig. 6.1).*

3.2.2 Human Influence on Carbon Circulation

Before the industrial revolution, human influence on carbon circulation was small. However, atmospheric CO_2 concentration started to rise from the mid-18th century and onwards, mainly due to the burning of fossil fuels and deforestation but also as a result of rising cement production.

In Fig. 8, the red figures denote changes in the reservoirs and flows over and above preindustrial values. The figures for reservoirs refer to 2011 while flows are yearly averages during the period 2000–09. At the bottom of the picture, we see that the stock of fossil fuel in the ground has been depleted by 365 ± 30 GtC since the beginning of industrialization. The flow to the atmosphere due to fossil-fuel use and cement production is reported to be 7.8 ± 0.6 GtC per year. In addition, changed land use adds 1.1 ± 0.8 GtC per year to the flow of carbon to the atmosphere. In the other direction, the net flows from the atmosphere to the terrestrial biosphere and to the oceans have increased. All in all, we note that while the fossil reserves have shrunk, the amount of carbon in the atmosphere has gone from close to 600 to around 840 GtC and currently increases at a rate of 4 GtC per year. A sizeable but somewhat smaller increase has taken place in the oceans while the amount of carbon in the vegetation has remained largely constant.

We see that the gross flows of carbon are large relative to the additions due to fossil-fuel burning. Furthermore, the flows may be indirectly affected by climate change, creating feedback mechanism. For example, the ability of the biosphere to store carbon is affected by temperature and precipitation. Similarly, the ability of the oceans to store carbon is affected by the temperature. Deposits of carbon in the soil may also be affected by climate change. We will return to these mechanisms below.

3.2.3 The Reserves of Fossil Fuel

The extent to which burning of fossil fuel is a problem from the perspective of climate change obviously depends on how much fossil fuel remains to (potentially) be burnt. This amount is not known and the available estimates depends on definitions. The amount of fossil resources that eventually can be used depends on estimates of future findings as well as on forecasts about technological developments and relative prices. Often, reserves are defined in successively wider classes. For example, the US Energy Information Agency defines four classes for oil and gas. The smallest is *proved reserves*, which are reserves that geologic and engineering data demonstrate with reasonable certainty to be recoverable in future years from known reservoirs under existing economic and operating conditions. As technology and prices change, this stock normally increases over time. Successively larger ones are *economically recoverable resources*, *technically recoverable resources*, and *remaining oil and natural gas in place*.

Given different definitions and estimation procedures the estimated stocks differ and will change over time. Therefore, the numbers in this section can only be taken as indications. Furthermore, reserves of different types of fossil fuels are measured in different units, often barrels for oil, cubic meters or cubic feet for gas, and tons for coal. However, for our purpose, it is convenient to express all stocks in terms of their carbon content. Therefore nontrivial conversion must be undertaken. Given these caveats, we calculate from BP (2015) global proved reserves of oil and natural gas to be approximately 200 GtC and 100 GtC, respectively.[al] At current extraction rates, both these stocks would last approximately 50 years. Putting these numbers in perspective, we note that the

[al] BP (2015) reports proved oil reserves to 239,8 Gt. For conversion, we use IPCC (2006, table 1.2 and 1.3). From these, we calculate a carbon content of 0.846 GtC per Gt of oil. BP (2015) reports proved natural gas reserves to be 187.1 trillion m^3. The same source states an energy content of 35.7 trillion BtU per trillion m^3 equal to 35.9 trillion kJ. IPCC (2006) reports 15.3 kgC/GJ for natural gas. This means that 1 trillion m^3 natural gas contains 0.546 GtC. For coal, we use the IPCC (2006) numbers for antracite, giving 0.716 GtC per Gt of coal. For all these conversions, it should be noted that there is substantial variation in carbon content depending on the quality of the fuel and the numbers used must therefore be used with caution.

atmosphere currently contains over 800 GtC. Given the results in the previous sections, we note that burning all proved reserves of oil and natural gas would have fairly modest effects on the climate.[am] Again using BP (2015), we calculate proved reserves of coal to around 600 GtC, providing more potential dangers for the climate.

Using wider definitions of reserves, stocks are much larger. Specifically, using data from McGlade and Ekins (2015) we calculate ultimately recoverable reserves of oil, natural gas and coal to close to 600 GtC, 400 GtC and 3000 GtC.[an] Rogner (1997) estimates coal reserves to be 3500 GtC with a marginal extraction cost curve that is fairly flat. Clearly, if all these reserves are used, climate change can hardly be called modest.

3.2.4 A Linear Carbon Circulation Model

A natural starting point is a linear carbon circulation model. Let us begin with a two-stock model as in Nordhaus and Boyer (2000). We let the variables S_t and S_t^L denote the amount of carbon in the two reservoirs, respectively: S_t for the atmosphere and S_t^L for the ocean. Emissions, denoted E_t, enter into the atmosphere. Under the linearity assumption, we assume that a constant share ϕ_1 of S_t flows to S_t^L per unit of time and, conversely, a share ϕ_2 of S_t^L flows in the other direction implying

$$\begin{aligned} \frac{dS_t}{dt} &= -\phi_1 S_t + \phi_2 S_t^L + E_t, \\ \frac{dS_t^L}{dt} &= \phi_1 S_t - \phi_2 S_t^L. \end{aligned} \tag{11}$$

Eq. (11) form a linear system of differential equations, similar to Eqs. (9) and (10). However, there is a key difference: additions of carbon to this system through emissions get "trapped" in the sense that there is no outflow from the system as a whole, reflecting the fact that one of the characteristic roots of the system in (11) is zero.[ao] This implies that if E settles down to a positive constant, the sizes of the reservoirs S and S^L will not approach a steady state, but will grow forever. If emissions eventually stop and remain zero, the sizes of the reservoirs will settle down to some steady-state values, but these values will depend on the amount of emissions accumulated before that. This steady state satisfies a zero net flow as per

$$0 = -\phi_1 S + \phi_2 S^L, \tag{12}$$

[am] As we will soon see, a substantial share of burned fossil fuel quickly leaves the atmosphere.

[an] See footnote al for conversions.

[ao] If we were to also define a stock of fossil fuel in the ground from which emissions are taken, total net flows would be zero. Since it is safe to assume that flows into the stock of fossil fuel are negligible, we could simply add an equation $\frac{dR_t}{dt} = -E_t$ to the other equations, which would thus capture the depletion of fossil reserves.

implying that

$$\frac{S}{S^L} = \frac{\phi_2}{\phi_1}$$

and that the rate of convergence is determined by the nonzero root $-(\phi_1 + \phi_2)$.

As we have seen above, CO_2 is mixed very quickly into the atmosphere. CO_2 also passes quickly through the ocean surface implying that a new balance between the amount of carbon in the atmosphere and the shallow ocean water is reached quickly.[ap] The further transport of carbon to the deep oceans is much slower, motivating a third model reservoir: the deep oceans. This is the choice made in recent versions of the DICE and RICE models (Nordhaus and Sztorc, 2013), which use a three-reservoir linear system similar to (11).

3.2.5 Reduced-Form Depreciation Models

Although the stock-flow model has a great deal of theoretical and intuitive appeal, it runs the risk of simplifying complicated processes too much. For example, the ability of the terrestrial biosphere to store carbon depends on temperature and precipitation. Therefore, changes in the climate may have an effect on the flows to and from the biosphere not captured in the model described earlier. Similarly, the storage capacity of the oceans depends (negatively) on the temperature. These shortcomings could possibly be addressed by including temperature and precipitation as separate variables in the system. Furthermore, also the processes involved in the deep oceans are substantially more complicated than what is expressed in the linear model. In particular, the fact that carbon in the oceans exists in different chemical forms and that the balance between these has an important role for the dynamics of the carbon circulation is ignored but can potentially be of importance.

An important problem with the linear specification (see, Archer, 2005 and Archer et al., 2009) is due to the so-called Revelle buffer factor (Revelle and Suess, 1957). As CO_2 is accumulated in the oceans, the water is acidified. This dramatically limits its capacity to absorb more CO_2, making the effective "size" of the oceans as a carbon reservoir decrease by approximately a factor of 15 (Archer, 2005). Very slowly, the acidity decreases and the preindustrial equilibrium can be restored. This process is so slow, however, that it can be ignored in economic models. IPCC (2007, p. 25, Technical Summary), take account of the Revelle buffer factor and conclude that "About half of a CO_2 pulse to the atmosphere is removed over a time scale of 30 years; a further 30% is removed within a few centuries; and the remaining 20% will typically stay in the atmosphere for many thousands of years." The conclusion of Archer (2005) is that

[ap] This takes 1–2 years IPCC (2013).

a good approximation is that 75% of an excess atmospheric carbon concentration has a mean lifetime of 300 years and the remaining 25% remain several thousands of years.[aq]

A way of representing this is to define a depreciation model. Golosov et al. (2014) define a carbon depreciation function. Let $1-d(s)$ represent the amount of a marginal unit of emitted carbon that remains in the atmosphere after s periods. Then postulate that

$$1-d(s)=\varphi_L+(1-\varphi_L)\varphi_0(1-\varphi)^s. \tag{13}$$

The three parameters in (13) are easily calibrated to match the three facts in the earlier IPCC quote; we do this in Section 5. A similar approach is described in IPCC (2007a, table 2.14). There,

$$1-d(s)=a_0+\sum_{i=1}^{3}\left(a_i e^{-\frac{s}{\tau_i}}\right), \tag{14}$$

with $a_0=0.217$, $a_1=0.259$, $a_2=0.338$, $a_3=0.186$, $\tau_1=172.9$, $\tau_2=18.51$, and $\tau_3=1.186$, where s and the τ_is are measured in years. With this parametrization, 50% of an emitted unit of carbon has left the atmosphere after 30 years, 75% after 356 years, and 21.7% stays forever. It is important to note that this depreciation model is appropriate for a marginal emission at an initial CO_2 concentration equal to the current one (around 800 GtC). The parameters of the depreciation function should be allowed to depend on initial conditions and inframarginal future emissions. If emissions are very large, a larger share will remain in the atmosphere for a long time. To provide a measure for how sensitive the parameters are, note that of an extremely large emission pulse of 5000 GtC, which is more than 1 × the current accumulated emissions, around 40% remains after a thousand years, as opposed to half as much for a much smaller pulse.[ar]

3.2.6 A Linear Relation Between Emissions and Temperature

As discussed earlier, it may be too simplistic to analyze the carbon circulation in isolation. The storage capacity of the various carbon sinks depends on how the climate develops. One might think that including these interactions would make the model more complicated. However, this does not have to be the case. In fact, there is evidence that various feedbacks and nonlinearity in the climate and carbon-cycle systems tend to cancel each other out, making the combined system behave in a much simpler and, in fact, linear way.[as] In order to briefly discuss this, let us defined the variable CCR_m (Carbon-Climate Response) as the change in the global mean temperature over some specified time interval m per unit of emissions of fossil carbon into the atmosphere over that same time interval

[aq] Similar findings are reported in IPCC (2013, Box 6.1).

[ar] See IPCC (2013, Box 6.1).

[as] This subsection is based on Matthews et al. (2009).

$$CCR_m \equiv \frac{T_{t+m} - T_t}{\int_t^m E_s ds}.$$

Given our previous discussions in this and the previous sections, one would think that this variable is far from a constant: the dynamic behavior of the climate and the carbon cycle will in general make the CCR_m depend on the length of the time interval considered. For example, since it takes time to heat the oceans, the temperature response could depend on whether the time interval is a decade or a century. Similarly, since also the carbon dynamics are slow, the extra CO_2 concentration induced by a unit of emission tends to be lower the longer the time interval considered. Furthermore, the CCR_m might depend on how much emissions have already occurred; higher previous emissions can reduce the effectiveness of carbon sinks and even turn them into net contributors. The marginal effect on temperature from an increase in the CO_2 concentration also depends on the level of CO_2 concentration due to the logarithmic relation between CO_2 concentration and the greenhouse effect.

Quite surprisingly, Matthews et al. (2009) show that the dynamic and nonlinear effects tend to cancel, making it a quite good approximation to consider the CCR_m as a constant, CCR, independent of both the time interval considered and the amount of previous emissions. Of course, knowledge about the value of CCR is incomplete but Matthews et al. (2012) quantify this knowledge gap and argue that a 90% confidence interval is between 1 and 2.5°C per 1000 GtC.[at] This means that we can write the (approximate) linear relationship

$$T_{t+m} = T_t + CCR \int_t^m E_s ds.$$

To get some understanding for this surprising result, first consider the time independence. We have shown in the previous chapter that when the ocean is included in the analysis, there is a substantial delay in the temperature response of a given forcing. Thus, if the CO_2 concentration permanently jumps to a higher level, it takes many decades before even half the final change in temperature has taken place. On the other hand, if carbon is released into the atmosphere, a large share of it is removed quite slowly from the atmosphere. It happens to be the case that these dynamics cancel each other, at least if the time scale is from a decade up to a millennium. Thus, in the shorter run, the CO_2 concentration and thus forcing is higher but this is balanced by the cooling effect of the oceans.

Second, for the independence of CCR with respect to previous emissions note that the Arrhenius law discussed in the previous chapter implies a logarithmic relation

[at] IPCC (2013) defines the very similar concept, the Transient Climate Response to cumulative carbon Emissions (TCRE), and states that it is likely between 0.8 and 2.5°C per 1000 GtC for cumulative emissions below 2000 GtC.

between CO_2 concentration and the temperature. Thus, at higher CO_2 concentrations, an increase in the CO_2 concentration has a smaller effect on the temperature. On the other hand, existing carbon cycle models tend to have the property that the storage capacity of the sinks diminishes as more CO_2 is released into the atmosphere. These effects also balance—at higher levels of CO_2 concentration, an additional unit of emissions increases the CO_2 concentration more but the effect of CO_2 concentration on temperature is lower by roughly the same proportion.

Given a value of CCR, it is immediate to calculate how much more emissions can be allowed in order to limit global warning to a particular value. Suppose, for example, we use a value of $CCR = 1.75$. Then, to limit global warming to 2°C, we cannot emit more than $(2/1.75) \times 1000 = 1140$ GtC, implying that only around 600 GtC can be emitted in the future. If, on the other hand, we use the upper limit of the 95% confidence interval ($CCR = 2.5$) and aim to reduce global warming to 2°C, accumulated emissions cannot be more than a total of 800 GtC of which most is already emitted.

3.3 Damages

In this section, we discuss how the economy is affected by climate change. Since economic analysis of climate change tends to rely on cost-benefit calculation, it is not only a necessary cornerstone of the analysis but arguably also a key challenge for climate economics. For several reasons, this is a very complicated area, however. First, there is an almost infinite number of ways in which climate change can affect the economy. Second, carbon emissions are likely to affect the climate for a very long time: for thousands of years. This implies that the quantitative issue of what weight to attach to the welfare of future generations becomes of key importance for the valuation. Third, global climate change can potentially be much larger than experienced during the modern history of mankind. Historical relations between climate change and the economy must therefore be extrapolated significantly if they are to be used to infer the consequences of future climate change. Fourth, many potential costs are to goods and services without market prices.

The idea that the climate affect the economy is probably as old as the economy itself, or rather as old as mankind. That the distribution of weather outcomes—the climate—affects agricultural output must have been obvious for humans since the Neolithic revolution. The literature on how the climate affects agriculture is vast and not reviewed here. It is also well known that in a cross-country setting, a hotter climate is strongly associated with less income per capita. Also within countries, such a negative relation between temperature and income per capita can be found (Nordhaus, 2006). However, Nordhaus (2006) also finds a hump-shaped relation between output density, ie, output per unit of land area, and average temperature. This suggests that a method of adaptation is geographic mobility. An overview is provided in Tol (2009). A more recent economic literature using modern methods emphasizing identification is now rapidly expanding. The focus is broad and climate change is allowed to have many different effects, including a heterogeneous effect on the economic productivity of different production sectors, effects on health, mortality, social unrest, conflicts, and much more. Dell et al. (2014) provide an overview of this newer literature.

Climate change thus likely has extremely diverse effects, involving a large number of different mechanisms affecting different activities differently. The effects are spatially heterogeneous and have different dynamics. Despite this, it appears important to aggregate the effects to a level that can be handled by macroeconomic models.[au]

3.3.1 Nordhaus's Approach

Early attempts to aggregate the economic impacts of climate change were carried out in Nordhaus (1991).[av] Nordhaus (1992, 1993) constructed the path-breaking integrated assessment model named DICE, ie, a model with the three interlinked systems—the climate, the carbon cycle, and the economy.[aw] This is a global growth model with carbon circulation, and climate module, and a damage function. This very early incarnation of the damage function assumed that the economic losses from global warming were proportional to GDP and a function of the global mean temperature, measured as a deviation from the preindustrial average temperature. Nordhaus's assumption in the first version of DICE was that the fraction of output lost was

$$D(T) = 0.0133\left(\frac{T}{3}\right)^2.$$

Nordhaus underlines the very limited knowledge that supported this specification. His own study (Nordhaus, 1991) studies a number of activities in the United States and concludes that these would contribute to a loss of output of 0.25% of US GDP for a temperature deviation of 3°C. He argues that a reasonable guess is that the this estimate omits important factors and that the United States losses rather are on the order of 1% of GDP and that the global losses are somewhat larger. Nordhaus (1992) cites Cline (1992) for an estimate of the power on temperature in the damage function but chooses 2 rather than the cited 1.3.

Later work (Nordhaus and Boyer, 2000) provided more detailed sectorial estimates of the damage function. Here, the aggregation includes both damages that accrue to market activities and those that could affect goods, services, and other values that are not traded.

[au] Macroeconomic modeling with large degrees of heterogeneity is developing rapidly, however. In the context of climate economy modeling, see eg, Krusell and Smith (2015) for a model with nearly 20,000 regions.

[av] Other early examples are Cline (1992), Fankhauser (1994), and Titus (1992).

[aw] DICE stands for Dynamic Integrated Climate-Economy model.

An attempt to value the risk of catastrophic consequences of climate change is also included. Obviously, this is an almost impossible task, given the little quantitative knowledge about tail risks. Nordhaus and Boyer use a survey, where climate experts are asked to assess the probability of permanent and dramatic losses of output at different increases in the global mean temperature.

The latest version of DICE (Nordhaus and Sztorc, 2013) instead goes back to a more ad-hoc calibration of the damage function. Based on results in a survey in Tol (2009) and IPCC (2007b) depicted in Fig. 9, they postulate a damage function given by

$$D(T) = 1 - \frac{1}{1 + 0.00267T^2}. \tag{15}$$

The functional form in (15) is chosen so that damages are necessarily smaller than 1 but for the intended ranges of temperature, it may be noted that $1 - \frac{1}{1+0.00267T^2} \approx 0.023\left(\frac{T}{3}\right)^2$.[ax] Thus, the functional form remains similar to the first version of DICE but the estimated damages at three degrees have increased from 1.3% to 2.3% of global GDP.

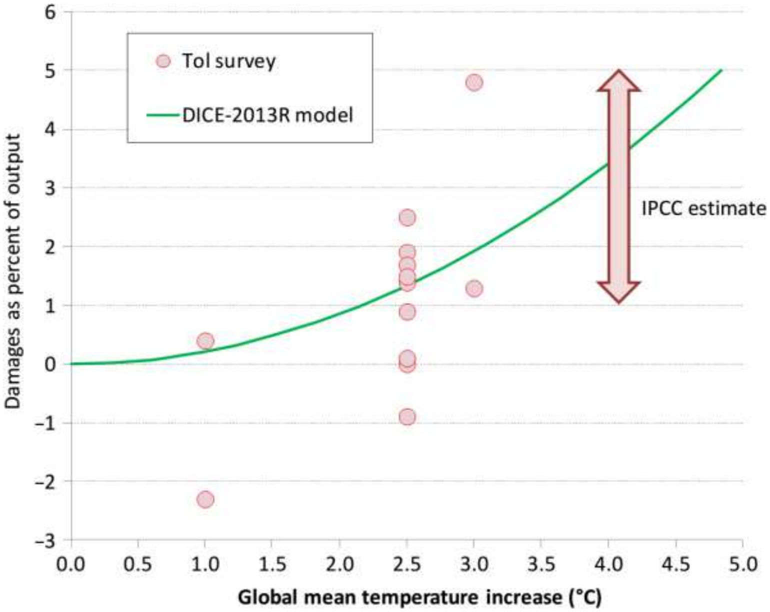


Fig. 9 Global damage estimates. Dots are from Tol (2009). The solid line is the estimate from the DICE-2013R model. The arrow is from the IPCC (2007b, page 17). *Reprinted from Nordhaus, W.D., Sztorc, P., 2013. DICE 2013R: introduction and users manual. Mimeo, Yale University.*

[ax] It is important to note that Nordhaus and Sztorc (2013) warn against using their damage function for temperature deviations over 3°C.

Nordhaus has also developed models with multiple regions, RICE (Regional Integrated Climate-Economy model). The later versions of this model have different damage functions defined for 12 regions. Here, the linear-quadratic function of the global mean temperature is appended with a threshold effect at a four-degree temperature deviation: at this level, the exponent on the temperature is increased to six. Separate account is also taken for sea-level rise, whose damages are described using a linear-quadratic function.

Similar aggregate damage functions are used in other global integrated assessment models; prominent examples are WITCH, FUND, and PAGE.[ay]. Specifically, WITCH has quadratic, region-specific damage functions for eight global regions. FUND uses eight different sectorial damage functions defined for each of 16 regions. PAGE, which was used in the highly influential Stern report (Stern, 2007), uses four separate damage functions for different types of damages in each of eight regions. A special feature of the damage functions in FUND is that the exponent on the global mean temperature is assumed to be a random variable in the interval [1.5–3].

3.3.2 Explicit Damage Aggregation

The damage functions described so far has only been derived to a limited degree from a "bottom-up approach" where explicit damages to particular regions and economic sectors are defined and aggregated. To the extent that such an approach has been used, the final results have been adjusted in an ad-hoc manner, often in the direction of postulating substantially larger damages than found in the explicit aggregation. Furthermore, the work has abstracted from general-equilibrium effects and simply added estimated damages sector by sector and region by region. Obviously this is problematic as the welfare consequences of productivity losses to a particular sector in a particular region depend on the extent to which production can move to other regions or be substituted for by other goods.

An example of a detailed high-resolution modeling of climate damages where (regional) general equilibrium effects are taken into account is the PESETA project, initiated by the European Commission.[az] Damages estimated are for coastal damages, flooding, agriculture, tourism, and health in the European Union. A reference scenario there is a 3.1°C increase in the temperature in the EU by the end of this century relative to the average over 1961–90. The resulting damages imply an EU-wide loss of 1.8% of GDP. The largest part of this loss is due to higher premature mortality in particular in south-central EU.[ba] In the northern parts of the EU, welfare gains associated mainly with lower energy expenditures are approximately balanced by negative impacts in

[ay] See Bosetti et al. (2006), Tol (1995), and Chris et al. (1993) for descriptions of WITCH, FUND, and PAGE, respectively.

[az] See Ciscar et al. (2011) for a short description.

[ba] France, Austria, Czech Republic, Slovakia, Hungary, Slovenia, and Romania.

human health and coastal area damages.[bb] Clearly, these effects are small relative to the expectations for economic growth over this period as well as compared to fears of dramatic impacts often expressed in the policy debate about climate change.

3.3.3 Top-Down Approaches

An alternative approach to the bottom-up approach is to estimate a reduced-form relation between aggregate measures like GDP, consumption, and investments and climate. The idea here is to associate natural historical variation in climate to changes in the aggregate variables of interest. Most of this work thus focuses on short-run changes in temperature as opposed to climate change. Examples of this approach are Dell et al. (2012) who examine how natural year-to-year variation in a country's temperature affects its GDP. Using data from 1950–2003, they find strong and persistent effects of a temporary deviation in temperature, with a point estimate of 1.4% of GDP per degrees Celsius—*but only in poor countries*. A similar result, but using global variation in the temperature, is reported by Bansal and Ochoa (2011). Krusell and Smith (2015), however, find that positive temperature shocks affect the level of GDP but not its rate of growth, and they do not find evidence of a difference between rich and poor countries.

Another approach is taken in Mendelsohn et al. (1994). Instead of attempting to measure a direct relation between climate and output, ie, estimating a production function with climate as an input, the focus is here on agricultural land prices. They label this a *Ricardian approach*.The advantage of this is that adaptation, for example changed crops, can be taken into account. The finding is that higher temperature, except in the fall, is associated with lower land prices. However, the strength in this relation is lower than what is suggested by estimates based on traditional production function analysis. This indicates that the latter underestimates the potential for adaptation.

Burke et al. (2015) estimate empirical relations between economic activity and climate by assuming that local damages are a function not of global temperature but of local temperature. That is, heterogeneity here is built in not in terms of differences in responses to global temperature changes but simply through how local climates are very different to start with. If a region is very cold, warming can be beneficial, and if a region is very warm, further warming will likely be particularly detrimental. In line with Nordhaus (2006), a hump-shaped relation between economic activity and average yearly temperature is then estimated, with a maximum around 12–13°C. If this relation is taken as a causal relation from climate to productivity, it can be used to measure the long-run consequences of climate change. However, the use of the relation to evaluate long-run consequences precludes a study of short- and medium-run costs. This holds in particular for the costs of geographic reallocation of people, an area where little is known. In line with Burke et al. (2015) and Krusell and Smith (2015) postulate a unique damage function of local

[bb] This area is defined by Sweden, Finland, Estonia, Lithuania, Latvia, and Denmark.

temperature for a large number of regions and impose the condition that this function generate Nordhaus's estimated aggregate damages for warming of 1°C, 2.5°C, and 5°C. They find a somewhat lower ideal temperature than do Burke et al. but that the losses from having local temperatures far from the ideal value can be very large.

3.3.4 Remarks

The section on damage measurements in this chapter is short and does not do full justice to the literature. However, even a very ambitious survey would make clear that the research area of damage measurement is at a very early stage and provides frustratingly little guidance for cost-benefit analysis. On the one hand, most of the evidence points to rather limited aggregate global damages, at last for moderate degrees of climate change. On the other hand, it is not possible to rule out large damages, at least if climate change is more than moderate. After all, if the damages from climate change cannot be measured and quantified, how can we arrive at policy recommendations? There is no quick answer; much more research on this is clearly needed. In the absence of more solid evidence there is unfortunately ample room for extreme views—on both sides of the climate debate—to make claims about damage functions that support any desired action. We therefore prefer to proceed cautiously and to base our calibrations of damage functions on the evidence that, after all, has been gathered and put together. But before moving on to a description of the approach we take here, let us make some remarks about some mechanisms we will be abstracting from and that could nevertheless prove to be important.

One aspect of damages concerns the long run: is it possible that a warmer climate hurts (or helps?) long-run economic development, and might it even affect the growth rate of output? The work by Dell et al. (2012) as well as Burke et al. (2015) suggest such effects might be present on the local level, though without providing evidence on mechanisms. For an overall growth-rate effect on world GDP, there is as far as we know no evidence. Clearly, any growth effects—by naturally adding effects over time—will lead to large total effects, and that regions at different ends of the distribution would diverge in their levels of production and welfare, and it is not clear that our growth data support this conclusion. At the same time, the large implied effects make it all the more important to dig deeper and understand whether growth effects could actually be present. To be clear, our null hypothesis is that there are no effects on long-run growth rates of climate change.

Relatedly, it is common—following Nordhaus's lead—to describe damages as essentially proportional to GDP. This formulation, which to an important extent appears to be untested, has some important implications. One is that higher GDP ceteris paribus leads to higher damages. Another is that, since lower GDP means less to consume and consumption (typically, in macroeconomic models) is assumed to be associated with diminishing marginal utility, the welfare losses from a unit of damage measured in consumption units are lower the higher is GDP. Thus, if future generations will have higher GDP than we have today, there are two opposing forces: the total damages in consumption units will be higher but each of those units will hurt future generations less. As we shall see, under reasonable assumptions on utility, those two forces cancel, or roughly cancel. However, there are various ways to depart from Nordhaus approach. One is to assume that damages occur in consumption units but are not (linearly) proportional to GDP (eg, our capital stock could be damaged). Another is to think of damages as occurring to specific consumption bundles that may not display the same degree of diminishing returns as consumption as a whole (examples include effects on leisure, health, or longevity). Damages can also occur in the form of changes in the distribution of resources and in other ways that are not easily thought of in terms of an aggregate damage function proportional to GDP.

Climate change can also lead to social conflict, as it changes the values of different activities and, more generally, "endowments." One channel occurs via migration: if a region is hit hard by a changed climate and people migrate out, history tells us that the probability of conflict in the transition/destination areas will rise (see eg, Miguel et al., 2004, Burke et al., 2009, Jia, 2014, Harari and La Ferrara, 2014, and Burke et al., 2015, for an overview). At the same time, migration is also one of the main ways humans have to adapt to a changing climate. In fact, one view is that "populations can simply move toward the poles a bit" and hence drastically limit any damages from warmer weather; see Desmet and Rossi-Hansberg (2015) for an analysis that takes the migration mechanism seriously (see also Brock et al., 2014). A related aspect is that climate change will have very diverse effects. It may be true that aggregated damages are small as a share of GDP and that those who lose a lot could be compensated by other, losing less or nothing at all. However, such global insurance schemes do not exist, at least not presently. The extent to which there are compensating transfers will likely to greatly impact any reasonable cost-benefit analysis of climate change and policies against it.

Tipping points are often mentioned in the climate-economy area and earlier we discussed some possible tipping points in the natural-science sections. Damages can also have tipping points in various ways and on some level a tipping point is simply a highly nonlinear damage function. One example leading to tipping points is the case of rising sea levels due to the melting of the ice caps. Clearly, some areas may become flooded and uninhabitable if the sea level rises enough, and the outcome is thus highly nonlinear. This argument speaks clearly in favor of using highly nonlinear damage functions on the local level, at least when it comes to some aspects of higher global temperatures. However, the sea-level rise equally clearly does not necessarily amount to a global nonlinearity in damages. Suffice it to say here that very little is known on the topic of global tipping points in damages. We will proceed with the null that a smooth convex aggregate damage function is a good starting point: we follow Nordhaus in this respect as well.

On an even broader level, let us be clear that different approaches are needed in this area. Bottom-up structural approaches like the PESETA project are very explicit and allow extrapolation, but they are limited to a certain number of factors and may miss important other mechanisms. Reduced-form micro-based approaches allow credible identification but may also miss important factors and general-equilibrium effects. Reduced-form aggregate approaches are less likely to miss mechanisms or general-equilibrium effects but necessarily involve a small number of observables and are much harder to interpret and extrapolate from. There is, we believe, no alternative at this point other than proceeding forward on all fronts in this important part of the climate-economy research area.

3.3.5 The Operational Approach: A Direct Relation

We now discuss a very convenient tool for the rest of the analysis in this chapter: a way of incorporating the existing damage estimates into our structural integrated-assessment models. In Section 3.1.1, we have noted that the relation between the CO_2 concentration and the greenhouse effect is concave (it is approximately logarithmic). The existence of feedbacks is likely to imply an amplification of the direct effect, but in the absence of known global threshold effects, the logarithmic relation is likely to survive. Above we have also noted that that modelers so far typically have chosen a convex relation between temperature and damages: at least for moderate degrees of heating, a linear-quadratic formulation is often chosen. Golosov et al. (2014) show that the combination of a concave mapping from CO_2 concentrations to temperature and a convex mapping from temperature to damages for standard parameterizations imply an approximately constant marginal effect of higher CO_2 concentration on damages as a share of GDP. Therefore, they postulate

$$D(T(S)) = 1 - e^{-\gamma(S-\bar{S})}, \tag{16}$$

where S is the amount of carbon in the atmosphere at a point in time and $\bar{S}$ is its preindustrial level. This formulation disregards the dynamic relation between CO_2 concentration and temperature. It also disregards the possibility of abrupt increases in the convexity of the damage mapping and threshold effects in the climate system. These are important considerations, in particular when large increases in temperature are considered. However, the approximation provides a very convenient benchmark by implying that the marginal damage measured as a share of GDP per marginal unit of carbon in the atmosphere is constant and given by γ.[bc] Measuring S in billions of tons of carbon (GtC), Golosov et al. (2014) show that a good approximation to the damages used to derive the damage function in DICE (Nordhaus, 2007) is given by (16) with $\gamma = 5.3 \cdot 10^{-5}$.

In Fig. 10, we show an exponential damage function with this parameter. Specifically, the figure shows the implied damage function plotted against temperature using the relationship $T(S) = 3\frac{\ln S - \ln S_0}{\ln 2}$, ie, using a climate sensitivity of 3 degrees. Comparing this damage function to the Nordhaus function as depicted in Fig. 9, we see that the former is

[bc] Output net of damages is $e^{-\gamma(S-S_0)}Y$. Marginal damages as a share of net-of-damage output then become $[d((1-e^{-\gamma(S-S_0)})Y)/dS]/e^{-\gamma(S-S_0)}Y = \gamma$.

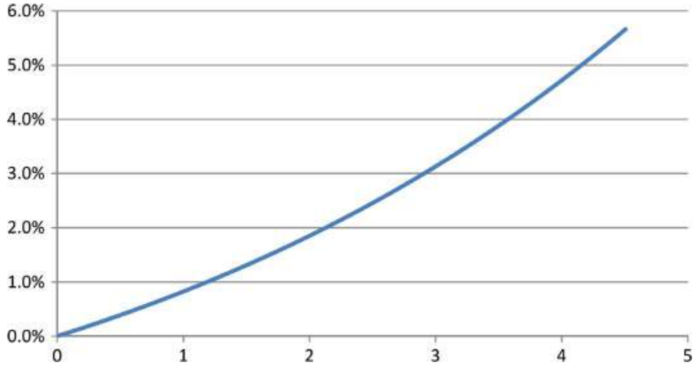


Fig. 10 Damage function using $T(S) = 3\frac{\ln S - \ln S_0}{\ln 2}$ and $D(T(S)) = 1 - e^{-\gamma(S-\bar{S})}$.

slightly less convex.[bd] While the exponential damage function implies a constant marginal loss of 0.0053% per GtC, the quadratic formulation implies increasing marginal loss up to approximately 4°C. However, in the important range 2.5–5.0°C, the marginal loss is fairly constant within the range 0.0053% and 0.0059% per GtC.

4. A STATIC GLOBAL ECONOMY-CLIMATE MODEL

Our discussion of integrated assessment models comes in two parts. The first part—in the present section—introduces an essentially static and highly stylized model, whereas the second part presents a fully dynamic and quantitatively oriented setup. The simple model in the present section can be viewed as a first step and an organizational tool: we can use it to formally discuss a large number of topics that have been studied in the literature. Moreover, for some of these topics we can actually use the model for a quantitative assessment, since it has most of the features of the macroeconomic structure in the later section. The model is thus a static version of Golosov et al. (2014) and it is also very similar to Nordhaus's DICE model.

We consider a world economy where the production of output—a consumption good—is given by

$$c = A(S)k^{\alpha}n^{1-\alpha-\nu}E^{\nu} - \zeta E.$$

Here, $A(S)$ denotes global TFP, which we take to be a function of the amount of carbon in the atmosphere, S. Moreover, we normalize so that S measures the excess carbon concentration, relative to a preindustrial average, $\bar{S}$. That is, the actual concentration is $S + \bar{S}$, whereas we will only need to use S in our modeling. The discussion in Section 3.3 allows us to use this notation and, moreover, to use a simple functional form that we argue is a

[bd] Reducing the exponent on temperature to 1.5 and increasing the constant in front of temperature to 0.0061 in (15) produces a damage function very close to the exponential one.

decent approximation to the complex system mapping the amount of carbon in the atmosphere to temperature and then mapping temperature, with its negative impacts on the economy, to TFP. We will thus use

$$A(S) = e^{-\gamma S},$$

with $\gamma > 0$. Recall from the previous discussions that the map from S to T is logarithmic, so it features decreasing marginal impacts of increased atmospheric carbon concentration on temperature. The estimated mapping from T to TFP, on the other hand, is usually convex, so that the combined mapping actually can be described with the negative exponential function. Thus, damages are $(1 - e^{-\gamma S})k^{\alpha}n^{1-\alpha-\nu}E^{\nu}$, which is increasing and concave in S. (Note that we let energy, E, be capitalized henceforth, to distinguish it from Euler's number, e, used in the exponential damage function.) Though we argue above that this form for the damage function is a good one, it is straightforward to change it in this simple model, as we will below in one of our model applications. The exponential function is also useful because it simplifies the algebra and thus helps us in our illustrations. We will occasionally refer to γ as the *damage elasticity of output*.

The inputs in production include capital and labor, which we take to be exogenously supplied in the static model. The production function is Cobb–Douglas in the three inputs. As for capital and labor entering this way, we just use the standard macroeconomic formulation. The substitution elasticity between the capital-labor composite and energy is also unitary here, which is not far from available estimates of long-run elasticities, and we think of the static model as a short-cut representation of a long-run model. The short-run elasticity is estimated to be far lower, as discussed in Section 2.3.

We also see that the generation of output involves a cost ζE of producing energy. We will discuss in detail below how energy is generated but the simple linear form here is useful because it allows us to illustrate with some main cases. One of these cases is that when energy is only produced from oil. Much of the oil (say, the Saudi oil) is very cheap to produce relative to its market price, so in fact we can think of this case as characterized by $\zeta = 0$. Oil exists in finite supply, so this case comes along with an upper bound on energy: $E \leq \bar{E}$.

A second case is that when energy comes from coal. Coal is very different because its market price is close to its marginal cost, so here we can think of ζ as a positive deep parameter representing a constant marginal cost in terms of output units (and hence the cost of producing energy in terms of capital and labor, and energy itself, has the same characteristics as does the final-output good). Coal is also only available in a finite amount but the available amount here is so large that we can think of it as infinite; in fact, if we were to use up all the coal within, say, the next 500 years, the implied global warming will be so high that most analysts would regard the outcome as disastrous, and hence the presumption in this case is that not all of the amount will be used up (and hence considering the available amount to be infinite is not restrictive). In reality, the supply of fossil fuel is of course not dichotomous: a range of fuels with intermediate extraction costs exists (see the discussion earlier in Section 3.2.3).

A third case is that with "green energy," where a constant marginal cost in terms of output is also a reasonable assumption. Finally, we can imagine a combination of these three assumptions and we will indeed discuss such possibilities below, but it is useful to consider coal and oil first separately first.

Turning to the mapping between energy use and atmospheric carbon concentration, the different energy sources correspond to different cases. In the case of oil and coal, we will simply assume that $S = \phi E + \bar{S}$, where $\bar{S}$ is the part of carbon concentration that is not of anthropogenic origin. As constants in TFP do not influence any outcomes here, we normalize $\bar{S}$ to equal zero. The equation thus states that carbon concentration is increased by the amount of emissions times ϕ. The constant ϕ represents the role of the carbon cycle over the course of a model period—which we will later calibrate to 100 years—and captures the fraction of the emissions during a period that end up in the atmosphere. A explained in Section 3.2, the depreciation structure of carbon in the atmosphere, though nontrivial in nature, can be rather well approximated linearly. Emissions, in turn, are proportional to the amount of fossil fuel used.[be]

We consider a consumer's utility function that, for now, only has consumption as an argument. Hence, so long as it is strictly increasing in consumption the model is complete.

We will discuss outcomes in a market economy of this sort where the consumer owns the capital and supplies labor under price taking, just like in standard macroeconomic models. Firms buy inputs, including energy, in competitive markets and energy is produced competitively. Formally, we can think of there being two sectors where isoquants have the same shape but where in the consumption-goods sector firms solve

$$\max_{k,l,E} e^{-\gamma S} k^{\alpha} n^{1-\alpha-\nu} E^{\nu} - wn - rk - pE,$$

where we denote wages and rental rates by w and r, respectively, and where p is the price of energy; the consumption good is the numéraire. In the energy sector the firms thus solve

$$\max_{k,l,E} p \frac{e^{-\gamma S}}{\zeta} k^{\alpha} n^{1-\alpha-\nu} E^{\nu} - wn - rk - pE.$$

It is straightforward to show, because the Cobb–Douglas share parameters are the same in the two sectors and inputs can be moved across sectors without cost, that this delivers

[be] Constants of proportion are omitted and are inconsequential in this simple model. In a more general framework one must take into account how oil and coal differ in the transformation between the basic carbon content and the resulting emissions as well as how they differ in productive use. We discuss these issues below when we consider coal and oil jointly.

$p=\zeta$ (whenever energy is nontrivially produced, so in the coal and green-energy cases, $1/\zeta$ becomes the TFP in the energy sector relative to that in the final-goods sector). Note also that GDP, y, equals the production of the consumption good, since energy here is an intermediate input.[bf]

Note that in both of the above profit maximization problems firms do not choose S, ie, they do not perceive an effect on TFP in their choice, even though $S=\phi E$ in equilibrium. This is as it should be: the climate damage from emissions are a pure, and global, externality. Markets fail to take this effect into account and optimal policy should be designed to steer markets in the right direction.

The associated planning problem thus reads

$$\max_E e^{-\gamma\phi E}k^\alpha n^{1-\alpha-\nu}E^\nu-\zeta E;$$

here, clearly, the externality is taken into account. In the case of oil, for which $\zeta=0$ is assumed, there is an additional constraint for the planner, namely that $E\leq\bar{E}$.

We will now discuss the solution to this problem for the different cases, starting with the case of oil.

4.1 The Case of Oil

Here, $\zeta=0$ and the energy-producing sector is trivial. Under laissez-faire, all of the oil is supplied to the market and its price will be given by its marginal product: $p\equiv\bar{p}=\nu e^{-\gamma\phi\bar{E}}k^\alpha n^{1-\alpha-\nu}\bar{E}^{\nu-1}$. To the extent $\bar{E}$ and $\gamma\phi$ are large, this will involve an allocation with large damages to welfare.

The planner, on the other hand, may not use up all the oil. It is straightforward to see that the solution to the planner's problem is a corner solution whenever $\bar{E}<\nu/(\gamma\phi)$: the planner would then, like the markets, use up all the available oil. Thus, there is a negative by-product of emissions but it is not, at its maximal use, so bad as to suggest that its use should be limited. (In fact, as we shall argue below, this is not an unreasonable conclusion for oil given a more general, calibrated structure.) If, on the other hand, $\bar{E}\geq\nu/(\gamma\phi)$, the solution is interior at an E that solves $E=\nu/(\gamma\phi)$.

4.1.1 Optimal Taxes

What are the policy implications of this model? For a range of parameter values—for $\bar{E}<\nu/(\gamma\phi)$—no policy is needed. At the same time, taxes are not necessarily harmful: if we think of a unit tax on the use of oil (the firms, whose maximization problems are displayed earlier), so that users of oil pay $p+\tau$ per unit instead of p, all tax rates on

[bf] We do not explicitly have a home sector demanding energy. We take GDP to include housing services and to the extent they can be thought of as produced according to the market production function, these energy needs are included, but other home energy needs (such as gasoline for cars) are simply abstracted from.

oil less than $\bar{p}$ will deliver the optimal outcome (recall that the price of oil is a pure rent and the tax will therefore not affect the allocation). If the unit tax is exactly equal to $\bar{p}$, the market price of oil will be zero and oil producers are indifferent between producing or not. At this level there is still an equilibrium which delivers the optimal amount of oil, namely, when all producers choose to produce; otherwise, not enough oil is used.

So suppose instead that $\bar{E} > \nu/(\gamma\phi)$. Now a tax is needed, and the tax should be set so that $p=0$; the price is zero at the socially optimal use of oil. Otherwise, no oil producer would restrict its production and the outcome would be $\bar{E}$. With a tax that is high enough that the price oil producers receive is zero, ie,

$$\tau = \nu e^{-\nu} k^{\alpha} n^{1-\alpha-\nu} \left(\frac{\nu}{\gamma\phi}\right)^{\nu-1},$$

there exists an equilibrium where precisely oil output is equal to $\nu/(\gamma\phi) < \bar{E}$.

4.1.2 Pigou and the Social Cost of Carbon: A Simple Formula

A different way of getting at optimal policy here is to directly compute the optimal tax of carbon to be that direct damage cost of a unit of emission that is not taken into account by markets. This "marginal externality damage" is referred to in the literature as the *social cost of carbon*.[bg] Moreover, the concept needs to be sharpened as the marginal externality damage can be computed at different allocations. We thus refer to the *optimal social cost of carbon* (OSCC) as the marginal externality damage of a unit of carbon emission evaluated at the optimal allocation. Let the optimal carbon amount be denoted E^*. Given Pigou's principle (Pigou, 1920), the OSCC is the way to think about optimal tax policy, so the tax to be applied is

$$\tau^* = \gamma\phi e^{-\gamma\phi E^*} k^{\alpha} n^{1-\alpha-\nu} (E^*)^{\nu},$$

since this is the derivative of the production function with respect to E where it appears as an externality, evaluated at E^*. The idea here is that this tax always allows the government to achieve the optimal outcome as a competitive equilibrium with taxes. To check that this is consistent with the brute-force analysis earlier, note first that for the case where $E^* = \bar{E}$, $\tau^* = \gamma\phi y^* < \nu y^*/\bar{E}$, where y^* is the optimal level of output. Thus, in equilibrium $p = \nu y^*/\bar{E} - \gamma\phi y^* > 0$, which is consistent with all oil being sold. For the case where $\bar{E} > \nu/(\gamma\phi)$, the optimal tax formula $\tau^* = \gamma\phi y^*$ implies, at the interior solution $E^* = \nu/(\gamma\phi)$, that $p + \tau^* = \nu y^*/E^* = \gamma\phi y^*$ so that $p = 0$. In other words, oil producers are indifferent between producing or not and E^* is therefore an optimal choice.

[bg] The terminology is perhaps a little misleading since one might be led to think that the social cost is the sum of the private and the externality cost, ie, the total cost. Instead "social" just refers to the part not taken into account by the market.

More generally, it is important to understand that Pigou pricing proceeds in two steps: (i) work out the optimal allocation, by solving the planning problem; and (ii) find the OSCC at this allocation and impose that tax. The first step is straightforward in principle but can be challenging if the planning problem is not convex, eg, because the damage function is highly nonlinear; in such a case, there may in particular be multiple solutions to the planner's first-order conditions. The second step has a potential difficulty if for a given tax there are multiple market equilibria. The simple baseline model here does not admit multiple equilibria for a given tax rate but such models are not inconceivable. One important case may be where there are coordination problems in which technology a society chooses—perhaps between a fossil and a green technology. We discuss such cases later.

The OSCC formula that we derived says that the optimal unit tax on carbon is proportional to the value of GDP at the optimal allocation, with a constant of proportionality given by $\gamma\phi$. This result is an adaptation of the finding in Golosov et al. (2014) who derive the OSCC to be proportional to GDP in a much more general setting—a dynamic model that is calibrated to long-run data. The constant of proportionality in that model is also a function of other parameters relating to intertemporal preferences and the carbon cycle, both elements of which are dynamic modeling aspects. They also find this result to be very robust to a number of modeling changes. We shall review these results later but it is important to note already at this point that the core of the proportionality of the OSCC to output can be explained within the structure of the simple static model here.

4.1.3 Costs of Carbon When Taxes are not Optimally Set

Let us emphasize what the OSCC formula says and does not say. It tells us what the marginal externality cost of carbon is, provided we are in an optimal allocation. However, as there appear to be damages from global warming on net and very few countries have carbon taxes, the real world is not at an optimal allocation with respect to carbon use, and this fact suggests that there is another measure that might be relevant: what the marginal externality cost of carbon is today, in the suboptimal allocation. So let SCC, the *social cost of carbon*, be a concept that can be evaluated at any allocation, and suppose we look at the laissez-faire allocation.

One can, conceptually, define a SCC in more than one way. We will define it here as the marginal externality damage of carbon emissions *keeping constant behavior in the given allocation*. This is an important qualification, because if an additional unit of carbon is emitted into the atmosphere, equilibrium decisions will change—whether we are in an optimal allocation or not—and if the given allocation is not optimal, the induced changes in behavior will, in general, have a first-order effect on utility. Hence, an alternative definition would, somehow, take the induced changes in decisions into account. (If the allocation is optimal, these effects can be ignored based on an envelope-theorem argument.)

Let us thus compute the SCC for the case of our static model. Let us assume $\bar{E} > \nu/(\gamma\phi)$, so that there is excessive carbon use. Then the SCC, $\gamma\phi y$, is lower than the OSCC, $\gamma\phi y^*$. This is of course true since $y^* > y$ by definition: the planner's aim is precisely to maximize GDP in this simple model and laissez-faire markets fail to. Note also that the percentage difference between the two measures here is only a function of $\bar{E}$ and E^* and not of other indicators of the "size" of the economy, such as the amount of capital or labor.

Depending on the allocation we are looking at, the SCC may in general be higher or lower than the OSCC. There is also no presumption that the laissez-faire SCC have to be higher than the OSCC, which one might imagine if the marginal damages of emissions rise with the level of emissions. In the simple static model we just looked at here, however, the SCC is always be below the OSCC, because damages appear in TFP and are of a form that implies proportionality to output; the OSCC is chosen to maximize output in this setting, so the OSCC must then be higher than the SCC. In contrast, in our dynamic model in Section 5, although the SCC will be proportional to current output there too, the SCC will typically be above the OSCC. The reason there is that current output tends to be rising with higher current fossil use—it is primarily future output that will fall with current emissions, due to the incurred damages—implying that the SCC will be higher for higher levels of current emissions, and in particular the SCC will be higher than the OSCC since the latter dictates lower emissions. The comparison between the SCC and the OSCC is of practical importance: suppose we are in a laissez-faire allocation today, and that econometricians have measured SCC, ie, damages from emissions based on our current allocation. Then this SCC measure is not of direct relevance for taxation; in fact, for the calibrated dynamic model, we would conclude that the optimal tax is below the econometricians' laissez-faire SCC estimates.

Most of the integrated-assessment literature on the social cost of carbon computes the cost as is indicated above, ie, as a marginal cost at an optimal allocation and, more generally, comparisons between suboptimal and optimal allocations are rather unusual. The simple model here does allow such comparisons (as does the dynamic benchmark model described later). Thus define the *percentage consumption equivalent* as the value λ such that $u(c^*(1-\lambda)) = u(c)$, where c^* is the optimal consumption level and c any suboptimal level. Thus we can compute the laissez-faire value for λ in the simple model (i) to be 0, in the case where there is little enough carbon that all of it should be used ($\bar{E} > \nu/(\gamma\phi)$); and (ii), in the case where too much carbon is available, to satisfy

$$1-\lambda = \frac{e^{-\gamma\phi\bar{E}} k^{\alpha} n^{1-\alpha-\nu} \bar{E}^{\nu}}{e^{-\gamma\phi E^*} k^{\alpha} n^{1-\alpha-\nu} (E^*)^{\nu}}$$
$$= e^{-\gamma\phi(\bar{E}-\frac{\nu}{\gamma\phi})} \left(\frac{\gamma\phi\bar{E}}{\nu}\right)^{\nu}.$$

It is straightforward to verify that λ is increasing in $\bar{E}$ here. Note, however, that variables such as capital or labor do not enter, nor would the size of the population if it were introduced as a separate variable. So the "size" of the economy is not important for this measure.

4.2 The Case of Coal

Here, $\zeta > 0$ and we interpret E as coal. Laissez faire now always involves an interior solution for E and it is such that its (private) benefit equals its (private) cost $p = \zeta = \nu e^{-\gamma\phi E} k^{\alpha} n^{1-\alpha-\nu} E^{\nu-1}$. The planner chooses a lower amount of E: E^* is chosen so that the private benefit of coal minus its social cost equals its private cost:

$$-\gamma\phi e^{-\gamma\phi E^*} k^{\alpha} n^{1-\alpha-\nu} (E^*)^{\nu} + \nu e^{-\gamma\phi E^*} k^{\alpha} n^{1-\alpha-\nu} (E^*)^{\nu-1} = \zeta.$$

Notice here that when coal production becomes more productive (ζ falls), markets use more coal. The same is true for the planner, since the left-hand side of the above equation must be decreasing at an optimum level E^* (so that the second-order condition is satisfied): if ζ falls, the left-hand side must fall, requiring E^* to rise. Thus, technical improvements in coal production imply higher emissions.

4.2.1 Optimal Taxes and the Optimal Social Cost of Carbon

Recall that, in the benchmark model, we think of coal as produced at a constant marginal cost in terms of output goods. Given that GDP, y, equals consumption or $e^{-\gamma\phi E} k^{\alpha} n^{1-\alpha-\nu} (E)^{\nu} - \zeta E$, we can write the equation determining the optimal coal use as

$$-\gamma\phi(y^* + \zeta E^*) + \nu(y^* + \zeta E^*)/E^* = \zeta.$$

Hence, the optimal social cost of carbon, OSCC, is now $\gamma\phi y^*(1 + \zeta E^*/Y^*) = \gamma\phi y^*(1 + \frac{pE^*}{y^*})$. So it is not quite proportional to GDP (as it was in the case of oil) but rather to GDP plus firms' energy costs as a share of GDP. In practice, energy costs are less than 10% of GDP so a rule of thumb that sets the unit tax on coal equal to $\gamma\phi$ times GDP is still approximately correct.

4.2.2 Costs of Carbon When Taxes are not Optimally Set

What is the social cost of carbon at the laissez-faire allocation? It is $\gamma\phi(y + \zeta E)$, where y is laissez-faire GDP and E is laissez-faire carbon use, where we know that $y < y^*$ and $E > E^*$. Unlike in the case of oil, it is not clear whether this amount is smaller than the OSCC. The subtlety here is that the production of coal itself—an intermediate input—is hampered by a damage from climate change and thus the total externality from coal production is not just $\gamma\phi y$.

Consumption in the laissez-faire allocation is lower by a fraction λ that satisfies

$$1-\lambda=\frac{e^{-\gamma\phi E}k^{\alpha}n^{1-\alpha-\nu}E^{\nu}-\zeta E}{e^{-\gamma\phi E^*}k^{\alpha}n^{1-\alpha-\nu}\left(\bar{E}^*\right)^{\nu}-\zeta E^*}=\frac{e^{-\gamma\phi E}k^{\alpha}n^{1-\alpha-\nu}E^{\nu}}{e^{-\gamma\phi E^*}k^{\alpha}n^{1-\alpha-\nu}\left(\bar{E}^*\right)^{\nu}}\frac{1-\nu}{1-\nu+\gamma\phi E^*},$$

where for the second equality we have used the equilibrium and planner's conditions, respectively. This expression is, unlike in the oil example, not explicit in terms of primitives. In general, it depends nontrivially on the size of the economy (of course, one can derive first-order conditions determining both E and E^* as a function of primitives but, for the latter, not in closed form).

4.2.3 Coal Production Only Requires Labor: Our Benchmark Model

The case where coal is produced at a constant marginal cost in terms of output units is somewhat less tractable than the following alternative: coal production does not require capital and does not experience TFP losses from climate change. Ie, $E=\chi n_E$, where n_E is labor used in coal production and χ is a productivity parameter. This case is less realistic but given that energy production is a rather small part of firms' costs, it is convenient to use this specification for some purposes. In this case, we have output given as

$$y=e^{-\gamma\phi\chi n_E}k^{\alpha}\left(1-n_E\right)^{1-\alpha-\nu}\left(\chi n_E\right)^{\nu},$$

where total labor is now normalized: $n=1$. In a laissez-faire allocation, we obtain that $n_E=\frac{\nu}{1-\alpha}$. The planner's allocation delivers optimal n_E^* from

$$-\gamma\phi\chi+\frac{\nu}{n_E^*}=\frac{1-\alpha-\nu}{1-n_E^*}.$$

It is straightforward to check that higher productivity in producing coal will increase emissions both in the laissez-faire allocation and in the optimal one.

Here, moreover, the social cost of carbon will be exactly proportional to GDP, as in the oil case: $\gamma\phi y^*$. The reason is that no indirect externality (through the production of fossil fuel) is involved in this case. Similarly, we can solve for laissez-faire measures of the cost of carbon and the welfare gap relative to the full optimum.

In what follows, when we focus on coal production or oil production that occurs at positive marginal cost, we will use this formulation since it allows for simpler algebra without forsaking quantitatively important realism.

4.3 Calibration

We will now calibrate the static model. This is of course heroic, given that so many aspects of the climate-economy nexus feature dynamics, but the point here is merely to show that the static model can be thought of in quantitative terms. It is also possible to compare the results here to those in the calibration of the fully dynamic model in Section 5.2.

So let the heroics begin by calling our model period 100 years. The benchmark model will have coal as the only source energy; as we will argue later, the stock of oil is rather small relative to the stock of coal, and we leave out renewables for now (in the dynamic model in the later section, we calibrate the production of energy services as using three sources: oil, coal, and green). We assume that coal is produced from labor alone as in the previous section, and the model thus has five parameters: γ, ϕ, α, ν, and χ. We thus need five observations to pin these down.

Output being a flow, we can straightforwardly set α and ν based on average historic data; we select 0.3 and 0.04, respectively (see Hassler et al., 2015). For the rest of the model parameters, let us relate the model's laissez-faire equilibrium to some other observables. We thus need to relate the equilibrium outcomes for the key variables—E, S, n_E, and y—to relevant data targets. A business-as-usual scenario with continuously increasing emissions can lead to increases of the temperature of around 4°C at the end of the century.[bh] We interpret business as usual as our laissez-faire allocation. Let us use this information to find out the associated atmospheric concentration and emissions implied to generate this result, given our model. Arrhenius's formula gives

$$4 = \Delta T = \lambda \frac{\log \frac{S+\bar{S}}{\bar{S}}}{\log 2} = 3 \frac{\log \frac{S+600}{600}}{\log 2},$$

which allows us to solve for S as roughly 900 (GtC, in excess of the preindustrial level 600). What are the corresponding emissions required? The model says $S = \phi E$. To select ϕ, use the estimated linear carbon depreciation formula in Section 3.2.5 above for computing the average depreciation from emitting a constant amount per decade. This amounts to a straight average of the consecutive depreciation rates and a value for ϕ of 0.48: the atmospheric carbon concentration rises by about one half of each emitted unit.

To calibrate γ, let us take IPCC's upper estimate from Fig. 9: at a warming of 4 ∘C, they report a total loss of 5% of GDP. This is a flow measure and thus easy to map into our present structure. We thus need $e^{-\gamma S}$ to equal 0.95. This delivers $\gamma = 5.7 \cdot 10^{-5}$.

It remains to calibrate the parameter χ of the coal sector: its labor productivity. We can find it as follows. To reach 900 GtC, one needs to emit 900/0.48 units given the calculation above. In the model solution, $n_E = \nu/(1-\alpha)$. This means that $900/0.48 = \chi n_E = \chi \cdot 0.04/0.7$, which delivers a χ of approximately 32,813.

4.4 A Few Quantitative Experiments with the Calibrated Model

We now illustrate the workings of the simple baseline model with coal with a few quantitative experiments. The chief purpose is to check robustness of the main results.

[bh] Scenario RCP8.5 from IPCC's 5th Assessment Report.

Similar exercises could be carried out in all of the applications that follow (dealing with uncertainty, tipping points, tax-vs-quota policy comparisons, and so on). We have left such quantitative analysis out for brevity but for each application it would be valuable to use the baseline calibration as discussed here, calibrate the new parameters relevant to the application, and then produce output in the form of tables and graphs. Indeed, such exercises appear ideal for teaching the present material.

Starting out from the calibrated benchmark, let us vary two of the parameters within reasonable ranges. We first look at the effect of the damage elasticity of output, varying it from a half of its estimated value to much higher ones. We see that a doubling of the damage elasticity a little more than doubles the GDP gap between laissez-faire and the optimum. For damages 10 × higher than the baseline estimate, the loss of GDP is almost a quarter of GDP.

Externality cost	$1-\frac{y}{y^*}$
$\gamma/2$	0.0037
γ	0.0177
2γ	0.0454
4γ	0.0983
6γ	0.1482
8γ	0.1954
10γ	0.2400

Turning to carbon depreciation, the robustness looks at a tighter range around the baseline calibration as compared to that for damages (the uncertainty about damages, after all, is much higher). Modest changes in carbon depreciation, as depicted in the table later, do nevertheless have some impact: a change of ϕ by 25 percentage point changes the output gap by about seven tenths of a percent and temperature by a little over half a degree.

1–carbon depreciation	ΔT	$1-\frac{y}{y^*}$
0.75ϕ	3.2624	0.0107
0.95ϕ	3.8340	0.0164
ϕ	3.9658	0.0177
1.05ϕ	4.0938	0.0192
1.15ϕ	4.3388	0.0219
1.25ϕ	4.5707	0.0247

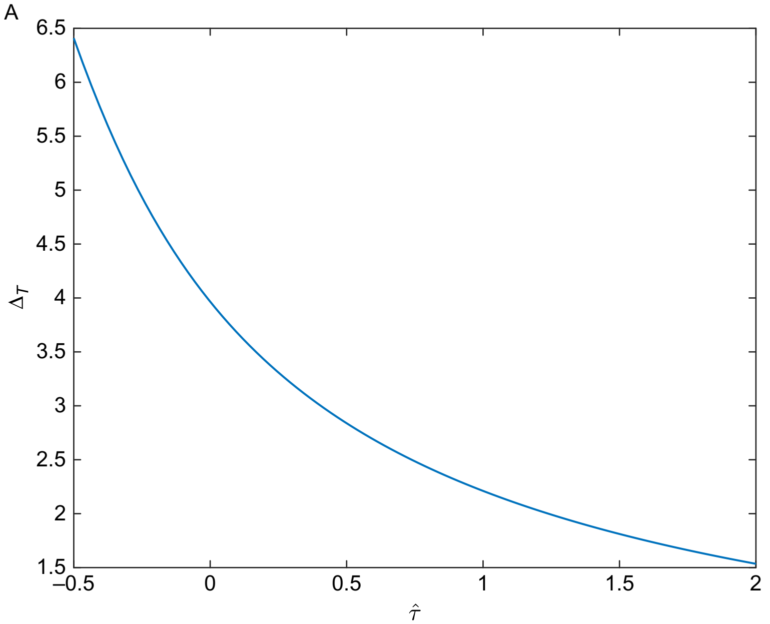


Fig. 11 Outcomes as a function of the tax-GDP rate, $\hat{\tau}$h. (A) Temperature change.

Finally, let us look at a more complete range of suboptimal taxes for the baseline calibration. The table and figures below illustrate by varying the tax, measured as a percent of GDP. Fig. 11 illustrates rather clearly that the model is more nonlinear for negative than for positive taxes: if the tax is turned into a sizeable subsidy the warming and output losses are substantial.

$(\tau/y)/(\tau^*/y^*)$	ΔT	$1-\frac{y}{y^*}$	n_E
−0.5	6.4084	0.0975	0.1294
0	3.9658	0.0177	0.0571
0.5	2.8365	0.0024	0.0353
1	2.2110	0	0.0254
2	1.5346	0.0035	0.0162

4.5 Summary: Core Model

We have built a simple static model which can be used to think about the key long-run aspects of carbon emissions and climate change. Though only a full dynamic, and much more complex, model can do the analysis of climate change full justice, our simple model does have some features that makes it quantitatively reasonable. The mapping from

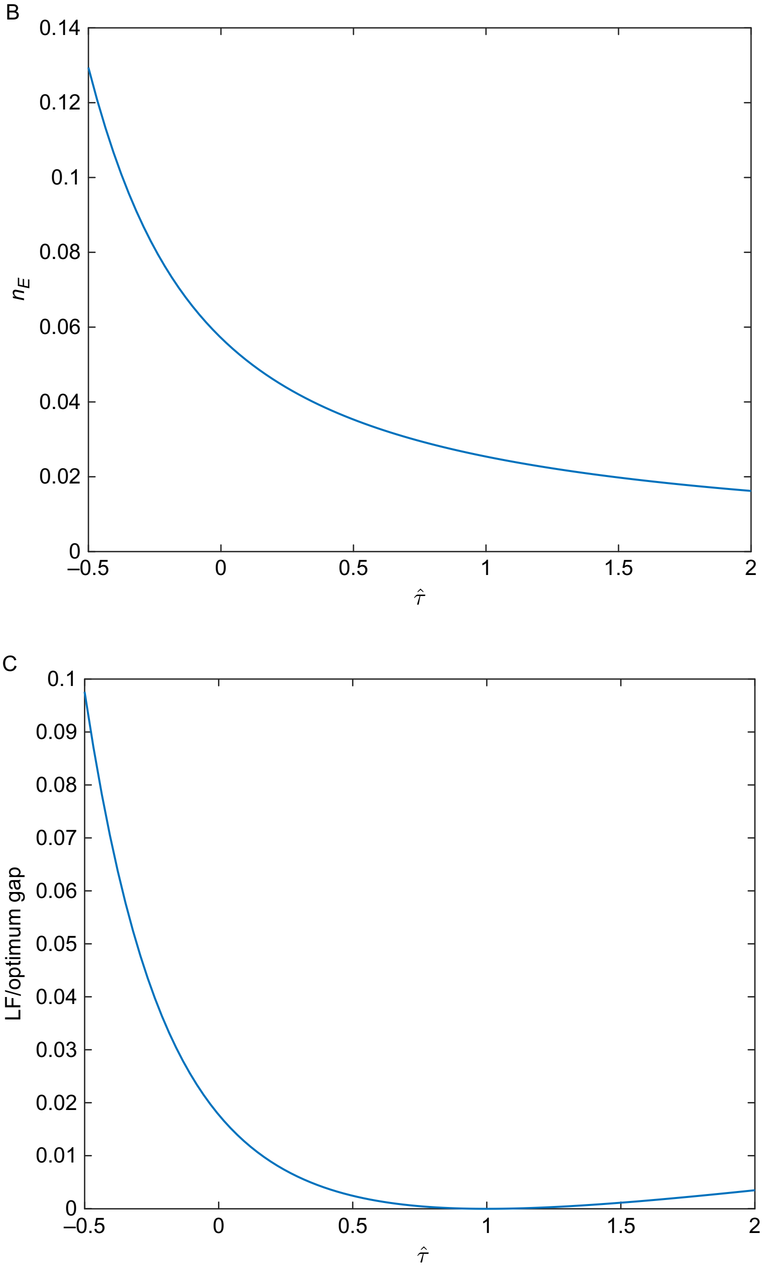


Fig. 11—cont'd (B) Labor in coal sector. (C) Output gap.

emissions to damages is described with a simple closed form but it captures the key features of this mapping in much more elaborate dynamic models, such as Nordhaus's DICE and RICE models. The role of fossil fuels in the economy is also described in a very rudimentary way but it too is the most natural starting point in dynamic quantitative models.

The simple model implies that the optimal social cost of carbon—the marginal externality damage at the optimal allocation—is proportional to GDP; this result is exactly true in some special cases of the model and approximately true otherwise. Also more generally, evaluated as a fraction of output, the (marginal) social cost of carbon (ignoring indirect effects on behavior of raising emissions) is independent of the allocation at which it is measured. This also means that the social cost of carbon is lower in the laissez-faire allocation than in the optimal allocation, because in the static model where damages appear to TFP optimal output by definition is higher than laissez-faire output. This feature will disappear in a dynamic model—where laissez-faire output tends to be higher (in the short run) than in the optimal allocation because less energy is used—and in a model where damages do not affect output, eg, by affecting utility directly. We will of course look at these kinds of extensions below. Moreover, in the simple static model we formulated here, the utility loss from not using taxes to curb carbon use, expressed in percentages of consumption, is scale-independent.

Next, we use the simple model to address some issues that have featured prominently in the literature. These include the choice of policy instruments—in particular the comparison between price and quantity regulations (taxes vs quotas)—along with extensions to consider utility damages, uncertainty, tipping points, technological change, and more.

4.6 Utility Damages

We can, instead of or in addition to the damages to TFP, imagine that higher global temperatures affect welfare directly. This could occur in a variety of ways, through effects on health, the value of leisure, or more generally perceived life quality. Ignoring TFP damages for simplicity, consider first a utility function of a specific functional form:

$$u(c, E) = \log c - \gamma S,$$

where, again, $S = \phi E$ is carbon concentration in excess of the preindustrial level. Here, thus, atmospheric carbon concentration, and hence emissions, influence utility linearly, whereas consumption has decreasing marginal utility. This means that the value of one less unit of emissions in terms of consumption increases as the economy gets richer: $u_E/u_c = \gamma\phi c$. This implies, immediately, that the social cost of carbon in this economy is identical to that above: it is proportional to output. Thus, if the utility cost has the structure just assumed, the implications for how to tax carbon remain the same as in the more common case of TFP damages. In fact, we can now interpret the formulation with TFP damages as possibly coming from two sources: direct damages to TFP and utility damages.

With the remaining parts of the economy unchanged (except that we now view TFP as unaffected by emissions), we can solve for the laissez-faire equilibrium exactly as before. For sake of illustration, let us focus on coal and on the case where energy is produced linearly from labor. The social planner's problem is to solve

$$\max_{n_E} \log\left(k^{\alpha}(n-n_E)^{1-\alpha-\nu}(\chi n_E)^{\nu}\right)-\gamma\phi\chi n_E.$$

The problem simplifies to solving

$$\max_{n_E}(1-\alpha-\nu)\log(n-n_E)+\nu\log n_E-\gamma\phi\chi n_E.$$

The first-order condition gives $\frac{\nu}{n_E}=\frac{1-\alpha-\nu}{n-n_E}+\gamma\phi\chi$, which is the exact same equation as in the corresponding model with TFP damages.

What is the optimal tax/the OSCC in this model? The consumption-good firm's first-order condition for energy (assuming a unit tax τ) is $p+\tau=\nu k^{\alpha}(n-E/\chi)^{1-\alpha-\nu}E^{\nu-1}$, whereas the energy firm's first-order condition reads $p\chi=w$, with $w=(1-\alpha-\nu)k^{\alpha}(n-E/\chi)^{-\alpha-\nu}E^{\nu}$. This delivers $\frac{1-\alpha-\nu}{\chi}k^{\alpha}(n-E/\chi)^{-\alpha-\nu}E^{\nu}+\tau=\nu k^{\alpha}(n-E/\chi)^{1-\alpha-\nu}E^{\nu-1}$, from which we see that $\tau^{*}=\gamma\phi y^{*}$ is the optimal tax here as well.

More generally, the SCC at any consumption/energy allocation here can be obtained as $-u_E(c,E)/u_c(c,E)=\gamma\phi c$, and since consumption is GDP in the static model we again have that the SCC equals $\gamma\phi y$. We can, finally, define the utility loss in the laissez-faire allocation, measured in terms of a percentage consumption loss (ie, from $u(c^{*}(1-\lambda), E^{*})=u(c,E)$). We obtain $\log(1-\lambda)=\log\frac{c}{c^{*}}-\gamma\phi(E-E^{*})$ and thus that $1-\lambda=e^{-\gamma\phi(E-E^{*})}\frac{c}{c^{*}}$ which has the same form as before and, thus, is scale-independent.

4.7 Other Damage Functions

Our assessment in the section earlier on damages from climate change is that this is the subarea in the climate-economy literature with the most striking knowledge gaps. Integrated assessment models differ to some extent in how they formulate damages as a function of climate (temperature) and how they parameterize their functions but the functional form used in Nordhaus's work (the DICE and RICE models) is the most common one. One possibility is that the overall damage *levels* are very different from the most common estimates in the literature, and another is that the functional-form assumptions are wrong. For this discussion, let us use the utility-damage formulation just outlined, and where we argued that $\log c-\gamma S$ is a formulation that is quantitatively close to that used by Nordhaus, given that this function should be viewed as a composition of the mapping from emissions to atmospheric carbon concentration and the mapping from the latter to damages. Let us therefore think about the choice of damage functions in terms of the more general formulation $\log c - \Gamma(S)$, with Γ being a more nontrivial function.[bi] The function Γ, if truly described globally, should probably be increasing for positive values of S (since $S=0$ corresponds to the preindustrial concentration) and convex. For sufficiently low values of S (below 0), the function ought to be decreasing, since there is a reasonable notion of an "appropriate" climate: human beings could not survive if it is too cold either.

A concrete argument for a convex $\Gamma(S)$, rather than the linear one we use in our benchmark, is based on the arguments in Section 3.2.6: there appears to be an approximate reduced-form relationship between the global temperature and the unweighted cumulative amount of past anthropogenic emissions (since the industrial era began), which is *linear*. This was labeled the *CCR* (Carbon-Climate Response) formulation. Then take, say, Nordhaus's global damage function mapping temperature to output losses as given, and combine it with this approximate linear relationship. The resulting $\Gamma(S)$ must then be convex.[bj]

With the more general damage function $\Gamma(S)$, all the earlier analysis goes through with the only difference being that $\Gamma'(S)$ now replaces γ earlier. Obviously, Γ could be calibrated so that $\Gamma'(S)=\gamma$ (with a standard calibration for γ) for current total emission levels, so the added insights here are about how the OSCC (and optimal tax) and the SCCs evolve as GDP evolves.

The SCC in this case becomes $\Gamma'(S)y$, where y again is GDP. Thus, to the extent Γ is convex, the optimal tax (as well as the SCC more generally) would not just be proportional to output but it would also increase with emissions; how much it would increase simply depends on the degree of convexity of Γ. Moreover, imagine an exogenous improvement in TFP. Such a shock would now increase the OSCC (the optimal tax) through two channels. The first channel was present before: a direct positive effect on y (leading to a higher tax by the same percentage amount). The second channel is an indirect effect via a higher demand for E. In terms of the decentralized economy, a higher TFP would, for a given tax, make firms demand a higher E, and since $\Gamma'(S)$ is increasing, this would then call for a further increase in the optimal tax rate.[bk]

[bi] We maintain logarithmic curvature without loss of generality.

[bj] Note, however, that the approximate linearity appears to be in somewhat of a conflict with Arrhenius's insight that the temperature change is proportional to the logarithm of the atmospheric carbon concentration (thus, a concave function). The conflict is not as strong as it seems, however. Our approximation that $\Gamma(S)$ is linear relies on a description of a carbon cycle that is rather realistic (eg, has more complex dynamics) and that uses Arrhenius's formula, which still has widespread acceptance. The upshot of this really is that the just-mentioned convexity after all cannot be very strong.

[bk] This discussion is a reminder that the optimal-tax formula $\tau^* = \Gamma'(S^*)y^*$ is not a closed form, since S^* and y^* are endogenous.

Similarly, the percentage consumption equivalent loss in welfare λ from remaining at laissez-faire can be computed from

$$\log(1-\lambda) = \log\frac{c}{c^*} - (\Gamma(S) - \Gamma(S^*)).$$

To the extent Γ is convex, this expression potentially increases faster in $S - S^*$ (and, more generally, depends on both these emission levels separately).

Now consider a highly nonlinear damage function, and let us investigate whether such a case poses a difficulty for the Pigou approach to the climate problem. Consider the possibility that at a low level of emissions, so for a low S, the social cost of carbon is actually zero: $\Gamma'(S) = 0$. However, $\Gamma(S)$ is at the same time increasing rapidly for higher values of S, after which it again levels off and becomes flat: $\Gamma'(S) = 0$ also for high enough values of S. The latter amounts to a "disaster" outcome where more atmospheric carbon concentration actually does not hurt because all the horrible events that could happen have already happened given that S is so high. Here, though low emissions have a zero SCC, such low emissions are not what Pigou's formula would prescribe: they would prescribe that the SCC equal the net private benefits from emissions, and they are high for low emission levels. The net private benefits of emissions are, in particular, globally declining here (and, since damages appear in preferences and not to production in the particular case under study, always positive). So instead, it is optimal to raise emissions to a point with a S^* such that $\Gamma'(S^*)$ is positive, perhaps one where Γ is increasing rapidly. The example shows that although a rapidly rising damage function in some sense poses a threat, the Pigou approach still works rather well. A key here is that for any given tax rate, the market equilibrium is unique; in the argument earlier, this manifested itself in the statement that the net private benefits from emissions are globally declining. They may not be, ie, there may be multiple market equilibria, but such cases are unusual. We consider such examples in Section 4.14.1 in the context of coordination problems in technology choice.

In conclusion, the model is well-designed also for incorporating "more convex" damage functions, and the qualitative differences in conclusions are not major nor difficult to understand. The key conclusion remains: more research on the determination and nature of damages—including the mechanisms whereby a warmer climate imposes costs on people—is of utmost importance in this literature, and integrated assessment modeling stands ready to incorporate the latest news from any such endeavors.

4.8 Tipping Points

A tipping point typically refers to a phenomenon either in the carbon cycle or in the climate system where there is a very strong nonlinearity. Ie, if the emissions exceed a certain level, a more drastic effect on climate, and hence on damages, is realized. As discussed earlier in the natural-science part of the chapter, one can for example imagine a departure from the Arrhenius approximation of the climate model. Recall that the Arrhenius approximation was that the temperature increase relative to that in the preindustrial era is proportional to the logarithm of the atmospheric carbon concentration (as a fraction of the preindustrial concentration), where the constant of proportionality—often labeled λ—is referred to as climate sensitivity. One way to express a tipping point is that λ shoots up beyond some critical level of carbon concentration. Another is that the carbon cycle has a nonlinearity making ϕ a(n increasing) function of S, due to carbon sinks becoming less able to absorb carbon. Finally, we can imagine that damages feature a stronger convexity beyond a certain temperature point; for example, sufficiently high temperature and humidity make it impossible for humans and animals to survive outdoors.

Notice that all these examples simply amount to a different functional form for damages than that assumed earlier (whether damages appear to TFP or to utility). Thus, one can proceed as in the previous section and simply replace the total damage γS by a damage function $\Gamma(S)$, where this function has a strong nonlinearity. One could imagine many versions of nonlinearity. One involves a kink, whereby we would have a linear function $\gamma_{lo}S$ for $S \leq \underline{S}$ and $\gamma_{hi}S$ for $S > \underline{S}$, with $\gamma_{lo} << \gamma_{hi}$. A second possibility is simply a globally more convex (and smooth) function Γ. One example is Acemoglu et al. (2012), who assume that there is something labeled "environmental quality" that, at zero, leads to minus infinity utility and has infinitely positive marginal utility (without quantitative scientific references). One can also imagine that there is randomness in the carbon cycle or the climate, and this kind of randomness may allow for outcomes that are more extreme than those given by a simple (and deterministic) linear function γS. Finally, the $\Gamma(S)$ function could feature an irreversibility so that it attains a higher value if S ever has been above some threshold, thus even if S later falls below this threshold.

As discussed in the previous subsection, the formulation with a tipping point does not change the analysis of the laissez-faire equilibrium. It does, however, alter the social planner's problem. In particular, in place of γ as representing the negative externality of emissions in the planner's first-order condition we now have $\Gamma'(S)$ and this derivative may be very high. It is still possible to implement the optimum with a carbon tax, though it will no longer just be proportional to the optimal level of GDP and may respond nonlinearly to any parametric change, as discussed earlier. Suppose, for example, that γ becomes "infinite" beyond some $\underline{S}$. Then, from the perspective of a government choosing the optimal tax rate on carbon emissions, the objective function would have highly asymmetric payoffs from the tax choice: if the tax rate is chosen to be too low, the damage would be infinite, and more generally changes in the environment (such as increases in the capital stock or labor input, which would increase the demand for energy) would necessitate appropriate increases in the tax so as to avoid disaster.

Overall, in order to handle tipping points in a quantitative study based on an integrated assessment model one would need to calibrate the nonlinear damage function.

In terms of our first example, how would one estimate $\underline{S}$? As we argued in the natural-science sections 3.1.2 and 3.3.4 earlier, our interpretation of the consensus is that whereas a number of tipping points have been identified, some of which are also quantified, these are tipping points for rather local systems, or systems of limited global impact in the shorter run. To the extent there is a global (and quantitatively important) tipping point, there does not appear to be a consensus on where it would lie in S space. Therefore, at this point and in waiting for further evidence either on aggregate nonlinearities in the carbon cycle or climate system or in how climate maps into economic costs, we maintain a linear formulation (or, in the case damages appear in TFP, in the equivalent exponential form). Performing comparative statics on γ is of course very important and we return to it later.

4.9 Uncertainty

It is possible to analyze uncertainty in a small extension of the simple benchmark model. Suppose we consider a prestage of the economy when the decisions on emissions need to be made—by markets as well as by a fictitious planner. We then think of utility as of the expected-utility kind, and we begin by using a utility formulation common in dynamic macroeconomic models: $u(c) = \log c$. Thus, the objective is $E(\log(c))$. Uncertainty could appear in various forms, but let us simply consider a reduced-form representation of it by letting γ, the damage elasticity of output, be random. That is, in some states of nature emissions are very costly and in some they are not. Recall that the uncertainty can be about the economic damages given any temperature level or about how given emissions influence temperature.

For the sake of illustration, we first consider the simplest of cases: γ is either high, γ_{hi}, or low, γ_{lo}, with probabilities π and $1-\pi$, respectively. The emissions decision has to be made—either by a planner or by actors in decentralized markets—ex-ante, but there is no "prior period" in which there is consumption or any other decisions than just how high to make E. We consider the case of coal here, and with coal production requiring labor only, without associated TFP damages.

Looking at the planning problem first, we have

$$\max_E \pi \log\left(e^{-\gamma_{hi}\phi E}k^\alpha\left(1-\frac{E}{\chi}\right)^{1-\alpha-\nu}E^\nu\right) + (1-\pi)\log\left(e^{-\gamma_{lo}\phi E}k^\alpha\left(1-\frac{E}{\chi}\right)^{1-\alpha-\nu}E^\nu\right).$$

Save for a constant, this problem simplifies to

$$\max_E -(\pi\gamma_{hi} + (1-\pi)\gamma_{lo})\phi E + (1-\alpha-\nu)\log\left(1-\frac{E}{\chi}\right) + \nu \log E.$$

A key feature of this maximization problem is that the damage elasticity appears only in expected value! This means that the solution of the problem will depend on the expected value of γ but not on any higher-order properties of its distribution. This feature, which of course holds regardless of the distributional assumptions of γ, will not hold exactly if coal/oil is produced with constant marginal cost in terms of final output (as in our very first setting above), but approximately the same solution will obtain in any calibrated version of the model since the fossil-fuel costs are small as a fraction of output.

Notice that the "certainty equivalence" result obtains here even though the consumer is risk-averse. However, it obtains for logarithmic utility only. If the utility function curvature is higher than logarithmic, the planner will take into account the variance in outcomes: higher variance will reduce the choice for E.[bl] Formally, and as an example, consider the utility function $c^{1-\sigma}/(1-\sigma)$ so that the planner's objective is

$$\mathbf{E}_{\gamma}\frac{\left(e^{-\gamma E}k^{\alpha}\left(1-\frac{E}{\chi}\right)^{1-\alpha-\nu}E^{\nu}\right)^{1-\sigma}}{1-\sigma}.$$

Since E is predetermined, we can write this as

$$\frac{\left(k^{\alpha}\left(1-\frac{E}{\chi}\right)^{1-\alpha-\nu}E^{\nu}\right)^{1-\sigma}}{1-\sigma}\mathbf{E}_{\gamma}e^{-\gamma E(1-\sigma)}.$$

Assume now that γ is normally distributed with mean $\overline{\mu}$ and variance σ_{μ}^{2}. Then we obtain the objective

$$\frac{\left(e^{-\Gamma(E)}k^{\alpha}\left(1-\frac{E}{\chi}\right)^{1-\alpha-\nu}E^{\nu}\right)^{1-\sigma}}{1-\sigma},$$

with

$$\Gamma(E)=-\overline{\gamma}E+\frac{\sigma_{\mu}^{2}E^{2}(1-\sigma)}{2}.$$

Thus, the objective function is a monotone transformation of consumption, with consumption determined as usual in this model except for the fact that the damage expression γE is now replaced by $\Gamma(E)$, a convex function for $\sigma>1$ (higher curvature than logarithmic). To the extent that the variance σ_{μ}^{2} is large and σ is significantly above 1, we thus have uncertainty play the role of a "more convex damage function," as discussed earlier. We see that the logarithmic function that is our benchmark does apply as a special case.

[bl] The asset pricing literature offers many utility functions that, jointly with random processes for consumption, can deliver large welfare costs; several of these approaches have also been pursued in the climate-economy literature, such as in Barro (2013), Gollier (2013), Crost and Traeger (2014), and Lemoine (2015).

4.9.1 The Dismal Theorem

In this context let us briefly discuss the so-called *Dismal Theorem* derived and discussed by Weitzman in a series of papers (eg, Weitzman, 2009; see also the discussion in Nordhaus, 2009). Weitzman provides conditions under which, in a rather abstract context where governmental action could eliminate climate uncertainty, expected utility is minus infinity in the absence of appropriate government action. Thus, one can (as does Weitzman) see this as an argument for (radical) government action. His result follows, very loosely speaking, if the uncertainty has fat enough tails, the risk aversion is high enough, and the government is able to entirely eliminate the tail uncertainty, but the details of the derivation depend highly on specifics. In our present context, a normal distribution for γ is clearly not fat-tailed enough and the only way for the government to shut down tail risk is to set E to zero. However, imagine that the economy has an amount of free green energy, denoted $\widetilde{E}$, ie, the production function is $e^{-\gamma E}k^{\alpha}\left(1-\dfrac{E}{\chi}\right)^{1-\alpha-\nu}(\widetilde{E}+E)^{\nu}$; then setting $E=0$ still allows positive output. Now imagine that γ has a distribution with fat enough tails, ie, one allowing infinitely high values for γ and slowly decreasing density there. Then expected utility will become infinite if σ is large enough.[bm]

The Dismal Theorem is not connected to data, nor applied in a quantitatively specified integrated assessment model. It relies fundamentally on a shock structure that allows infinitely negative shocks (in percentage terms), and our historical data is too limited to allow us to distinguish the shape of the left tail of this uncertainty in conjunction with the shape of marginal utility near zero; at this point, it seems hard enough to be sure of the mean of the shocks.

4.10 Taxes vs Quotas

In the discussion earlier, we have been focusing on a tax as the obvious candidate policy instrument. Indeed the damage externality is a pure externality for which the Pigou theorem applies straightforwardly. What are alternative policies? The Coase theorem applies too as well but it does not seem possible in practice to define property rights for the atmosphere (into which emissions can then be made, in exchange for a payment to the owner). What about regulating quantities? Indeed the "cap-and-trade" system, which is a quota-based mechanism, has been the main system proposed in the international negotiations to

[bm] A simpler, reduced form setting is that where consumption is given by a t distribution (which has fatter tails than the normal distribution), representing some risk which in this case would be labeled climate risk. Then with power utility, $u(c)=c^{1-\sigma}/(1-\sigma)$, and if σ is high enough, the marginal utility at zero goes to infinity fast enough that expected utility is minus infinity. This point was original made by Geweke (2001). If the government can shut down the variance, or otherwise provide a lower bound for consumption, it would then be highly desirable.

come to a global agreement on climate change. A cap-and-trade system is indeed in place in Europe since 2005.[bn] There is a debate on whether a tax or a quota system is better, and here we will only allude to the main arguments. Our main purpose here, instead, is to make a few basic theoretical points in the comparison between the two systems. These points are also relevant in practice.

Before proceeding to the analysis, let us briefly describe the "-and-trade" part, which we will not subject to theoretical analysis. If a region is subject to a quantity cap—emissions cannot exceed a certain amount—the determination of who gets to emit how much, among the users of fossil energy in the region, must still be decided on. The idea is then to allocate *emission rights* and to allow trade in these rights. The trading, in theory at least, will then ensure that emissions are made efficiently. The initial allocation of emission rights can be made in many ways, eg, through grandfathering (giving rights in proportion to historic use) or auctions. To analyze the trading system formally we would need to introduce heterogeneity among users, which would be straightforward but not yield insights beyond that just mentioned.

The first, and most basic, point in comparing quotas and taxes is that, if there is no uncertainty or if policies can be made contingent on the state of nature, both instruments can be used to attain any given allocation.[bo] If a tax is used, the tax applies to all users; if a quota is used, regardless of how the initial emission rights are used, the market price of an emission right will play the role of the tax: it will impose an extra cost per unit emission and this cost will be the same for all users, provided the market for emission rights works well.

Second, suppose there is uncertainty and the policy cannot be made state-contingent. This is a rather restrictive assumption—there is no clear theoretical reason why policies could not change as the state of nature changes—but still an interesting one since it appears that political/institutional restrictions of this sort are sometimes present. To analyze this case, let us again consider uncertainty and an ex-ante period of decisions. To capture the essence of the restriction we assume that the only decision made ex-ante is the policy decision. A policy could be either a unit tax or a quantity cap. We assume that the quantity cap is set so that it is always binding ex-post, in which case one can view the government as simply choosing the level of emissions ex-ante.

The choice between a tax and a quota when there is uncertainty (or private information on the part of "the industry") has been studied extensively in the environmental literature since Weitzman (1974) and similar analyses are available in other parts of economics (eg, Poole, 1970). One can clearly provide conditions under which one

[bn] The European Union Emission Trading System (EU ETS) was launched in 2005 covering about half the CO_2 emissions in the union (Ellerman and Buchner, 2007).

[bo] This statement requires a qualification for taxes in the (rather unusual) cases for which a Pigou rule is not sufficient, as discussed already.

policy or the other is better, along the lines of Weitzman's original paper. Weitzman considered a cost and a benefit of a pollutant, each of which depended on some random variable, and the two random variables were assumed to be independent. He then showed that what instrument would be best depended on the relative slopes of the marginal benefits and cost curves. Follow-up papers relaxed and changed assumptions in a variety of directions, but there appear to be no general theorems that apply in the climate-change application to conclude decisively in one way or the other. In fact, we know of no quantitatively parameterized dynamic model that looks at the issue so what we will do here is simply provide a straightforward example using our simple static model and then discuss a couple of separate, and we believe important, special cases.

For our example, we use one type of uncertainty only: that of the cost of producing fossil fuel, χ. With the calibrated model and a uniform distribution around the calibrated value for χ we obtained the ex-ante utility levels for a range of taxes and for a range of emissions, both committed to before the randomness is realized. Fig. 12 shows the results: a range of tax values around the optimal tax outperform the optimal quota. In this case, the precommitted tax rate is a fixed value. If it could be set as a proportion of output, which is ruled out now by assumption since the tax cannot be state-contingent

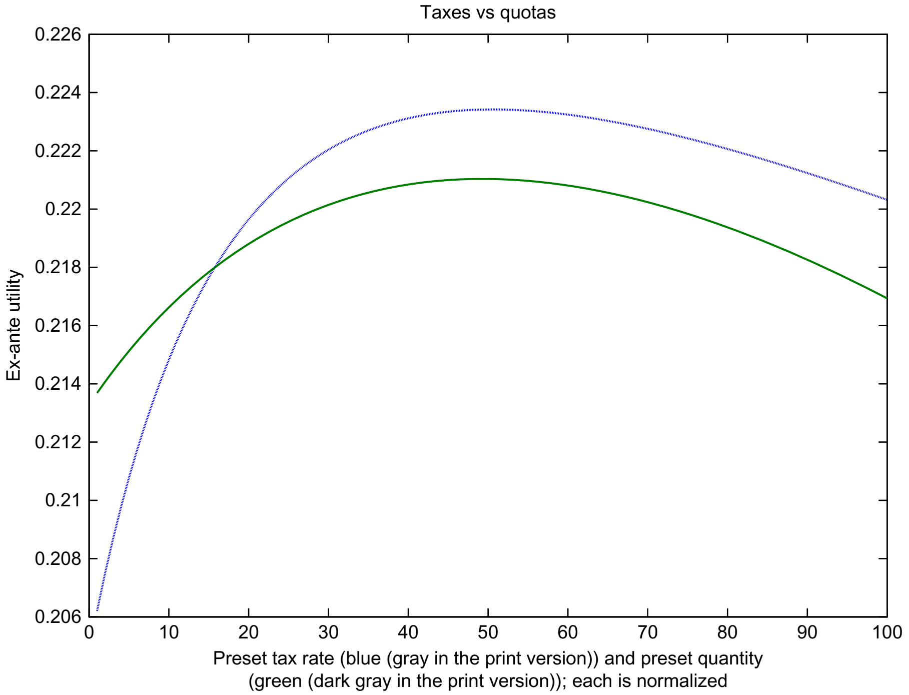


Fig. 12 Utility from precommitting to a unit tax (blue (gray in the print version), with the tax on the *x*-axis) or a quantity cap (green (dark gray in the print version), with the quantity cap on the *x*-axis).

but output will be, it would be fully optimal also ex-post, since the best tax ex-post is always a fraction $\gamma\phi$ of output. Apparently, the ex-post randomness of output is not significant enough to overturn this result. It is straightforward to look at other types of shocks. Shocks to γ deliver more similar welfare outcomes for (optimal and precommitted) taxes and quotas.

Now suppose that we consider a case of a tipping point and that the uncertainty is coming from energy demand (through, say, a separate, exogenous and random TFP factor) or from the cost of coal production (through χ). If the tipping point is known to be $\underline{E}$, and $\Gamma(E)$ is equal to zero for $E < \underline{E}$ but positive and very high otherwise, what is then the best policy from an ex-ante perspective? Clearly, a policy with an emissions cap would simply be set at $\underline{E}$, a cap that may or may not bind ex-post: if the demand for energy is low, or the cost of producing it is high, the ex-post market solution will (efficiently) be to stay below $\underline{E}$, and otherwise the cap will (efficiently) bind. A tax will not work equally well. One can set the tax so that the economy stays below the tipping point, but in case the energy demand is low, or its production costs are high, ex-post, output will be inefficiently depressed. Thus, when we are dealing with *asymmetric* payoffs of this sort (relative to the amount of emissions), a quantity cap is better.

The previous example would have emissions rights trading at a positive price sometimes and at a zero price otherwise. Thus, the system with a quantity cap leads to a random cost for firms of emitting carbon dioxide (beyond the price the firms pay the energy producers). Variations in the supply of emissions rights, decided on by regulatory action, influence the price of the trading rights as well. The experience in Europe since the cap-and-trade system illustrates these points well: carbon prices have fluctuated between over 30 euro and virtually zero since the system started. Such fluctuations are observed also in other regions with cap-and-trade systems (eg, New Zealand). Clearly, since optimal carbon pricing should reflect the social cost of carbon, such fluctuations are only efficient if the social cost of carbon experiences fluctuations. Damages from carbon emissions are likely not experiencing large fluctuations, but our assessments of how large they are of course change over time as scientific knowledge accumulates. The recent large drops in the price of emission rights can therefore be viewed as problematic from a policy perspective.

A cap-and-trade system could be augmented with a "central emission bank" that would have as its role to stabilize the price of emission rights by trading actively in this market, hence avoiding the large and inefficient swings observed in the EU system. Notice, however, that we would then be very close in spirit to a tax system: a tax system would just be a completely stable (provided the chosen tax is stable) way of implementing a stable price of emissions for firms.[bp]

[bp] This and other issues in this policy discussion are covered in Hassler et al. (2016).

4.11 Carbon Taxation in the Presence of Other Distortionary Taxes

Suppose the government needs to raise revenue and needs to do this in a distortionary manner; the most common example would involve labor taxation and it is also a form of taxation that can be studied in the baseline model here by the addition of valued leisure. How, then, will the optimal carbon tax change? For example, suppose preferences are $\log c+\psi\log l$, where l is leisure, so that the labor input in the final-goods sector would be $1-n_E-l$ (and, as before, n_E in the coal sector). Suppose also that the government has a distortionary tax on labor income, τ_l. Taxes are used to pay for an exogenous amount G of consumption good (that does not enter agents' utility). Lump-sum taxation is ruled out (but lump-sum transfers are not), and thus the setup mimics a typical second-best situation in public finance.[bq]

Consider first a planning solution where the government is unrestricted and can just mandate quantities. Thus, it maximizes

$$\log\left(e^{-\gamma\phi\chi n_E}(1-n_E-l)^{1-\alpha-\nu}(\chi n_E)^{\nu}-G\right)+\psi\log l$$

by choice of n_E and l. This delivers two first-order conditions. One is familiar from our baseline model:

$$-\gamma\phi\chi_E-\frac{1-\alpha-\nu}{1-n_E-l}+\frac{\nu}{n_E}=0.$$

The other is the standard macro-labor condition

$$-\frac{1}{c}\cdot\frac{(1-\alpha-\nu)y}{1-n_e-l}+\frac{\psi}{l}=0,$$

which says that the marginal utility of consumption times the marginal product of labor has to equal the marginal utility of leisure (in the expression, of course, y denotes $e^{-\gamma\phi\chi n_E}(1-n_E-l)^{1-\alpha-\nu}(\chi n_E)^{\nu}$ and $c=y-G$). These two first-order conditions can be solved for first-best levels of n_E and l given any G.

Now consider in contrast a competitive equilibrium which is laissez-faire with regard to the taxation of carbon and which only uses labor taxes to raise revenue. Then, the two conditions above would be replaced, first, by the laissez-faire condition for coal

$$-\frac{1-\alpha-\nu}{1-n_E-l}+\frac{\nu}{n_E}=0$$

[bq] One can also consider an alternative assumption: there is no need to raise revenue ($G=0$), there is an exogenous tax rate on labor income, $\tau>0$, and any tax revenues are rebated back lump-sum.

and, second, a distorted macro-labor condition

$$-\frac{1}{c}\cdot\frac{(1-\alpha-\nu)y(1-\tau_l)}{1-n_E-l}+\frac{\psi}{l}=0,$$

with the additional constraint that the government budget balances: $\tau_l(1-\alpha-\nu)y/(1-n_E-l)=G$. These three conditions now determine n_E, l, and τ_l and do not deliver the first best. In particular, one can think of two "wedges" defining different departures from the first best: the externality wedge due to climate damages and the tax wedge on labor supply (these are defined as the differences between the left-hand sides of the above equations with taxes and the corresponding ones from the first-best first-order conditions).

Now suppose we increase the carbon tax marginally from 0. Then (i) the climate wedge would become smaller and (ii) because τ_l falls—the government budget now reads $\tau_l(1-\alpha-\nu)y/(1-n_E-l)+\tau\chi n_E=G$ so that $\tau>0$ allows a lower τ_l—the labor wedge would fall as well. Hence relative to a laissez-faire situation from the perspective of coal, introducing coal taxation involves a *double dividend*: it diminishes the climate externality and it reduces the labor distortion. This is an often-discussed point in the climate literature; for example Jorgenson et al. (2013b,a) argue that the double dividends are quantitatively important for the United States and China, respectively.[br] Of course, the extent to which labor taxes can be reduced depends on the size of the coal tax base.

What, then, will the best level of carbon taxation be? Will carbon taxes be higher than in the absence of distortionary labor taxation? It would be straightforward to derive an answer in the present model by maximizing consumer welfare—with the same objective as that used by the planner—subject to the macro-labor first-order condition above, $\tau\chi/y-\frac{1-\alpha-\nu}{1-n_E-l}+\frac{\nu}{n_E}=0$ for the market's marginal condition for coal, and the government's budget constraint. One can derive a marginal condition for the planner's choice of τ which involves the setting of a weighted combination of wedges to zero; this condition can be solved numerically, together with the other equations, for the endogenous variables. The final level of taxes in this second-best solution is hard to characterize in terms of primitives but some intuition can perhaps be gleaned. If the use of coal is complementary with labor (which it is in the Cobb–Douglas formulation of production), on the margin the reduction of coal will hurt labor supply because it lowers the marginal product of labor. This speaks for a second best with a coal tax that is lower than in the absence of distortionary labor taxation. If coal were instead complementary with leisure (say because people burn coal to heat their homes when not working), this effect would go in the opposite direction on the margin. However, exactly how all these effects play

[br] One can also identify a third dividend from introducing coal taxation: the reduce in local pollution from the burning of coal, a factor which appears of first-order relevance particularly in China.

out depends on the details of preferences and technology. For recent work on these issues that in addition also addresses distortions due to capital taxation, see Schmitt (2014), who pursues this approach in a dynamic model closely related to the setup here, and Barrage (2015), who looks at a closely related setting and uses a primal approach to taxation.[bs]

4.12 A More Detailed Energy Sector

We set out with a stylized description of energy production using either oil, coal, or some green alternative. In practice it is not either or; rather, these sources can all be used and are partially, but not fully, substitutable. Some integrated assessment models include very complex energy systems (eg, WITCH or MERGE; the latter is described in Manne et al., 1995). One way to incorporate multiple energy sources explicitly is to keep one kind of energy as an input into production but let this energy itself be produced from an array of sources, including fossil fuel. Thus, consider the CES technology

$$E = \left(\kappa_o E_o^{\rho} + \kappa_c E_c^{\rho} + (1 - \kappa_o - \kappa_c) E_g^{\rho} \right)^{\frac{1}{\rho}},$$

where E_i is the energy produced from source i, with $i = o$ representing oil (and natural gas), $i = c$ representing coal, and $i = g$ representing energy generated without fossil fuel.[bt] This description is still stylized but it allows us to look into some interesting issues. The parameter $\rho \in (-\infty, 1]$ regulates the (constant) elasticity of substitution between the different energy sources.[bu] The κ_is are share parameters regarded as exogenous in all of our analysis. We continue to think about the production of oil, coal, and green energy as in the previous discussion.

It is straightforward to check that the social cost of carbon is still γy with this formulation. Thus, this extension is not interesting from the perspective of optimal policy. Its value, instead, is to deliver a much richer view of what the cost is of remaining at laissez-faire, or in any case far from the optimum, because this cost turns out to crucially depend on the elasticity of substitution between the different kinds of energy.

First, and just for illustration, let us look at the case where there is just oil and coal, ie, where there is no green energy. Clearly, then, if the degree of substitutability between oil and coal is very low, the difference between laissez-faire and the optimum is small. Consider the extreme case: a Leontief function, ie, $\rho = -\infty$. Then if the total stock of oil is small enough that the optimum involves using it all, the laissez-faire and optimal

[bs] As is typically the case, in dynamic analyses it makes a difference whether the government has commitment or not; Schmitt considers cases without commitment.

[bt] It would be natural to consider a slight extension of this formulation with a nested CES between a composite of oil and coal, on the one hand, and green energy on the other. Thus, oil and coal would form a separate CES aggregate and one could consider the quantitatively reasonable case with a high degree of substitutability between oil and coal and a lower one between the oil–coal composite and green energy.

[bu] The elasticity is $1/(1-\rho)$.

allocations are identical. With some more substitutability, the laissez-faire allocation is not optimal, because coal use should be reduced given the externality and its unlimited supply (recall its constant marginal cost in terms of labor). However, the difference is still limited. In practice, however, oil and coal are rather good substitutes, so let us instead (again, for illustration only) consider the opposite extreme case: perfect substitutability ($\rho = 1$). Then the level of coal is determined very differently: laissez faire is far from the optimum (provided γ is large). Thus, in this case there will be significant total losses from government inaction.

According to available estimates, the remaining amount of (low-cost) oil left is quite limited, in particular in comparison with the amount of remaining coal, so oil is not of key importance for climate change.[bv] What is of importance, however, is the substitutability with green energy. So, second, let us consider fossil fuel (interpreted as coal) vs green energy. In a metastudy, Stern (2012) reports a long-run elasticity of substitution of 0.95, as an average of oil–coal, oil–electricity, and coal–electricity elasticity measures. Thus, this unweighted average is close to a Cobb–Douglas specification. In this case, there can be a rather significant difference between the optimum and laissez-faire; relatedly, price incentives, or the effects of imposing a tax, are large if there is a nontaxed good that is a close substitute.[bw] However, it is conceivable that green technology in the future will be a very good substitute with fossil fuel. Considering a higher elasticity than the unitary Cobb–Douglas elasticity is therefore a relevant robustness check. In this case, the difference between the optimum and laissez-faire is rather large. For example, Golosov et al. report, using a calibrated dynamic counterpart of the model here, that an elasticity of 2 leads laissez-faire coal use 100 years from now to rise to levels that imply exhaustion of all the coal deposits and would likely have catastrophic consequences for the climate. In contrast, in the optimum, coal use in 100 years is *lower* than it is today, and the climate as a result is rather manageable.

By definition, in the case of green energy vs fossil fuel, the observation that a high elasticity of substitution leads to large welfare losses from not imposing a carbon tax (or a quota) at the same time means that there is a large potential social benefit from climate change action. A closely related implication is that there are, in such a case, strong incentives—high social payoffs—from doing research to come up with green alternatives. We turn to this issue in Section 4.14.

4.13 The Substitutability Between Energy and Other Inputs

What aspects of the earlier analysis are influenced by the nature of the production function? We have assumed a Cobb–Douglas structure in part for simplicity and part because the energy share, though having gone through large swings over shorter periods of time,

[bv] See McGlade and Ekins (2015) for supply curves of different types of fossil fuel.

[bw] The Cobb–Douglas case is very similar to the case with only coal considered above.

has remained fairly stable over the longer horizon (recall Fig. 1 in Section 2). It is nevertheless necessary to also discuss departures from unitary elasticity. In this discussion, we will maintain the assumption of a unitary elasticity between the capital and labor inputs, thus confining attention to a different elasticity between the capital-labor composite, on the one hand, and energy on the other.

Consider the aggregate production function $e^{-\gamma S}F(Ak^{\alpha}n^{1-\alpha}, A_E E)$, where F is CES and A and A_E are technology parameters, thus maintaining the assumption that damages appear as decreases in TFP. The social cost of carbon with this formulation will then obey the same structure as before, ie, the marginal externality damage of fossil fuel (through increased emissions E) is $\gamma\phi y$. What is different, however, is the difference between the laissez-faire allocation and the optimum or, expressed differently, the consumption equivalent cost of a suboptimal allocation. Consider oil, ie, a fossil fuel with zero extraction costs in a finite supply $\bar{E}$. Assume that it is not optimal to use all of the oil, and let us simply examine the two extreme cases: Leontief and perfect substitutability.

We begin with the Leontief case. Here, output is given by $e^{-\gamma\phi E}\min\left\{Ak^{\alpha}n^{1-\alpha}, A_E E\right\}$. Ie, there is no substitutability between the capital-labor composite and oil. In laissez-faire, oil use is $\bar{E}$. It is easy to show from the planner's first-order condition that $E^{*}=1/(\gamma\phi)$ in this case.[bx] Recall from Section 4.1.3 that, under Cobb–Douglas, the optimal allocation is $E^{*}=\nu/(\gamma\phi)$ and that the ratio of optimal to laissez-faire output is $e^{\gamma\phi(\bar{E}-\nu/(\gamma\phi))}\left(\frac{\nu}{\gamma\phi\bar{E}}\right)^{\nu}>1$. Now we obtain $e^{\gamma\phi(\bar{E}-1/(\gamma\phi))}\frac{1}{\gamma\phi\bar{E}}$. Because $-\nu+\nu\log\nu$ is decreasing we therefore conclude that in the Leontief case, the difference between the optimal and the laissez-faire allocation is smaller than under unitary elasticity. The fall in energy use is smaller, and this effect dominates the stronger impact on output of any given fall in energy.

Under perfect substitutability, we have output given by $e^{-\gamma S}\left(Ak^{\alpha}n^{1-\alpha}+A_E E\right)$ and we assume that capital and labor are in use. Now the planner's first-order condition leads to $E^{*}=1/(\gamma\phi)-Ak^{\alpha}n^{1-\alpha}/A_E$, which (as for the unitary-elasticity case) is a smaller amount than in the Leontief case. It is also possible to show that the wedge between optimal and laissez-faire output in this case is smaller than in the Leontief case.

In sum, we see that the energy use can be different than in the case with unitary elasticity between energy and other inputs. With production functions with very low substitution elasticity between energy and other inputs, energy use will dictate that energy use in the optimum fall more, but there is also a corresponding gain in a higher TFP. There does not, perhaps surprisingly, therefore appear to be a very strong effect on

[bx] This holds so long as there is an interior solution, ie, if $1/(\gamma\phi)<Ak^{\alpha}n^{1-\alpha}/A^{E}$. Note that there is abundance of capital and labor now: on the one hand, the market uses oil to the point where $E=Ak^{\alpha}n^{1-\alpha}$, so that there is excessive oil. On the other hand, the planner may want to decrease the oil use if the just stated inequality holds, so that from the planner's perspective, there is an abundance of capital and labor instead.

the net gap between optimal output and laissez-faire output as the elasticity of substitution between inputs is varied. This is comforting given that the Cobb–Douglas formulation is much easier to handle analytically.

4.14 Green Technology and Directed Technical Change

The existence of the green technology was taken as given earlier; green technologies of various sorts—versions of water and wind power—have of course existed since before the industrial revolution. These technologies have been improved and there are also new sources of electricity production that do not involve fossil fuels, such as nuclear power and solar power.[by] A central issue of concern in the area of climate change is the further development of these technologies and research toward new ones. In the macroeconomically oriented literature on climate change, various models have been developed, with early papers by Bovenberg and Smulders and others (see, eg, Bovenberg and Smulders, 1995). More recently, Acemoglu et al. (2012) provided a setting of directed technical change and made the point that there may be *path dependence* in R&D efforts toward the development of different energy technologies. We will now use the simple model to illustrate these facts and some other points that have been made in the literature.

A static model cannot fully do justice to the much more elaborate dynamic settings where many of the arguments in this part of the literature have been developed. It does, however, allow us to make a number of basic points. One simplification in our analysis here is that we will not explicitly describe a decentralized R&D sector.[bz] We will distinguish between two different kinds of technological developments: new techniques for the efficient use of energy ("energy saving") and new techniques for the production of energy. We begin with the latter.

4.14.1 Energy Production

We will mostly abstract from the determination of the overall efforts toward technological developments, which one could model as well (say as a tradeoff between these activities and using labor directly in production), and simply assume that there is an R&D input available in fixed supply; we set the total amount to 1 without loss of generality. The use of this input can be *directed* toward either improving the productivity in producing energy from fossil sources, m_c, or from green sources, m_g, with the constraint that $m_c + m_g = 1$. Eg, we can think of this choice as one between improving the drilling/extraction technologies for North Sea oil and technological improvements in the

[by] Nuclear power is problematic from an environmental perspective too but we do not discuss this issue here.

[bz] We could have developed such a version even in our static model but it would have complicated notation without adding much of significance.

cost-efficiency of solar-based units. The most straightforward setting would maintain the production function in a two-energy-input form:

$$e^{-\gamma E_c} k^{\alpha} n^{1-\alpha-\nu} \left(\lambda_c E_c^{\rho} + (1-\lambda_c) E_g^{\rho} \right)^{\frac{\nu}{\rho}},$$

with the production of energy given by

$$E_c = \chi_c n_c \quad \text{and} \quad E_g = \chi_g n_g$$

with $n + n_c + n_g = 1$. Along the lines indicated earlier, for given values of χ_c and χ_g, this model is straightforwardly solved either for the optimum or for a laissez-faire allocation.

A very simple way of modeling research into making energy production more efficient can now be expressed as follows:

$$\chi_c = \overline{\chi} m_c \quad \text{and} \quad \chi_g = \overline{\chi} m_g,$$

with $m_c + m_g = 1$. (If $\lambda_c = 1/2$, this setting is now entirely symmetric.)

A decentralized version of this model would have no agent—either the producer or the user of fossil fuel—take into account the negative externality. However, notice that there are increasing returns to scale in producing energy: double n_c, n_g, m_c, and m_g, and E_c and E_g more than double. A decentralized equilibrium here would then have a much more elaborate structure of varieties within each energy type, either with variety expansion à la Romer or fixed variety but creative destruction Aghion and Howitt (1992), monopolistic competition with profits, and then perfectly competitive research firms producing new varieties (in the Romer case) or product improvements (in the Aghion–Howitt case). We will not spell the variety structure out, but we will make the assumption that the aggregation across varieties is identical for fossil fuel and green energy, eg, implying identical markups across these two energy sectors. Finally, there would normally (in dynamic models) also be spillovers, mostly for tractability, but they are not needed here.[ca] We will, however, discuss spillovers later because there are substantive issues surrounding them.

A decentralized model such as that just described delivers equilibrium existence despite the technological nonconvexity but we omit the description of it for brevity; see Romer (1990) for the basic variety-expansion structure and Acemoglu (2009) for a more recent description of a range of endogenous-growth models and many of their uses. Monopolistic competition would distort the allocation, in the direction of underprovision of energy, which itself would be beneficial for counterbalancing the climate externality and thus to some extent relieve the government of the pressure to tax fossil fuel. In the laissez-faire equilibrium, in the case of symmetry between fossil fuel and

[ca] The reason they improve tractability is that if the researchers' output does not give the researcher herself dynamic gains, the R&D decision becomes static.

energy, the markets will produce whatever the total energy composite is in an efficient manner.[cb] Denoting this level E, the laissez-faire allocation will minimize $n_c + n_g$ subject to

$$E_c^\rho + E_g^\rho \geq E^\rho, \quad E_c = n_c \overline{\chi} m_c, \quad E_g = n_g \overline{\chi} m_g, \quad \text{and} \quad m_c + m_g = 1.$$

The solution to this problem depends critically on ρ. So long as $\rho < 1/2$, ie, so long as the two sources of energy are poor enough substitutes, the solution is to set $n_g = n_e$ and $m_c = m_g = 1/2$; it is straightforward to compute the implied total labor use. If, on the other hand, $\rho > 1/2$, then the outcome is to set either $n_c = m_c = 0$ or $n_g = m_g = 0$, ie, a corner solution obtains, with another easily computed labor use. So if the energy inputs are substitutable enough, there are multiple equilibria. The multiplicity is knife-edge in this case since we assumed full symmetry. However, the essential insight here is not multiplicity but rather sensitivity to parameters, as we will now elaborate on.

Suppose now, instead, that we change the setting slightly and assume

$$\chi_c = \overline{\chi}_c m_c \quad \text{and} \quad \chi_g = \overline{\chi}_g m_g,$$

ie, we assume that there are two separate constants in the two research production functions. Then, in the case where ρ is high enough, there will be full specialization but the direction of the specialization will be given by the relative sizes of $\overline{\chi}_c$ and $\overline{\chi}_g$. If the former is higher, the energy will be produced by fossil fuel only; if the latter is higher, the energy will be produced by green energy only. If the economy experienced a small change in these parameters switching their order, we would have a complete switch in the nature of the energy supplies. Crucially, now, note that we can think of $\overline{\chi}_c$ and $\overline{\chi}_g$ as given by historical R&D activities. Then we can identify the kind of *path dependence* emphasized in Acemoglu et al. (2012). These authors argued that temporary efforts, via subsidies/taxes, to promote the research on "clean goods"—those produced using green energy—would have permanent effects on our energy supplies by managing to shift our dependence on fossil fuel over to a dependence on green energy.[cc] This can be thought of, in terms of this model, as having managed to make $\overline{\chi}_g > \overline{\chi}_c$ by past subsidies to green R&D. Acemoglu et al. used a dynamic model with details that differ from those here—among other things, they assumed much stronger convexities in damages so that a switch to green energy was necessary or else utility would be minus infinity—but this is the gist of their argument.

One can question whether the substitutability is strong enough for the path-dependence argument to apply. For example, Hart (2013) argues that there are strong

[cb] The assumption of symmetry across the two energy sectors, and hence identical markups, is an important assumption behind this result.

[cc] In their analysis, the authors use a notion of two kinds of goods, one clean and one dirty, with labels deriving from the energy source used to produce them. The setting we use here, with an energy composite relevant for the whole economy, is of course also an abstraction but we prefer it because it lends itself more easily to calibration and comparison with data.

complementarities in research across dirty and clean technologies. These complementarities could, in practice, take the form of external effects/spillovers. For example, research into improving electric cars can be helpful for improving the efficiency of cars running on gasoline or diesel, and whether these complementarities are fully paid for or not in the marketplace is not obvious. A way of expressing this formally within our simple framework is a further generalization of our framework as follows:

$$\chi_c = \overline{\chi}_c m_c^{\zeta} m_g^{1-\zeta} \quad \text{and} \quad \chi_g = \overline{\chi}_g m_g^{\zeta} m_c^{1-\zeta}.$$

To the extent ζ is not too much higher than 1/2 here, there are strong complementarities in technology development and path dependence would not apply. Hart (2013) argues this is the relevant case, but it would be hard to argue that the case is settled. Aghion et al. (2014), furthermore, show that there is empirical support for persistence, though whether these effects are strong enough to generate the kind of path dependence emphasized in Acemoglu et al. (2012) is still not clear.

Turning, finally, to the planning problem in these economies, it is clear that the planner faces a tradeoff between the forces discussed here and the climate externality generated by fossil fuel. The setting is rather tractable and it is straightforward to determine the optimal mix of energy supplies. We leave out the detailed analysis for brevity.

4.14.2 Energy Saving

Research into alternative (green) energy supplies is definitely one way of decreasing our fossil-fuel use. Another is energy saving. To formalize this idea, let the energy composite be written in a somewhat more general way, again emphasizing two energy sources (c and g) only:

$$E = \left(\lambda_c (A_c E_c)^{\rho} + (1-\lambda_c)\left(A_g E_g\right)^{\rho}\right)^{\frac{1}{\rho}}.$$

The technology factors A_i here indicate the "efficiency" with which different energy sources are used. Note, parenthetically, that there is a direct parallel with how we treated energy vs a capital-labor composite in Section 2. Now the A_is introduce asymmetry between the different energy sources through another channel, and moreover we can think of them as being chosen deliberately. One interpretation of these choices is temporary decisions to save on energy, eg, by directing effort toward closing windows or making sure machines don't run unnecessarily. Another interpretation emphasizes research toward energy efficiency that are of a permanent nature. One example is the development of more fuel-efficient cars; another is to develop methods for using less jet fuels when airplanes land. In parallel with our treatment of energy production, we then add the equations

$$A_c = \bar{A}_c m_c^{\zeta} m_g^{1-\zeta} \quad \text{and} \quad A_g = \bar{A}_g m_g^{\zeta} m_c^{1-\zeta},$$

again with the constraint $m_c + m_g = 1$.[cd] With this structure as well, market allocations may end up with specialization for a range of parameter configurations, as will the solution to the planning problem, and path dependence is again possible.

An important concern in the modeling of energy saving or the efficiency of producing energy is that there is a natural upper limit to efficiency. For example, light produced with LED has almost reached the efficiency limit and the same is true for electrical engines. However, this does not mean that we are close to maximal energy efficiency in the production of transportation services. For the transportation example it is less appropriate to capture efficiency through A_g; rather, improvements come about through increasing general energy efficiency (say, a coefficient in front of E in the overall production function). The limits to efficiency are normally not made explicit in economic models but arguably should be in quantitative applications.

4.14.3 Are Subsidies for Green Technology Needed?

To attain the optimal allocation, the planner will of course need to tax the use of fossil fuel. What other taxes and subsidies might be necessary? To the extent there is monopoly power, and the energy sources undersupplied, subsidies are needed. Should the green R&D sector be subsidized? Following Pigou's principle, it should be to the extent there are positive spillovers. So in the absence of technology spillovers in the green R&D sector, there would actually be no reason to subsidize. Moreover, if there are spillovers but they are identical for the two sorts of energy, it is not clear that green technology should receive stronger subsidies than should fossil-fuel technology, so long as fossil fuel is taxed at the optimal rate.

In a second-best allocation, of course, matters are quite different. Suppose no coal tax is used. Then subsidies to the production of green energy, or to the development of new green technologies, would be called for. In political debates, subsidies to the development of green technology appear to be quite popular, and our analysis is in agreement with this view insofar as an optimal (global) carbon tax is not feasible. In practical policy implementation, though less so in debates, it also appears that coal subsidies are popular, perhaps not as per-unit instruments but as support in the construction of plants. A study (Hassler and Krusell, 2014) in fact claims that the average global tax on carbon is set at about the right magnitude but with the wrong sign—owing to large subsidies for coal production across the world.

The view expressed in Acemoglu et al. (2012) appears to contrast somewhat with ours. They argue, based on their model of path dependence, that subsidies to green technology are necessary for attaining an optimum and that carbon taxes would not suffice.

[cd] One can also state these constraints using other functional forms, such as $(\bar{A}_c A_c)^{\zeta} + (\bar{A}_g A_g)^{\zeta} \leq A^{\zeta}$. It is an empirical matter what formulation works best, and it is probably fair to say that the literature is so far silent on this issue.

They obtain this result not only because their model features strong intertemporal spillovers to R&D but also because they make assumptions such that if the "clean good" does not take over from the "dirty good," the climate damages will be infinitely costly (thus, they have strong nonconvexities in their damage function, a tipping point of sorts). Moreover, their model has a second-best structure with spillovers and very limited patent lives. How can we understand this result from the perspective of Pigou taxation? Recall that we pointed out that Pigou taxation may not work if there are multiple market equilibria, and the kind of setting Acemoglu et al. describe has a feature of this kind. The simplest parallel in our static model is the coal-green setup we described in Section 4.14.1. There, we looked at a planning problem with a choice between two energy sources. So suppose that $\overline{\chi}_c = \overline{\chi}_g = \chi$ there, and let us imagine a market allocation where the labor productivity of coal and green energy production, χm_c and χm_g, respectively, derive from variety expansion in patent efforts (m_c and m_g) driven by monopoly profits for intermediate, specialized goods. Suppose, moreover, that there are no research spillovers in this setting: this assumption is perhaps natural in a static model (but less so in a dynamic one). In this framework, then, there would be two equilibria if ρ, the parameter guiding the key energy elasticity, is high enough. Suppose, moreover, that damages are to preferences, as in Section 4.6, and with highly nonlinear features, as discussed in Section 4.7: the marginal damages are first zero for a range of low emission levels, then high and positive, and then again zero in a "disaster zone." Suppose, moreover, that if the economy ends up using coal, emissions will end up in the disaster zone. Then the Pigou procedure would amount to finding the optimal solution—that with green technology only—and an associated tax on carbon that is zero, since the marginal damage at zero emissions is zero. So here Pigou's procedure is highly problematic, since there are now two market outcomes given a zero tax on carbon, and one of them is a disaster outcome! Thus another instrument would be needed to select among the two market outcomes, and one option would be a large enough subsidy to green technology creation to rule out an equilibrium where markets engage in the research on coal technologies.[ce]

4.14.4 Green Technology as a Commitment Mechanism

Some argue that future decision makers cannot be trusted to make good decisions and that, therefore, to the extent we can affect their decisions with irreversible decisions made today, we should. Why would future decision makers not make good decisions? One reason is based on time-inconsistent discounting, as discussed earlier: the current decision maker may have lower discount rates between any two future cohorts than that between the current and next cohort, and if this profile of decreasing discount rates is shared by future cohorts—updated by the appropriate number of cohorts—then profiles are

[ce] With monopolistic competition, one would in general also need to encourage production to prevent undersupply for those technologies that end up being patented.

time-inconsistent. In particular, from the perspective of the current cohort, future cohorts look too impatient. Since future carbon taxes cannot literally be committed to today, then, the current cohort is restricted and appears to not be able to attain its preferred outcome.[cf] Another conceptually distinct reason for disagreements is that politicians (and possibly the voters who support them) may be "myopic"; Amador (2003) shows that rationality-based dynamic voting games in fact can lead to reduced forms characterized by time-inconsistent preferences of politicians.[cg] Finally, Weitzman (1998) provides further arguments for falling discount rates based on the idea that the true future discount rate may be uncertain.

If current decision makers cannot decide directly on the future use of fossil fuels, they may be able to at least influence outcomes, for example by investing in green technology that, ex-post, will tilt the decision makers in the future in the right direction. To illustrate, consider a model where production is given by

$$e^{-\gamma\phi\chi_E n_E}(1-n_E-n_g)^{1-\alpha-\nu}(\chi_E n_E+\chi_g n_g)^{\nu}.$$

$E=\chi_E n_E$ is coal-produced energy and $E_g=\chi_g n_g$ is green energy; we make the assumption, only for obtaining simpler expressions, that these two energy sources are perfect substitutes. Now assume that there is an ex-ante period where an irreversible decision can be made: that on n_g. The cost is incurred ex-post, so only the decision is made ex-ante. Moreover, it is possible to increase n_g ex-post but not decrease it: it is not possible to literally reverse the first decision.[ch] Finally, assume that the ex-ante decision maker perceives a different damage elasticity than the ex-post decision maker (they have different γs, with the ex-ante value higher than the ex-post value): this captures, in a simple way, the intertemporal disagreement.

We make two further simplifying assumptions, for tractability. First, we take the ex-post decision maker to perceive a damage elasticity of exactly 0 and the ex-ante decision maker to use the value $\gamma>0$. Second, we assume that $\chi_E>\chi_g$, ie, that—climate effects aside—the coal technology is a more efficient one for producing energy, regardless of the level at which the two technologies are used (due to the assumption of perfect substitutability). How can we now think about outcomes without commitment?

It is clear that the ex-post decision maker sees no reason to use the green technology at all. Facing a given amount of n_g that he cannot decrease (and will not want to increase), the level of n_E will be determined by the first-order condition

[cf] Karp (2005), Gerlagh and Liski (2012), and Iverson (2014) analyze optimal taxes in the presence of time-inconsistent preferences.

[cg] See also Azzimonti (2011) for a similar derivation.

[ch] We may think of this setup as a reduced-form representation for a case when an ex-ante investment in capital or a new technology makes it profitable to use at least n_g units of labor in green energy production, even if it the emission reduction is not valued per se. In a dynamic model, the cost of this investment would at least partly arise ex-ante, but this is not of qualitative importance for the argument.

$$\frac{1-\alpha-\nu}{1-n_E-n_g}=\frac{\nu\chi_E}{\chi_E n_E+\chi_g n_g}. \tag{17}$$

This expression delivers a linear (affine) and decreasing expression for n_E as a function of n_g: $n_E = h(n_g)$, with $h' < 0$ and independent of n_g.

What is the implied behavior of the ex-ante decision maker without commitment? She will want to maximize

$$e^{-\gamma\phi\chi_E h(n_g)}\left(1-h(n_g)-n_g\right)^{1-\alpha-\nu}\left(\chi_E h(n_g)+\chi_g n_g\right)^{\nu}$$

by choice of n_g, a decision that delivers a second-order polynomial equation as first-order condition, just like in the baseline case (though now with somewhat more involved coefficients in the polynomial). Does this first-order condition admit the first best outcome of the ex-ante decision maker? Such a first best would amount to the solution of the two first-order conditions

$$\gamma\phi\chi_E+\frac{1-\alpha-\nu}{1-n_E-n_g}=\frac{\nu\chi_E}{\chi_E n_E+\chi_g n_g} \tag{18}$$

and

$$\frac{1-\alpha-\nu}{1-n_E-n_g}=\frac{\nu\chi_g}{\chi_E n_E+\chi_g n_g} \tag{19}$$

which result from taking derivatives with respect to n_E and n_g, respectively. It is easy to see that these cannot deliver the same solution as the problem without commitment. For one, Eqs. (19) and (17) cannot deliver the same values for both n_E and n_g, since they differ in one place only and $\chi_E > \chi_g$. Thus, we are in a second-best world where the ex-ante decision maker uses her instrument but cannot, without an additional instrument, obtain her first-best outcome. Moreover, total energy use and/or total labor used to produce energy will be lower with the ex-ante decision on green energy than in the absence of it, comparing Eqs. (17) and (18). This model is stylized and it would appear that the specific predictions could change when moving to a more general setting. However, the second-best nature of the setting would remain.

4.14.5 The Green Paradox

The Green Paradox, a term coined by Sinn (2008), refers to the following logical chain. Decisions to subsidize green technology so as to speed up the research efforts in this direction will, if these efforts are successful, lead to better and better alternatives to fossil fuel over time. This, in turn, implies that fossil-fuel producers have an incentive to produce more in advance of these developments, given that their product is more competitive now than it will be in the future. As an extreme example, imagine that cold fusion is invented but takes one year to implement, so that one year from now we have essentially free, green energy in the entire economy. Then owners of oil wells will produce at maximum capacity today and, hence, there will be much higher carbon dioxide emissions than if cold fusion had not been invented. Hence the "paradox": green technology (appearing in the future) is good but therefore bad (in the short run).

Our static model fully cannot express the Green Paradox, of course, since the essence of the paradox has to do with how events play out over time. Consider therefore a very simple two-period version of the model that allows us to think about how the intertemporal decision for oil producers depends on the availability of green technology. We assume that consumers' preferences are linear so that the gross interest rate is given by $1/\beta$. We assume that fossil fuel is (free-to-produce) oil and that $\rho=1$, so that oil and green energy are perfect substitutes. We also assume that there is no green technology in the first period. A simplified production function thus reads $e^{-\gamma\phi_1 E_1} k^\alpha E_1^\nu$ for period 1 and $e^{-\gamma\phi_1(E_2+\phi_2 E_1)} k^\alpha (E_2+E_g)^\nu$ for period 2; for simplicity, we also abstract from the costs for producing green energy and set E_g to be exogenous, with $n=1$ in both periods). Here, ϕ_1 and ϕ_2 allow us to capture a carbon depreciation process that does not occur at a geometric rate, a feature we argued is realistic. Our notation reveals that capital cannot be accumulated in this example, but we will comment on accumulable capital later.

Given this setting, the price of oil in period 1 is given by $p_1=\nu e^{-\gamma\phi_1 E_1} k^\alpha E_1^{\nu-1}$ and in period 2 it is given by $p_2=\nu e^{-\gamma\phi_1(E_2+\phi_2 E_1)} k^\alpha (E_2+E_g)^{\nu-1}$. All of the available oil, $\bar{E}$, will be used up in the laissez-faire allocation and so oil use in the two periods will be given by the Hotelling condition, a condition we derived and analyzed in Section 2: $p_1=\beta p_2$. Recall that this equation expresses the indifference between producing a marginal unit of oil in period 1 and in period 2. This condition implies that E_1 can be solved for from $e^{-\gamma\phi_1 E_1} E_1^{\nu-1}=\beta e^{-\gamma\phi_1(\bar{E}-E_1(1-\phi_2))}(\bar{E}-E_1+E_g)^{\nu-1}$. Clearly, this equation has a unique solution and comparative statics with respect to E_g shows that more green energy in period 2 makes E_1 rise and E_2 fall. Hence the Green Paradox.

Is the move of emissions from period 2 to period 1 bad for welfare? The negative externality (SCC) of emissions in period 1 is $\gamma\phi_1(y_1+\beta\phi_2 y_2)$ and the present value of the corresponding externality in period 2 is $\gamma\phi_1\beta y_2$. In the absence of a green technology in period 2 ($E_g=0$) it is easy to show that $y_2<y_1$ in the laissez-faire allocation and, hence, at least for a range of positive values of E_g, the externality damage is higher for early emissions. Intuitively, emissions in period 2 have two advantages. One is that they hurt the economy only once: emissions in period 1 will, except for the depreciated fraction $1-\phi_2$, remain in the atmosphere—a significant factor given calibrated carbon-cycle dynamics—and hence also lower second-period TFP. The second advantage of emissions in the future is that their negative effect is discounted (to the extent we assume $\beta<1$). Note, finally, that the possibility of accumulating physical capital would not change any of these conclusions: with more green energy in the second period, capital accumulation with rise somewhat to counteract the initial effect, and it would work toward an increase in p_2, but this mechanism would not overturn our main observation.

Can the future appearance of green technology also make overall welfare go down in the laissez-faire allocation? This is much less clear, as an additional unit of E_g (for free) has a direct positive welfare effect.[ci] However, now consider competitive production of green energy under laissez-faire, at a unit labor cost χ_g. Here, a second-best argument would suggest that there is a negative "induced externality" of green energy production: since the economy is far from the optimum, and emissions in period 1 would be detrimental, any additional unit of E_g would have a negative side-effect on welfare. Hence, a(t least a small) *tax* on green energy production would be desirable! The reason for this perhaps counterintuitive effect—aside from the Green-Paradox logic—is that the total amount of fossil fuel used will still be $\bar{E}$: green technology, in this setting, will not curb the use of fossil fuel, only change the timing of emissions (in the wrong direction).

The previous example points to counterintuitive policy implications: green technology should be discouraged. However, aside from the assumptions that make the Green Paradox relevant, this result also relies on second-best analysis. In the social optimum, green technology should not be taxed (nor subsidized): there is, simply, no externality from producing green technology in this model. If green technology is developed in an R&D activity, then support of this activity (relative to other activities) may be called for, but only if there is an R&D externality to green technology development that is, in the appropriate sense, larger than the corresponding one for fossil-fuel technology developments. Hence, the optimum (in this economy, where oil is free to produce) involves fossil-fuel taxes but no net support to green technology.

Is the Green Paradox empirically relevant? The key assumption that leads to the paradox is that the accumulated use of fossil fuel is the same under laissez-faire as in the optimal allocation. In this case, suboptimality only comes from the speed at which the fossil reserves are used. That all reserves are used also in the optimal allocation is arguably reasonable when it comes to conventional oil with low extraction costs (eg, Saudi oil). However, it is not reasonable for nonconventional reserves and coal. Here, policy, including subsidies to the development of future green energy production, can and should affect how much fossil resources are left in ground. So suppose, instead, that we focus on fossil fuel in the form of coal and that we maintain our assumption that the marginal cost of coal is constant (in terms of labor or some other unit). Then an increase in E_g would lead to a lower demand for coal and hence have an impact on coal use: it would clearly induce lower coal production in the second period. Lower coal use, in turn, has a positive externality on the economy. Moreover, coal use in period 1 is not affected. Hence, the conclusion here is the opposite one: green energy has a positive effect on the economy (beyond its direct positive effect, to the extent it comes for free).

[ci] If there are strong nonlinearities, like a threshold CO_2 concentration level above which climate damages are catastrophic, then the introduction of a green technology in the second period could make laissez-faire welfare fall.

In addition, relative to a laissez-faire allocation it would be beneficial to subsidize, not tax, green energy production. Which case appears most relevant? We take the view that the latter is more relevant. The argument has two parts. First, the intertemporal reallocation of emissions emphasized in the Green-Paradox argument, though logically coherent, is not, by our measure, quantitatively important. The main reason is that the total amount of oil is rather small and its effect on climate is limited, and a reallocation of emissions due to oil over time is of second-order importance compared to being able to control the cumulated (over time) emissions. Second, if the fossil fuel is costly to extract then there would be lower emissions, as argued earlier, and in terms of the total amount of fossil fuel available, most of it is costly to extract (most of it is coal). Coal is produced at a price much closer to marginal cost and the Hotelling part of the coal price appears small. This argument, moreover, is quantitatively important given the large amounts of coal available.

4.15 Regional Heterogeneity

Nordhaus's basic DICE model is a one-region integrated assessment model, but there are by now several calibrated models in the literature with more than one region. His own RICE (R for Regional) model was perhaps the first multiregion model and it had 7 regions, defined by geographic and economic indicators; Krusell and Smith (2015) have developed a model at the extreme end of heterogeneity, treating one region as a 1-by-1-degree square with land mass on the global map. Regional models can serve a variety of purposes and we first briefly discuss the chief purposes. We then use a multiregion version of our basic model as an illustration; in particular, we use a simple version of Hassler and Krusell (2012) and look at some extensions.

A major purpose for looking at regional heterogeneity comes from recognizing that damages are very different in different parts of the world; some regions, such as Canada and most of Russia, are even expected to gain from a warmer climate. Thus, using a multiregion IAM as a simulation device, one can trace out the heterogeneous effects of climate change under different policy scenarios. Even if there is no agreement on a social welfare function for the world, surely policymakers are very interested in this heterogeneity.

Another purpose of a multiregion IAM is to look at the effects of regionally heterogeneous policies. Suppose the Western world adopts a strict carbon tax and the rest of the world does not. How effective will then the western policies be in combatting climate change, and what will its distributional consequences be?

Relatedly, one of the key concepts in policymakers' studies of climate change is *carbon leakage*. The idea here is simply that when carbon is taxed at higher rates in some regions than in others, the decreases in carbon use in the high-tax regions will presumably be (partially, or fully) offset by increases in carbon use in other regions. *Direct carbon leakage* would for example occur if the oil shipments are simply redirected away from low-tax to high-tax regions. But there can also be *indirect carbon leakage* in that the other factors of production (capital and/or labor) can move to where carbon taxes are lower—and hence carbon will be used more there as a result. Differential policies can also affect outcomes through trade (see, eg, Gars, 2012 and Hémous, 2013). Finally, when there is R&D in the development of fossil-fuel and green technologies, differential policies in this regard come into play as well (Hémous, 2013, looks at this case as well).

Still another important aspect of a multiregion IAM is its potential for discussing *adaptation* to climate change through the migration of people (along with other production factors).[cj] Adaptation is not just important in practice but it is important to think about from a theoretical and quantitative perspective since the damages from climate change really are endogenous and depend on how costly it is to migrate. If migration were costless, significant warming would potentially be less detrimental to human welfare since there are vast areas on our continents that are too cold today but, with significant warming, inhabitable. There is very little research on this issue so far (Brock et al., 2014 and Desmet and Rossi-Hansberg, 2015 are promising exceptions) but we believe it is an important area for future research and one with much potential. Empirical research on the costs of migration is also scant, but some work does exist (Feng et al., 2010 and, for a study of conflict in this context, see Harari and La Ferrara (2014); see also the review Burke et al., 2015).

4.15.1 A Two-Region IAM with Homogeneous Policy: Oil

Our simple model is easily extendable to include another region (or more). Let us look at a series of simple cases in order to illustrate some of the main points made in the literature.[ck] Let us first look at heterogeneous damages, so assume that production in region 1 is $e^{-\gamma_1 E}k_1^{\alpha}n_1^{1-\alpha-\nu}E_1^{\nu}$ whereas production in region 2 is $e^{-\gamma_2 E}k_2^{\alpha}n_2^{1-\alpha-\nu}E_2^{\nu}$. Energy is coming from fossil fuel only, and let us first assume that it is (costless-to-produce) oil available at a total amount $\bar{E}$ in a third region of the world, which supplies the oil under perfect competition (the third region thus plays no role here other than as a mechanical supplier of oil). Let us also for simplicity start out by assuming that the two regions are homogeneous in the absence of climate damages, so that $k_1 = k_2 = k$ and $n_1 = n_2 = n$. It is easy to work out a laissez-faire equilibrium for this world and we can look at different cases, the first of which is that when neither capital nor labor can move. Thus, the only trade that occurs takes the form that the oil-producing region sends oil to the two other regions and is paid in consumption goods; regions 1 and 2 do not interact, other than by trading in the

[cj] For a recent example, see Krusell and Smith (2015), who allow for the migration of capital.

[ck] It should be noted, however, that there are very few examples of multiregion IAM that are studied in full general equilibrium. Thus the number of formal results from the literature is therefore very limited relative to the number of informal conjectures.

competitive world oil market. All of the oil will be used and the equilibrium oil distribution will now be determined by the following condition:

$$e^{-\gamma_1 \bar{E}} E_1^{\nu-1} = e^{-\gamma_2 E} E_2^{\nu-1},$$

ie, by ($E_1 + E_2 = \bar{E}$ and)

$$\frac{E_1}{E_2} = e^{\frac{\gamma_2 - \gamma_1}{1-\nu} \bar{E}}.$$

Thus, the relative use of oil is higher in the country with lower climate damages.[cl] Suppose that region 1 experiences stronger damages. Clearly, then, region 1 is worse off and the damage has a small "multiplier effect" to the extent that its energy used is curbed: more energy is used in region 2. In other words, we would see lower TFP in region 1 but lower activity there also because of reduced energy use. Consumption is a fraction $1-\nu$ of output, with the remainder sent to the third, oil-producing region.

If we also allow capital to move—but maintain that the populations cannot move—the output effect will be somewhat strengthened as capital will also move to region 2 to some extent. If half of capital is owned by each region, this makes region 1 gain, however, because its GNP will rise even though its GDP will fall. In the real world, there are moving costs and cultural and other attachments to regions, so full and costless migration is probably not an appropriate assumption even in the long run (as the static model is supposed to capture a longer-run perspective).

Suppose now that regions 1 and 2 consider a common tax τ on carbon and suppose that this tax is collected in each country and redistributed back lump-sum to the local citizens. Would such a tax be beneficial? To regions 1 and 2, yes. The analysis depends on the size of the tax but suppose the tax is low enough that firms are not sufficiently discouraged from using oil that the total amount of oil use is lowered. Then the relative energy uses in the two regions will still satisfy the equations above and the levels will not change either. The price of oil, p, will satisfy

$$p = \nu y_1 / E_1 - \tau,$$

the first term of which is independent of the tax size (for a small enough tax). Hence, country i's consumption will now be $y_i - (p+\tau)E_i + \tau E_i = (1-\nu)y_i + \tau E_i$ so that consumption is strictly increasing in τ for both regions. Thus, the two regions can use the tax to shift oil revenues from the oil-producing region to its own citizens, without affecting output at all.[cm] When the tax is high enough that p reaches zero, the level of production responds to taxation: as producers now receive nothing for their oil, they

[cl] Of course this result depends on damages occurring to TFP; if they affect utility, oil use is identical in the two regions.

[cm] This argument is of course unrelated to any climate externality; the climate is unaffected by the taxation.

are indifferent as to how much to supply. At that tax level, the total energy supply will still be given by $\bar{E}$ and the equations above, but now consider a slightly higher tax, still with a zero price of oil. Then the total amount of energy E is then lower and is determined from

$$\tau = \nu e^{-\gamma_1 E} k^{\alpha} n^{1-\alpha-\nu} E_1^{\nu-1} \quad \text{and} \quad \frac{E_1}{E-E_1} = e^{\frac{\gamma_2-\gamma_1}{1-\nu}E}.$$

It is straightforward to show, if the γs are not too far apart, that these two equations imply a lower E and E_1 as τ is raised and that E_1/E_2 will rise. Now, for each region there would be an optimal τ and there would be a conflict between these two values. Generally, the region with a higher climate externality would favor a higher tax.

4.15.2 A Two-Region IAM with Homogeneous Policy: Coal

These discussions all refer to the case of oil, ie, a free-to-extract fossil fuel. Suppose we instead look at coal, and assume that coal is domestically produced: it costs $1/\chi_i$ units of labor per unit, as in most of our analysis earlier. We also assume that the transport costs for coal are inhibitive so that there is no trade at all. The only connection between the regions is thus the climate externality. In the absence of taxes the world equilibrium is then determined independently of the externality and according to

$$\frac{1-\alpha-\nu}{\chi_i - E_i} = \frac{\nu}{E_i}$$

for $i=1,2$.

Now the reason to tax in order to transfer resources away from a third region and to the home country is no longer applicable; the only reason to tax is the climate externality. As in the oil case, let us assume that any tax on coal is lump-sum transferred back to domestic consumers. What is then the best outcome for each of the two regions? The two countries can, in principle, act in a coordinated fashion so as to maximize overall welfare—by maximizing world output—and then choose a point on the Pareto frontier by the use of transfers. World output is maximized by setting the tax equal to the marginal damage externality in the world, ie, $\gamma_1 y_1 + \gamma_2 y_2$. Thus, the social planner chooses E_1 and E_2 to solve

$$\gamma_1 e^{-\gamma_1(E_1+E_2)} k_1^{\alpha} \left(1 - \frac{E_1}{\chi_1}\right)^{1-\alpha-\nu} E_1^{\nu} + \gamma_2 e^{-\gamma_2(E_1+E_2)} k_2^{\alpha} \left(1 - \frac{E_2}{\chi_2}\right)^{1-\alpha-\nu} E_2^{\nu} =$$

$$e^{-\gamma_1(E_1+E_2)} k_1^{\alpha} \left(1 - \frac{E_1}{\chi_1}\right)^{1-\alpha-\nu} E_1^{\nu} \left(\frac{1-\nu-\alpha}{\chi_1 - E_1} - \frac{\nu}{E_1}\right) =$$

$$e^{-\gamma_2(E_1+E_2)} k_2^{\alpha} \left(1 - \frac{E_2}{\chi_2}\right)^{1-\alpha-\nu} E_2^{\nu} \left(\frac{1-\nu-\alpha}{\chi_2 - E_2} - \frac{\nu}{E_2}\right).$$

The first line represents the global damage externality (which is also the optimal tax on coal); it has to be set equal to the net benefit of emissions in each of the two regions (the following two lines). The allocation will have lower E_1 and E_2 amounts (provided, at least, both γs are positive) than in the laissez-faire allocation.

Suppose, however, that the regions cannot use transfers to arrive at a Pareto-optimal allocation. Then an optimal allocation would be obtained by maximizing a weighted value of the utilities of consumers in the two regions. Often, macroeconomic models adopt the utilitarian approach. Assuming, as in a benchmark case above, logarithmic utility of consumption, and a utilitarian social welfare function, we would then need to solve

$$\max_{E_1,E_2} \log\left(e^{-\gamma_1(E_1+E_2)}k_1^{\alpha}\left(1-\frac{E_1}{\chi_1}\right)^{1-\alpha-\nu}E_1^{\nu}\right)+\log\left(e^{-\gamma_2(E_1+E_2)}k_2^{\alpha}\left(1-\frac{E_2}{\chi_2}\right)^{1-\alpha-\nu}E_2^{\nu}\right).$$

This problem delivers two simple first-order conditions:

$$\gamma_1+\gamma_2=\frac{1-\nu-\alpha}{\chi_1-E_1}-\frac{\nu}{E_1}=\frac{1-\nu-\alpha}{\chi_2-E_2}-\frac{\nu}{E_2}.$$

It is easy to see from these two equations the only parameters that influence emissions in country i are parameters specific to that country plus the damage elasticity parameter of the other country. Suppose now that we try to back out what tax on coal in country i would be necessary to attain this allocation. From the firm's first-order condition we obtain

$$\tau_i=e^{-\gamma_1(E_1+E_2)}k_i^{\alpha}\left(1-\frac{E_i}{\chi_i}\right)^{1-\alpha-\nu}E_i^{\nu}\left(\frac{1-\nu-\alpha}{\chi_i-E_i}-\frac{\nu}{E_i}\right).$$

Let us now evaluate the right-hand side at the utilitarian optimum as given by the previous equations. This delivers

$$\tau_i=(\gamma_1+\gamma_2)e^{-\gamma_1(E_1+E_2)}k_i^{\alpha}\left(1-\frac{E_i}{\chi_i}\right)^{1-\alpha-\nu}E_i^{\nu}.$$

Does this imply a uniform tax across countries? The answer is no. We obtain, in particular, that

$$\frac{\tau_1}{\tau_2}=\left(\frac{k_1}{k_2}\right)^{\alpha}\left(\frac{1-\dfrac{E_1}{\chi_1}}{1-\frac{E_2}{\chi_2}}\right)^{1-\alpha-\nu}\left(\frac{E_1}{E_2}\right)^{\nu}=\frac{\gamma_1}{\gamma_2}.$$

Clearly, this expression is not 1 in general. It depends on the ratio of capital stocks (note that E_1 and E_2 do not) and the expression involving the Es and χs is also not equal to 1 in general: it is above (below) 1 if χ_1 is above χ_2. In the latter case, the richer country imposes a larger tax on carbon. Note, however, that we obtain a common tax rate, ie, a *common tax on coal per output unit.*

We have learned from the earlier analysis (i) that the Pareto optimum involves a globally uniform tax on coal (along with some chosen lump-sum transfers across regions) but (ii) the utilitarian optimum assuming no transfers across regions does not, and instead prescribes—in the benchmark case we look at—a tax that is proportional to the country's output. It is straightforward to go through a similar exercise with population sizes differing across regions; in this case, the optimal tax rate in region i is equal to the region's per-capita income times the world's population-weighted γs.

4.15.3 Policy Heterogeneity and Carbon Leakage

International agreements appear hard to reach and it is therefore of interest to analyze policy heterogeneity from a more general perspective. So suppose region 1 considers a tax on its fossil fuel but knows that region 2 will not use taxes. What are the implications for the output levels of the two regions and for the climate implied by such a scenario? We again begin the analysis by looking at the case of oil, and we start off by assuming that neither capital nor labor can move across regions.

In a decentralized equilibrium, oil use in region 1 is given by

$$p+\tau=\nu e^{-\gamma_1(E_1+E_2)}k_1^\alpha n_1^{1-\alpha-\nu}E_1^{\nu-1}$$

and in region 2 it is given by

$$p=\nu e^{-\gamma_2(E_1+E_2)}k_2^\alpha n_2^{1-\alpha-\nu}E_2^{\nu-1}.$$

Thus, we can solve for E_1 and E_2 given $E_1+E_2\leq\bar{E}$. Clearly, we must have $p>0$—otherwise, region 2 would demand an infinite amount of oil—and so we first conclude that $E_1+E_2=\bar{E}$: there is no way for one country, however large, to influence total emissions. What the tax will do is change energy use across regions: region 1 will use less and region 2 more. Moreover, in utility terms region 1 is worse off and region 2 better off from this unilateral tax policy. This example illustrates direct (and full) carbon leakage: if one region taxes oil, oil use will fall in this region but there will be an exact offset elsewhere in the world.

In the coal example, the situation is rather different. The laissez-faire allocation is now given by

$$\tau_1=e^{-\gamma_1(E_1+E_2)}k_1^\alpha\left(1-\frac{E_1}{\chi_1}\right)^{1-\alpha-\nu}E_1^\nu\left(\frac{1-\nu-\alpha}{\chi_1-E_1}-\frac{\nu}{E_1}\right)$$

and

$$0 = \frac{1-\nu-\alpha}{\chi_2 - E_2} - \frac{\nu}{E_2}.$$

We see that coal use in region 2 now is independent of the tax policy in region 1.[cn] It is easy to show that region 1's coal use will fall and that, at least if both γs are positive and locally around $\tau_1 = 0$, welfare will go up in both regions. There will be an optimal tax, from the point of view of region 1's utility, and it is given by the SCC (computed ignoring the negative externality on region 2), ie, $\gamma_1 y_1$.

If one allows capital mobility, as in Krusell and Smith (2015), there will be indirect carbon leakage. In the case of oil, a tax in region 1 would act as a multiplier and tilt the relative oil use more across regions, ie, increase the leakage. In the case of coal, whereas there is no leakage when capital cannot flow, there is now some leakage: the lower use of coal will decrease the return to capital in region 1 and some capital will then move to region 2, in turn increasing emissions there. We thus see that the extent of leakage depends on (i) how costly fossil fuel is to extract and (ii) to what extent other input factor flow across regions.[co]

It would be straightforward to apply this model, and even dynamic versions of it as they can allow closed-form analysis, for a range of qualitative and quantitative studies. A recent example is Hillebrand and Hillebrand (2016), who study tax-and-transfer schemes in a dynamic multiregion version of the model.

4.15.4 More Elaborate Regional Models

Multiregion models of the sort discussed here can be applied rather straightforwardly, and without much relying on numerical solution techniques, in a number of directions. However, some extensions require significant computational work. One example is the case where the intertemporal cross-regional trade is restricted; a specific case is that where there are shocks and these shocks cannot be perfectly insured. Krusell and Smith (2014, 2015) study such models and also compare outcomes across different assumptions regarding such trade; in their models with regional temperature shocks, the model is similar to that in Aiyagari (1994), with the Aiyagari consumers replaced by regions, and where the numerical methods borrow in part from Krusell and Smith (1998). The Krusell and Smith (2015) model has regions represent squares that are 1 by 1 degree on the map; Nordhaus's G-Econ database with population and production on that level of aggregation can then

[cn] Our particular assumptions on how coal is produced explains why there is no effect at all on coal use in region 2: the costs and the benefits of coal are both lowered by the same proportion as a result of the tax in region 1. With coal produced with a constant marginal cost in terms of output (as opposed to in terms of labor), there would be a small effect on region 2's coal use.

[co] We did not consider the case where coal is costless to trade and potentially produced in a third region but it is straightforwardly analyzed.

be used to calibrate the model. Thus, the calibration makes the initial model output distribution match that in the data, and the marginal products of capital are assumed to be equal initially—these two restrictions are made possible by choosing TFP and capital-stock levels for each region. There is also heterogeneity in two aspects of how regions respond to climate change. One is that for any given increase in global temperature, the regional responses differ quite markedly according to certain patterns, as discussed in Section 3.1.4; Krusell and Smith use the estimates implied by a number of simulations of advanced climate models to obtain region-specific parameters. These estimated "climate sensitivities" are plotted by region on the global map in left panel of Fig. 13.

A second element is differences in damages from climate change across regions. In the latest version of their work and as mentioned in Section 3.3.3, Krusell and Smith use the assumption that there is a common, U-shaped damage function for all regions defined in terms of the local temperature, ie, there ideal temperature is the same at all locations. This common damage function has three parameters which are estimated to match, when the model is solved, the aggregate (global) damages implied by Nordhaus's DICE damage function for three different warming scenarios (1, 2.5, and 5 degrees of global warming). The estimates imply that an average daily temperature of 11.1°C (taken as a 24-h average) is optimal.

The right panel of Fig. 13 displays the model's predicted laissez-faire outcomes in year 2200. We see large gains in percent of GDP in most of the northern parts of the northern hemisphere and large losses in the south. Overall, the damage heterogeneity is what is striking here: the differences across regions swamp those obtained for any comparisons over time of global average damages. The results in this figure of course rely on the assumption that the damage function is the same everywhere so that warming implies gains for those regions that are too cold initially and losses for those that are too warm. This, however, seems like a reasonable assumption to start with and, moreover, is in line with recent damage-function estimates using cross-sectional data: see Burke et al. (2015). These results at the very least suggest that the returns from further research on heterogeneity should be rather high.

We already mentioned Hémous's (2013) work on the R&D allocation across regions, emphasizing the importance of understanding the determinants and consequences of the regional distribution of R&D and of trade in goods with different carbon content.[cp] Another very promising and recent line of research that we also made reference to above is that on endogenous migration pursued in Brock et al. (2014) and Desmet and Rossi-Hansberg (2015). The latter study, which is an early adopter of the kind of damage-function assumption (for both agriculture and manufacturing) used in the later study by Krusell and Smith (2015), assumes free mobility and that there is technology heterogeneity across regions, with operative region-to-region spillovers. The model

[cp] See also Acemoglu et al. (2014).

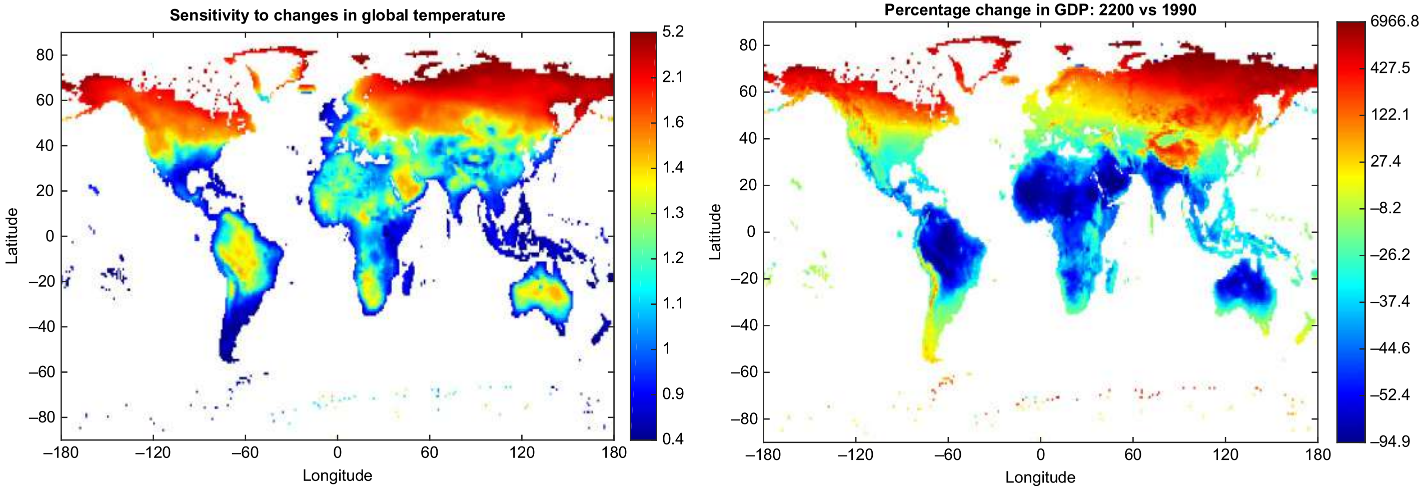


Fig. 13 *Left*: temperature increases for global warming of 1 degree. *Right*: simulation of Krusell and Smith (2015) model, future % GDP losses under laissez-faire.

structure used by Desmet and Rossi-Hansberg is particularly tractable for the analysis of migration, as it uses indifference conditions to distribute agents across space. In contrast, models where location is a state variable (in a dynamic sense) and moving is costly are much more difficult to characterize, as moving then is a highly multidimensional and nonlinear problem both with regard to state and control variables. Stylized two-region models like those studied herein and in Hémous's work can perhaps be solved for endogenous migration outcomes but full dynamics are probably very challenging to solve for.

5. DYNAMIC IAMS

Even though the static IAM setting analyzed in the previous section is useful in many ways, its value in quantitative evaluations is limited: climate change plays out very slowly over time—the dynamics of the carbon cycle especially—and the intertemporal economics aspects involving the comparison between consumption today and consumption far out in the future are therefore of essence. Thus, a quantitatively oriented integrated assessment model of economics and climate change needs to incorporate dynamics. In addition, there are some conceptual issues that cannot be properly discussed without a dynamic setting, such as time preferences.

To our knowledge, the first steps toward modern integrated assessment model appear in Nordhaus (1977). A little over a decade later, Nordhaus developed a sequence of dynamic models, all in the spirit of the simple model above, but formulated in sufficient complexity that numerical model solution is required. The core, one-region version of Nordhaus's model is DIce: a Dynamic Integrated Climate-Economy model, described in detail in Nordhaus and Boyer (2000). In one respect, almost all the dynamic IAMs, including Nordhaus's, are more restrictive than the setting in our previous section: they focus on a planning problem, ie, on characterizing optimal allocations. That is, decentralized equilibria without carbon policy, or with suboptimal carbon policy, are rarely analyzed, let alone explicitly discussed in dynamic models.[cq] In our present treatment, we insist on analyzing both optima and suboptimal equilibria, in large part because the quantitative assessments of the "cost of inaction" cannot be computed otherwise.

In what follows we will discuss a general structure for which we define the social cost of carbon and, under some additional assumptions, can derive a simple and directly interpretable formula for the tax. It is a straightforward extension of the results from the static model above. This material is contained in Section 5.1. In Section 5.2 we then make further assumptions, relying also on the finite-resource modeling from Section 2, and simplify the general structure so as to arrive at an easily solved, and yet quantitatively reasonable, model that can be used for positive as well as normative analysis. Throughout, the discussion follows Golosov et al. (2014) rather closely.

[cq] For an exception, see, eg, Leach (2007).

5.1 The Social Cost of Carbon in a General Dynamic Model

We now focus on how the SCC is determined in a dynamic setting that is reasonably general. For this, we use a typical macroeconomic model with a representative (for the global economy, at this point) agent, as in Nordhaus's DICE model, a production structure, and a specification of the climate system as well as the carbon cycle.

The representative agent has utility function

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t U(C_t),$$

where U is a standard, strictly concave utility function of (the one and only) consumption good C and where $\beta \in [0,1)$ is the discount factor. The resource constraint for the consumption good is more broadly a constraint for the final good, because like in most of the macroeconomic literature we treat consumption and investment as perfect substitutes. The constraint thus reads

$$C_t + K_{t+1} = Y_t + (1-\delta)K_t,$$

which involves a typical capital accumulation specification with geometric depreciation at rate δ and where Y denotes global output. Global output, in turn, is generated from

$$Y_t = F_{0,t}(K_{0,t}, N_{0,t}, \mathbf{E}_{0,t}, S_t).$$

Here, "0" represents the sector producing the final good. The function F_0 is assumed to display constant returns to scale in the first three inputs. $N_{0,\ t}$ is labor used in this sector and $\mathbf{E}_{0,t} = (E_{0,1,t}, \ldots, E_{0,I,t})$ denotes a vector of different energy inputs. We use a subindex t on the production function to indicate that there can be technical change over time (of various sorts and deterministic as well as stochastic). S, finally, is atmospheric carbon concentration, and it appears in the production function because it causes damages—through the effect of S on the climate (in particular through the temperature).

In our formulation here, as discussed earlier, we adopt the common assumption that damages only appear in the production function. Moreover, they only appear in the time-t production function through atmospheric carbon concentration at t, thus subsuming the mapping from S to temperature and that from temperature to output loss in one mapping. As we already argued, these assumptions are convenient in that they map neatly into Nordhaus's DICE model. We should remind the reader that the inclusion of only S_t in the damages at t captures a lack of dynamics; as we pointed out, this should still be a reasonable approximation to a more complex setting where, conceptually, one would include past values of S in the production function at t as a way of capturing the full dynamics. An extension to include such lagged variables is straightforward but would not greatly change the results as the temperature dynamics are rather quick.

Turning to energy production, we assume that there are $I_g - 1$ "dirty" energy sources (involving fossil fuel), $i = 1, \ldots, I_g - 1$, and a set of green sources, $i = I_g, \ldots, I$. Each component of $\mathbf{E}_{0,\,t}$, $E_{0,\,i,\,t}$ for $i = 1, \ldots, I$, is then produced using a technology $F_{i,\,t}$, which uses the three inputs capital, labor, and the energy input vector. Some energy sources, such as oil, may be in finite supply. For those i in finite supply, $R_{i,\,t}$ denotes the beginning-of-period stock at t and $E_{i,\,t}$ the total amount extracted (produced) at t. Thus, the exhaustible stock i evolves as

$$R_{i,t+1} = R_{i,t} - E_{i,t} \geq 0. \tag{20}$$

Production for energy source i, whether it is exhaustible or not, is then assumed to obey

$$E_{i,t} = F_{i,t}(K_{i,t}, N_{i,t}, \mathbf{E}_{i,t}, R_{i,t}) \geq 0. \tag{21}$$

The resource stock appears in the production function because the production costs may depend on the remaining resource stock. Notice, also, that S_t does not appear in these production functions: we assume that climate change does not cause damages to energy production. This, again, is a simplification we make mainly to adhere to the TFP damage specification that is common in the literature, but it also simplified formulas and improves tractability somewhat. Given that the energy sector is not so large, this simplification should not be a major problem for our quantitative analysis.

To close the macroeconomic part of the model, we assume that inputs are allocated across sectors without costs, again a simplifying assumption but one that appears reasonable if the period of analysis is as long as, say, 10 years. Thus we have

$$\sum_{i=0}^{I} K_{i,t} = K_t, \quad \sum_{i=0}^{I} N_{i,t} = N_t, \quad \text{and} \quad E_{j,t} = \sum_{i=0}^{I} E_{i,j,t}. \tag{22}$$

We assume that the sequence/process for N_t is exogenous.

Finally, we let the carbon cycle generally be represented by a function $\widetilde{S}_t$ as follows:

$$S_t = \widetilde{S}_t\left(E^f_{i,-T}, E^f_{-T+1}, \ldots, E^f_t,\right). \tag{23}$$

Here, T periods back represents the end of the preindustrial era and $E^f_s \equiv \sum_{i=1}^{I_g - 1} E_{i,s}$ is fossil emission at s and we recall that $E_{i,\,s}$ is measured in carbon emission units for all i. When we specialize the model, we will adopt a very simple structure for $\widetilde{S}_t$ that is in line with the discussion in the section earlier on the carbon cycle.

We are now ready to state an expression for the SCC. Using somewhat abstract (but obvious) notation, and denoting the social cost of carbon at time t, in consumption units at this point in time, by SCC_t, we have

$$\text{SCC}_t = \mathbb{E}_t \sum_{j=0}^{\infty} \beta^j \frac{U'(C_{t+j})}{U'(C_t)} \frac{\partial F_{0,t+j}}{\partial S_{t+j}} \frac{\partial S_{t+j}}{\partial E^f_t}. \tag{24}$$

Before we discuss this equation, let us emphasize—as we pointed out in the context of the static model—that this expression amounts to keeping decisions fixed as emissions are increased incrementally. Ie, this concept of the social cost of carbon does not correspond to a policy experiment (where presumably induced changes in decisions would add indirect damage effects, positive or negative). Golosov et al. (2014) derive this equation as part of an optimal allocation but then the interpretation really is that the right-hand side equals the OSCC$_t$.

Eq. (24) is easily interpreted. First, $\frac{\partial S_{t+j}}{\partial E_t^f}$ captures the carbon cycle dynamics: it tells us how much the atmospheric carbon content j periods ahead is increased by a unit emission at t. That amount of increase in S_{t+j} then changes final output in period $t+1$ by $\frac{\partial F_{0,t+j}}{\partial S_{t+j}}$ per unit. The total effect (the multiplication of these two factors), which is presumably negative, is the marginal damage in that period in terms of the final output good arising from a unit of emission at t. To translate this amount into utils at $t+j$ one multiplies by $U'(C_{t+j})$, and to bring the utils at $t+j$ back to time-t utils one multiplies by β^j: utility discounting. The division by $U'(C_t)$ then translates the amount back into consumption units at t. Finally, since one needs to take into account the effect of emissions at all points in time $t, t+1, \ldots$, one needs the infinite sum.

Conceptually, thus, Eq. (24) really is straightforward. However, in its general form it is perhaps not so enlightening. A key result in Golosov et al. (2014) is that with some assumptions that the authors argue are weak, one can simplify the formula considerably and even arrive at a closed-form expression in terms of primitive parameters. We present the assumptions one by one.

Assumption 1. $U(C) = \log C$.

Logarithmic utility, both used and relaxed in our static model, is very often used in macroeconomic models and seems appropriate as a benchmark. It embodies an assumption about the intertemporal elasticity of consumption but obviously also about risk aversion.

Assumption 2.

$$F_{0,t}(K_{0,t}, N_{0,t}, \mathbf{E}_{0,t}, S_t) = \exp(-\gamma_t S_t)\widetilde{F}_{0,t}(K_{0,t}, N_{0,t}, \mathbf{E}_{0,t}),$$

where we have normalized so that S is the atmospheric CO_2 concentration in excess of that prevailing in preindustrial times, as in the earlier section, and where γ can be time- and state-dependent.

This assumption was discussed in detail in Section 3.3: we argue that it allows a good reduced-form approximation to the most commonly used assumptions on the S-to-temperature and the temperature-to-damage formulations in this literature.

Assumption 3.

$$S_t = \sum_{s=0}^{t+T} (1 - d_s) E^f_{t-s} \tag{25}$$

where $d_s \in [0, 1]$ for all s.

A linear carbon cycle was also discussed Section 3.2.4 on carbon circulation above and argued to be a good approximation. The linear structure was also simplified further there, and we will use that simplification below.

Assumption 4.
C_t/Y_t does not depend on time.

This assumption, which is tantamount to that used in the textbook Solow model, is not an assumption on primitives as we usually define them. However, it is an assumption that can be shown to hold exactly for some assumptions on primitives—as those that will be entertained below—or that holds approximately in a range of extensions; see Barrage (2014). Major changes in saving behavior away from this assumption are needed to drastically alter the quantitative conclusions coming out of our SCC formula.

Now given these four assumptions only a minor amount of algebra suffices to arrive at a formula for the SCC, as well as for the optimal tax on carbon. It is

$$\text{SCC}_t = Y_t \left[\mathbb{E}_t \sum_{j=0}^{\infty} \beta^j \gamma_{t+j} (1 - d_j) \right]. \tag{26}$$

As can be seen, this formula is a straightforward extension of that arrived at for the static economy. As in the static economy, the formula for the tax as a fraction of output is a primitive: there, simply γ; here, a present value of sorts of future γs. Note, of course, here as well as for the static model, that if one needs to assign a specific value to the optimal tax, one would strictly speaking need to evaluate output at its optimal level, and the optimal level of output is not expressed in closed form here (and may be cumbersome to compute). However, given our quantitative analysis later, we note that the optimal tax rate does not alter current output so much. Hence, a good approximation to the optimal tax rate is that given by the expression in brackets in Eq. (26) times current output.[cr]

In the static economy, we assumed a Cobb–Douglas form for output, as we will in the next section as well for our positive analysis. However, Cobb–Douglas production

[cr] In the dynamic model, this approximation would overstate the exact value of the tax since optimal output in the short run will be lower than laissez-faire output. In the static model with TFP damages, the reverse inequality will hold.

is apparently not necessary for the result earlier. What is true is that Cobb–Douglas production, along with logarithmic utility and 100% depreciation for capital, are very helpful assumptions for arriving at a constant C/Y ratio (Assumption 4), but we also know that an approximately constant C/Y ratio emerges out of a much broader set of economies.

We note that, aside from the damage parameter γ, utility discounting and carbon depreciation now matter very explicitly as well. This is quite intuitive: it matters how long a unit of emitted carbon stays in the atmosphere and it also matters how much we care about the future. As for how γ appears, note that the formula is an expectation over future values—as in the static model, a certainty equivalence of sorts applies—but that one could also imagine γ as evolving over time, or incorporating different amounts of uncertainty at different points in time.[cs] Of course, suppose more information is revealed about γ as time evolves, the optimal tax will evolve accordingly (as, eg, in a specification where γ is assumed to follow a unit-root process).

A final expression of our SCC is obtained by (i) assuming that $\mathbb{E}_t\left[\gamma_{t+j}\right]=\overline{\gamma}_t$ for all j (as for example for a unit root process) and (ii) letting the $1-d_j$s be defined by Eq. (13) (which we argued gives a good account of the depreciation patterns). Then we obtain

$$\mathrm{SCC}_t/Y_t=\overline{\gamma}_t\left(\frac{\varphi_L}{1-\beta}+\frac{(1-\varphi_L)\varphi_0}{1-(1-\varphi)\beta}\right). \tag{27}$$

Here, the expression inside the parenthesis on the right-hand side can be thought of as the *discount-weighted duration of emissions*, an object that is stationary by assumption here.

A remarkable feature of the formula for the SCC as a fraction of output as derived here is that it depends on very few parameters. In particular, no production parameters appear, nor do assumptions about technology or the sources of energy. In contrast, we will see in the positive analysis below that such assumptions matter greatly for the paths of output, the climate, energy use, and the total costs of suboptimal climate policy. These are obviously important as well, so we need to proceed to this analysis. However, for computing what optimal policy is, straightforward application of the formula above works very well, and in some sense is all that is needed to optimally deal with climate change. To compute the optimal quantity restrictions is much more demanding, because then precisely all these additional assumptions are made, and to predict the future of technology (especially that regarding energy supply) is extremely difficult, to say the least. Section 5.2.3 calibrates the key parameters behind the formula above and Section 5.2.4 then displays the numerical results for the social cost of carbon.

[cs] Learning (about γ or the natural-science parameters) could also be introduced formally, as in the planning problem studied by Kelly and Kolstad (1999).

5.2 A Positive Dynamic Model

The positive dynamic model will be a straightforward extension of the static model in Section 4 in combination with the basic model from Section 2.3.2 (without endogenous technical change).

Thus we assume a production function that is Cobb–Douglas in capital, labor, and an energy input, along with TFP damages from climate:

$$Y_t = e^{-\gamma_t S_t} A_t K_t^{\alpha} N_{0t}^{1-\alpha-\nu} E_t^{\nu}. \tag{28}$$

Here, we maintain the possibility that γ changes over time/is random.

There are three energy-producing sectors, as in one of the extensions of the static model. Sector 1 thus produces "oil," which is in finite supply and is extracted at zero cost. The accounting equation $E_{ot} = R_t - R_{t+1}$ thus holds for oil stocks at all times. The second and third sectors are the "coal" and the "green" sectors, respectively. They deliver energy using

$$E_{i,t} = \chi_{it} N_{it} \text{ for } i = c, g. \tag{29}$$

Here, $N_t = N_{0t} + N_{ct} + N_{gt}$. We will focus on parameters such that coal, though in finite supply, will not be used up; hence, its Hotelling premium will be zero and there will be no need to keep track of the evolution of the coal stock.[ct] This specification captures the key stylized features of the different energy sectors while maintaining tractability. In practice, oil (as well as natural gas) can be transformed into useable energy quite easily but these resources are in very limited supply compared to coal. Coal is also more expensive to produce, as is green energy.

Here, energy used in production of the final good, E_t, then obeys

$$E_t = \left(\kappa_o E_{ot}^{\rho} + \kappa_c E_{ct}^{\rho} + \kappa_g E_{gt}^{\rho}\right)^{1/\rho} \tag{30}$$

with $\sum_{i=o,c,g} \kappa_i = 1$. As before, $\rho < 1$ regulates the elasticity of substitution between different energy sources; the κs are share parameters and also influence the efficiency with with the different energy sources are used in production. In addition, coal is "dirtier" than oil in that it gives rise to higher carbon emissions per energy unit produced. With E_{ot} and E_{ct} in the same units (of carbon emitted), the calibration therefore demands $\kappa_o > \kappa_c$.

The variables A_t, χ_{it}, and N_t are assumed to be exogenous and deterministic. Population growth is possible within our analytically tractable framework but we abstract from considering it explicitly in our quantitative exercises below, since A and N play the same

[ct] This will, under some specifications, require that a *back-stop technology* emerge at a point in the future, ie, a technology that simply replaces coal perfectly at lower cost.

role.[cu] Our final assumption, which is key for tractability, is that capital depreciates fully between periods ($\delta = 1$). This is an inappropriate assumption in business-cycle analysis but much less so when a model focusing on long-run issues; a model period will be calibrated to be 10 years.

5.2.1 Solving the Planner's Problem

For brevity, we do not state the planner's problem; it is implicit from the description earlier. The first-order conditions for C_t and K_t yield

$$\frac{1}{C_t} = \beta \mathbb{E}_t \frac{\alpha}{C_{t+1}} \frac{Y_{t+1}}{K_{t+1}}.$$

Together with the resource constraint

$$C_t + K_{t+1} = Y_t$$

we then obtain an analytical solution for saving as $K_{t+1} = \alpha\beta Y_t$ for all t. It follows that C_t/Y_t is equal to $1-\alpha\beta$ at all times, and we have therefore demonstrated that Assumption 4 is verified for this economy. A byproduct of our assumptions here, then, are that the formula for the optimal carbon tax, Eq. (26), holds exactly.

What is the planner's choice for the energy inputs, and what is the resulting effect on atmospheric carbon concentration and, hence, the climate? First, we assume that $\rho < 1$, and from this Inada property we then conclude that the energy choices will be interior at all times. Looking at the first-order conditions for E_t and E_{ot}, we obtain

$$\frac{\nu\kappa_o}{E_{ot}^{1-\rho} E_t^{\rho}} - \frac{\mathrm{SSC}_t}{Y_t} = \beta \mathbb{E}_t \left(\frac{\nu\kappa_o}{E_{o,t+1}^{1-\rho} E_{t+1}^{\rho}} - \frac{\mathrm{SSC}_{t+1}}{Y_{t+1}} \right), \tag{31}$$

where SSC_t/Y_t is, again, defined Eq. (26). This equation expresses Hotelling's formula in the case where there is a cost of using carbon: the damage externality (thus, playing a similar role to an extraction cost).

Looking at the other two energy source, by choosing $N_{i,\ t}$ optimally we obtain

$$\chi_{ct} \left(\frac{\nu\kappa_c}{E_{ct}^{1-\rho} E_t^{\rho}} - \frac{\mathrm{SCC}_t}{Y_t} \right) = \frac{1-\alpha-\nu}{N_t - \dfrac{E_{ct}}{\chi_{ct}} - \dfrac{E_{gt}}{\chi_{gt}}} \tag{32}$$

and

[cu] We formulate the utility function in terms of total consumption, and we do not adjust discounting for population growth. One might want to consider an alternative here, but we suspect that nothing substantial will change with this alternative.

$$\chi_{gt}\frac{\nu\kappa_g}{E_{gt}^{1-\rho}E_t^{\rho}}=\frac{1-\alpha-\nu}{N_t-\dfrac{E_{ct}}{\chi_{ct}}-\dfrac{E_{gt}}{\chi_{gt}}}. \tag{33}$$

From the perspective of solving the model conveniently, it is important to note now that SSC_t/Y_t is available in closed form as a function of primitives: the remaining system of equations to be solved is a vector difference equation but only in the energy choices. Ie, the model can be solved for energy inputs first, by solving this difference equation, and then the rest of the variables (output, consumption, etc.) are available in the simple closed forms given above.

To solve the vector difference equation—to the extent there is no uncertainty—is also simple, though in general a small amount of numerical work is needed.[cv] A robust numerical method goes as follows. With any given value for E_{ot}, the Eqs. (32) and (33) can be used to solve for E_{ct} and E_{gt}, and thus E_t. The solution is nonlinear but well defined. For any given initial stock of oil R_0, one can now use a simple shooting algorithm. The "shooting" part is accomplished by (i) guessing on a number for E_{o0}; (ii) deriving the all the other energy inputs at time 0; (iii) using the Hotelling Eq. (31), which is stated in terms of E_{o1} and E_1, to obtain E_{o1} as a function of E_1; (iv) combining this relation between E_{o1} and E_1 with Eqs. (32) and (33) evaluated for period 1 to obtain all the energy choices in period 1; and (v) going back to step (iii) to repeat for the next period. The so-obtained path for all energy inputs in particular delivers a path for oil extraction. To check whether the fired shot hits the target involves simply checking that the cumulated oil use exactly exhausts the initial stock asymptotically. If too much or too little is used up, adjust E_{o0} appropriately and run through the algorithm again.

If there is uncertainty about γ that is nontrivial and does not go away over time, one needs to use recursive methods, given the nonlinearity of the vector difference equation. It is still straightforward to solve, however, with standard versions of such methods.

5.2.2 Competitive Equilibrium

It is straightforward to define a dynamic (stochastic) general equilibrium for this economy as for the static model. All markets feature perfect competition. Firms in the final-goods sector make zero profits, as do firms in the coal and green-energy sectors. In the oil sector, there is a Hotelling rent, and hence profits. These profits are delivered to the representative consumer, who otherwise receive labor and capital income and, to the extent there is a tax on fossil fuel, lump-sum transfers so that the government budget balances. When taxes are used, we assume that they are levied on the energy-producing firms (oil and coal). The consumer's Euler equation and the return to capital satisfying the first-order condition for capital from the firm's problem deliver the constant saving rate $\alpha\beta$. The energy supplies (or, equivalently, the labor allocation) is then given by a set of

[cv] Solving the model with only coal or only green energy is possible in closed form.

conditions similar to those from the planning problem. Assuming that the carbon tax in period t is set as an exogenous fraction of output in period t, we then obtain from the energy producers' problems

$$\frac{\nu\kappa_o}{E_{ot}^{1-\rho}E_t^{\rho}}-\tau_t=\beta\mathbb{E}_t\left(\frac{\nu\kappa_o}{E_{o,t+1}^{1-\rho}E_{t+1}^{\rho}}-\tau_{t+1}\right), \tag{34}$$

$$\chi_{ct}\left(\frac{\nu\kappa_c}{E_{ct}^{1-\rho}E_t^{\rho}}-\tau\right)=\frac{1-\alpha-\nu}{N_t-\dfrac{E_{ct}}{\chi_{ct}}-\dfrac{E_{gt}}{\chi_{gt}}}, \tag{35}$$

and

$$\chi_{gt}\frac{\nu\kappa_g}{E_{gt}^{1-\rho}E_t^{\rho}}=\frac{1-\alpha-\nu}{N_t-\dfrac{E_{ct}}{\chi_{ct}}-\dfrac{E_{gt}}{\chi_{gt}}}. \tag{36}$$

Since this vector difference equation is very similar to the planner's vector difference equation, it can be solved straightforwardly with the same kind of algorithm. The laissez-faire allocation is particularly simple to solve.

5.2.3 Calibration and Results

In the spirit of quantitative macroeconomic modeling, the calibration of our model parameters is critical. Also in this part, we follow Golosov et al. (2014) in selecting parameter values. The calibration is important to review in some detail here, as calibration of this class of models is not standard in the macroeconomic literature. Given our assumptions, two parameters are easy to select: we assume that α and ν are 0.3 and 0.04, respectively; the value for the capital share is standard in the macroeconomic literature and the energy share is taken from the calibration in Hassler et al. (2015).

5.2.3.1 Discounting

As will be clear from our results, the discount factor matters greatly for what optimal tax to recommend. We do not take stand here but rather report our results for a range of values for β. Nordhaus's calibrations start from interest-rate data; interest rates should mirror the interest rate, if markets work, so to set $1/\beta-1=0.015$ is then reasonable. Stern, in his review on climate change, takes a very different view and uses what is essentially a zero rate: $1/\beta-1=0.001$. A view that sharply differs from the market view can be motivated on purely normative grounds, though then there may be auxiliary implications of this normative view: perhaps capital accumulation should then be encouraged more broadly, eg, using broad investment/saving subsidies. Sterner and Persson (2008), however, argue informally that it is possible to discount consumption and climate services—to the extent the latter enter separately in utility—at different rates.

A third and, we think, interesting argument for using a lower discount rate is that it is reasonable to assume that discounting is time-inconsistent: people care about themselves and the next generation or so with rates in line with observed market rates but thereafter, they use virtually no discounting. The idea would be that I treat the consumption of my grand-grand-grand children and that of my grand-grand-grand-grand children identically in my own utility weighting. If this is a correct description of people's preferences, and if people have commitment tools for dealing with time inconsistency, we would see it in market rates, but there are not enough market observations for such long-horizon assets to guide a choice of discount rates. Hence, it is not easy to reject a rate such as 0.1% (but, by the same token, there is no market evidence in favor of it either). If people have no commitment tools for dealing with time inconsistency, observed market rates today would be a mix of the short- and long-run rates (and very heavily weighted toward present-bias), thus making it hard to use market observations to back out the longer-run rates. These arguments can be formalized: it turns out that the present model—if solved with a simplified energy sector (say, coal only)—can be solved analytically also with time-inconsistent preferences (see Karp, 2005, Gerlagh and Liski, 2012, and Iverson, 2014).

5.2.3.2 The carbon cycle

We calibrate the carbon cycle, as indicated, with a linear system implying that the carbon depreciation rates are given by Eq. (13). Thus with the depreciation rate at horizon j given by $1-d_j=\varphi_L+(1-\varphi_L)\varphi_0(1-\varphi)^j$, we have to select three parameter: φ_L, φ_0, and φ. Recall the interpretation that φ_L is the share of of carbon emitted into the atmosphere that stays there forever, $1-\varphi_0$ the share that disappears into the biosphere and the surface oceans within a decade, and the remaining part, $(1-\varphi_L)\varphi_0$, decays (slowly) at a geometric rate φ. We set φ_L to 0.2, given the estimate in the 2007 IPCC report that about 20% any emission pulse remains in the atmosphere for several thousand years.[cw] Archer (2005), furthermore, argues that the excess carbon that does depreciate has a mean lifetime of about 300 years. Thus, we set $(1-\varphi)^{30}=0.5$, implying $\varphi=0.0228$. Third, the 2007 IPCC report asserts that about 50% of any CO_2 emission pulse into the atmosphere has left the atmosphere after about 30 years. This means that $d_2=0.5$ so that $1-\frac{1}{2}=0.2+0.8\varphi_0(1-0.0228)^2$, and hence $\varphi_0=0.393$. Finally, to set the initial condition for carbon concentration we showed above that the assumed depreciation structure is consistent with the existence of two "virtual carbon stocks" S_1 (the part that remains in the atmosphere forever) and S_2 (the part that depreciates at rate φ), with $S_{1,t}=S_{1,t-1}+\varphi_L E_t^f$ and $S_{2,t}=\varphi S_{2,t-1}+\varphi_0(1-\varphi_L)E_t^f$, and $S_t=S_{1,\ t}+S_{2,\ t}$. We choose starting values so that time-0 (ie, year-2000) carbon equals 802, with the division

[cw] Archer (2005) argues for a slightly higher number: 0.25.

$S_1 = 684$ and $S_2 = 118$; the value of S_1 comes from taking the preindustrial stock of 581 and adding 20% of accumulation emissions.[cx]

5.2.3.3 Damages

Turning to the calibration of damages, recall that we argued that for a reasonable range of carbon concentration levels the exponential TFP expression $e^{-\gamma S}$ is a good approximation to the composed S-to-temperature and temperature-to-TFP mappings in the literature. It remains choose γ, deterministic or stochastic. Here, in our illustrations, we will focus on a deterministic γ and only comment on uncertainty later. Following the discussion in the damage section earlier and Golosov et al. (2014), with S measured in GtC (billions of tons of carbon), an exponential function with parameter $\gamma_t = 5.3 \times 10^{-5}$ fits the data well.

5.2.3.4 Energy

Turning, finally, to the energy sector, we first need to select a value for ρ, which guides the elasticity of substitution between the energy sources. Stern (2012) is a metastudy of 47 studies of interfuel substitution and reports the unweighted mean of the oil–coal, oil–electricity, and coal–electricity elasticities to be 0.95. Stern's account of estimates of "long-run dynamic elasticities" is 0.72. In terms of our ρ, the implied numbers are -0.058 and -0.390, respectively, and the former will constitute our benchmark.

As for the different energy sources, for oil we need to pin down the size of the oil reserve. According to BP (2010), the proven global reserves of oil are 181.7 gigaton. However, these figures only refer to reserves that are economically profitable to extract at current conditions. Rogner (1997), on the other hand, estimates the global reserves of potentially extractable oil, natural gas, and coal taken together to be over 5000 Gt, measured as oil equivalents.[cy] Of this amount, Rogner reports around 16% to be oil, ie, 800 Gt. We use a benchmark that is in between these two numbers: 300 Gt. To express fossil fuel in units of carbon content, we set the carbon content in crude oil to be 846 KgC/t oil. For coal, we set it to the carbon content of anthracite, which is 716 KgC/t coal.[cz] As for coal, as implied by Rogner's (1997) estimates, the coal supply is enough for several hundreds of years of consumption at current levels, and hence we have assumed the scarcity rent to be zero.

[cx] These number include the preindustrial stock and, hence, do not strictly follow the notation above, where S_t denotes the concentration in excess of preindustrial levels.

[cy] The difference in energy content between natural gas, oil, and various grades of coal is accounted for by expressing quantities in oil equivalents.

[cz] IPCC (2006, table 1.2–1.3).

To calibrate κ_o and κ_c we use relative prices of oil to coal and oil to renewable energy, given by

$$\frac{\kappa_o}{\kappa_c}\left(\frac{E_{ot}}{E_{ct}}\right)^{\rho-1} \quad \text{and} \quad \frac{\kappa_o}{1-\kappa_o-\kappa_c}\left(\frac{E_{ot}}{E_{gt}}\right)^{\rho-1},$$

respectively. The average price of Brent oil was \$70 per barrel over the period 2005–09 (BP, 2010); with a barrel measuring 7.33 metric tons and a carbon content of 84.6%, the oil price per ton of carbon is then \$606.5. As for coal, its average price over the same period is \$74 per ton. With coal's carbon content of 71.6%, this implies a price of \$103.35 per ton of carbon.[da] The implied relative price of oil and coal in units of carbon content is 5.87.

As for renewables/green energy, there is substantial heterogeneity between different such sources. With unity as a reasonable value of the current relative price between green energy and oil, we employ data on global energy consumption to finally pin down the κs. Primary global energy use in 2008 was 3.315 Gtoe (gigaton of oil equivalents) of coal, 4.059 of oil, 2.596 of gas, and $0.712+0.276+1.314=2.302$ of nuclear, hydro, and biomass/waste/other renewables. Based on the IPCC tables quoted earlier, the ratio of energy per ton between oil and anthracite is then $\frac{42.3}{26.7}=1.58$, implying that 1 t of oil equivalents is 1.58 t of coal.[db] With these numbers and the value for ρ of -0.058, we can finally use the equations above to back out $\kappa_o=0.5008$ and $\kappa_c=0.08916$.

The parameters χ_{ct}, which determines the cost of extracting coal over time, are set based on an average extraction cost of \$43 per ton of coal (see IEA, 2010, page 212). Thus, a ton of carbon in the form of coal costs \$43/0.716. The model specifies the cost of extracting a ton of carbon as $\frac{w_t}{\chi_{ct}}$, where w_t is the wage. The current shares of world labor used in coal extraction and green energy production is very close to zero, so with total labor supply normalized to unity we can approximate the wage to be $w_t=(1-\alpha-\nu)Y_t$. With world GDP at \$700 trillion per decade and a gigaton of carbon (our model unit) costing $w_t/\chi_{ct}=(1-\alpha-\nu)Y_t/\chi_{ct}$ to produce delivers $43\cdot 10^9/0.716=0.66\cdot 700\cdot 10^{12}/\chi_{c0}$ and hence $\chi_{c0}=7693$. This means, in other words, that a share $\frac{1}{7693}$ of the world's labor supply during a decade is needed to extract one gigaton of carbon in the form of coal. The calibration of χ_{g0} comes from using the fact that χ_{g0}/χ_{c0} equals the relative price between coal and green energy, thus delivering $\chi_{g0}=7693/5.87=1311$ since the prices

[da] BP (2010) gives these estimates for US Central Appalachian coal.

[db] The amounts of oil and coal in carbon units is obtained by multiplying by the carbon contents 84.6 and 71.6%, respectively.

of oil and green are assumed to be equal and the relative price of oil in terms of coal is 5.87. Lastly, we posit growth in both χ_{ct} and χ_{gt} at 2% per year.[dc]

5.2.4 Results

We begin by reporting what our model implies for the optimal tax on carbon. Given our calibration, and expressed as a function of the discount rate, we plot the tax per ton of emitted carbon in Fig. 14, given annual global output of 70 trillion dollars.[dd]

Fig. 14 displays our benchmark as a solid line along with two additional lines representing two alternative values for γ, the higher one of which represents a "catastrophe scenario" with losses amounting to about 30% of GDP and the lower one representing an opposite extreme case with very low losses. The numbers in the figure can be compared to the well-known proposals in Nordhaus and Boyer (2000) and in the Stern review (Stern, 2007), who suggest a tax of \$30 and \$250 dollar per ton of carbon, respectively. As already pointed out, these proposals are based on very different discount rates, with Nordhaus using 1.5% per year and Stern 0.1%. For these two discount-rate values, the optimal taxes using our analysis are \$56.9 per ton and \$496 per ton, respectively, thus showing larger damages than in these studies. There are a number of differences in assumptions between the model here and those maintained in, say, Nordhaus's work; perhaps the most important one quantitatively is that we calibrate the duration of carbon in the atmosphere to be significantly higher.

The figure reveals that, to the extent the catastrophe scenario—which comes from a hypothesis Nordhaus entertained in a survey study—might actually materialize, there will

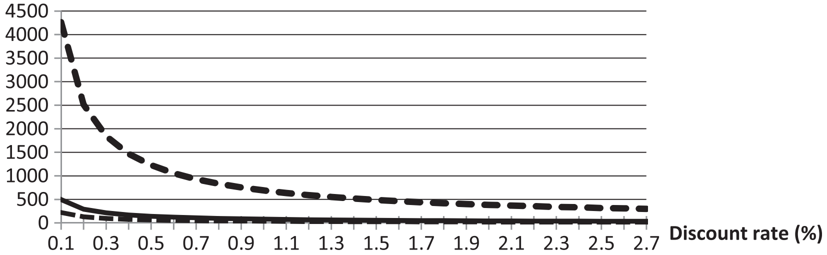


Fig. 14 Optimal tax rates in current dollars per ton of emitted fossil carbon vs yearly subjective discount rate.

[dc] Under our calibration, coal use does not go to zero, which contradicts it being a finite resource. Strictly speaking, one should instead, then, solve the model under this assumption and the implication that coal would have scarcity value. But we consider it quite likely that a competitive close and renewable substitute for coal is invented over the next couple of hundred years, in which case our solution would work well as an approximation.

[dd] The graphs are taken from Golosov et al. (2014).

be dramatic consequences on the level of the optimal tax: we see that the tax is roughly multiplied by a factor 20.

5.2.5 Positive Implications

Fossil fuel use in the optimal allocation and in the *laissez-faire* allocation are shown in Fig. 15. We base our results in this section on the discount rate 1.5%.

Looking at the comparison between the optimum and laissez faire, we see a markedly lower use of fossil fuel in the optimum.[de] In the laissez-faire scenario, there would be a continuous increase in fossil fuel use, but in the optimum the consumption of fossil fuel is virtually flat.

It is important to realize that the difference between the fossil-fuel use in the optimum and in laissez faire is almost entirely coming from a lower coal use in the former. In Figs. 16 and 17, we look separately at coal use and oil use in the optimal vs the laissez-faire allocations. Although the tax on carbon is identical for oil and coal in the optimal allocation, its effects are very different: coal use is simply curbed significantly—the whole path is shifted down radically—but oil use is simply moved forward slightly in time. With optimal taxes,

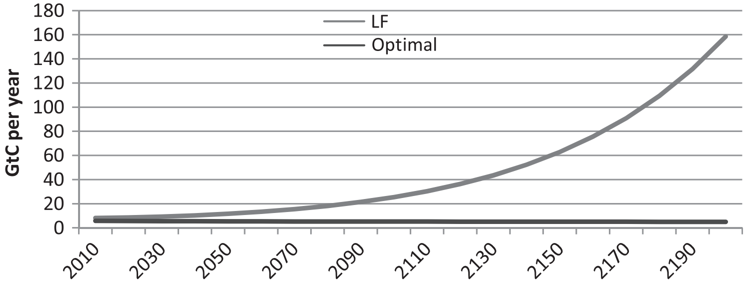


Fig. 15 Fossil fuel use: optimum vs laissez faire.

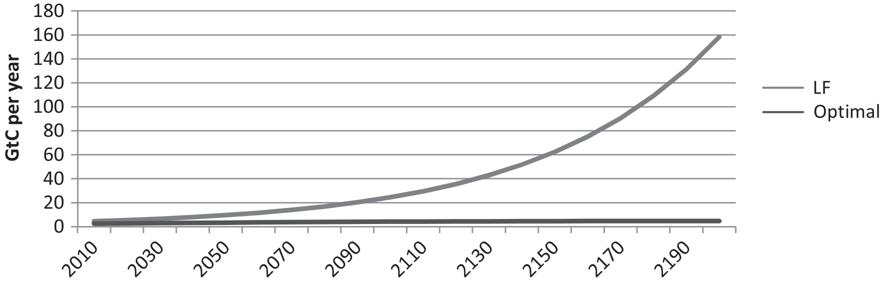


Fig. 16 Coal use: optimum vs laissez faire.

[de] The model predicts coal use in laissez faire of 4.5 GtC during the coming decade; it is currently roughly 3.8 GtC. It predicts oil use of 3.6 GtC, which is also close to the actual value for 2008 or 3.4 GtC.

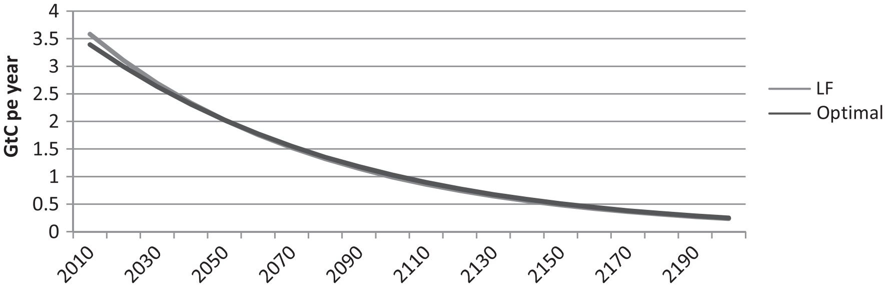


Fig. 17 Oil use: optimum vs laissez faire.

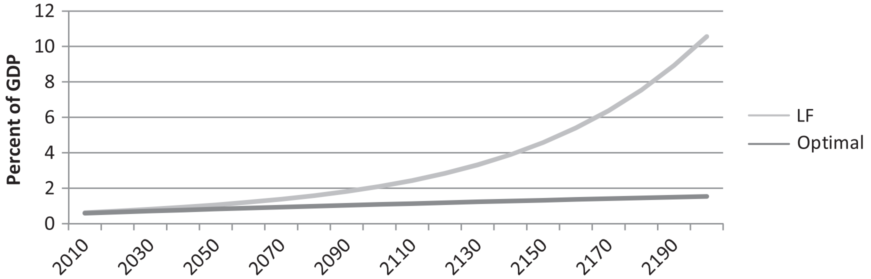


Fig. 18 Total damages as a percent of global GDP: optimum vs laissez faire.

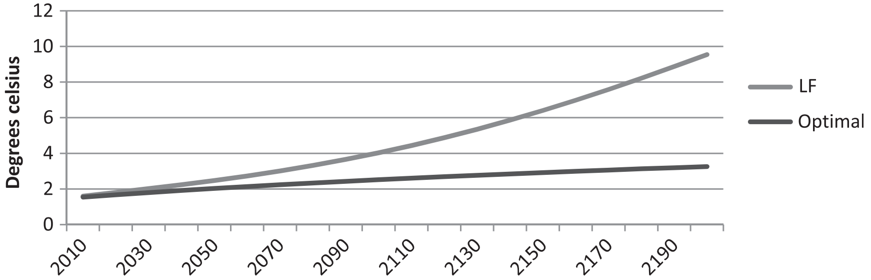


Fig. 19 Increases in global temperature: optimum vs laissez faire.

coal use would fall right now to almost half; a hundred years from now, laissez-faire coal use would be 7 × higher than optimally. Green energy use is very similar across the optimum and laissez-faire allocations.

Total damages are shown in Fig. 18. We note large, though not gigantic, gains from moving from laissez faire to the optimum allocation. The gains grow over time, with damages at a couple of percent of GDP in the laissez-faire allocation, thus about double its optimal value at that time. In 2200, the difference is a factor of six.

We can also back out the path for global temperature in the two scenarios, using the known mapping from S to temperature. Fig. 19 illustrates that laissez faire is associated

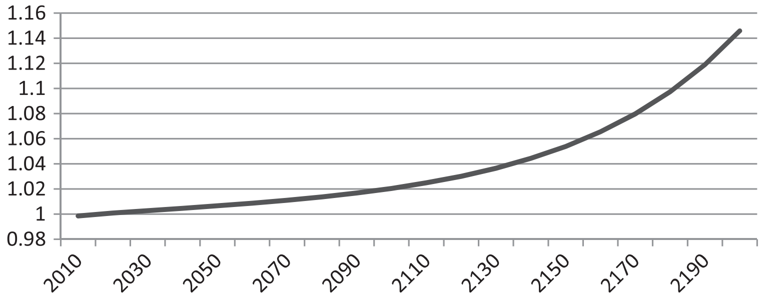


Fig. 20 Net output: optimum vs laissez faire.

with a temperature rise of 4.4°C a hundred years from now; in the optimum, heating is only 2.6 degrees. Toward the end of the simulation period, however, due to massive coal use, laissez faire predicts increased heating by almost 10°C; the optimum dictates about 3 degrees.

Finally, Fig. 20 displays the evolution of the (net-of-damage) production of final-good output (GDP). The intertemporal trade-off is clear here, but not as striking as one might have guessed: the optimal allocation involves rather limited short-run losses in GDP, with optimal output exceeding that of laissez faire as early as 2020. 100 years later, GDP net of damages is 2.5% higher in the optimum and in year 2200, it is higher by almost 15%.

5.2.6 Discussion

How robust are the quantitative results in Section 5.2.4? First, the tax formula appears remarkably robust. The point that only three kinds of parameters show up in the formula is a robustness measure in itself; eg, no details of the fossil-fuel stocks, production technologies, or population matter. Strictly speaking, these features begin mattering once one or more of the main assumptions behind the formula are not met, but they will only matter indirectly, eg, insofar as they influence the consumption-output path, and if their impact here is minor, the formula will be robust. In a technical appendix to the Golosov et al. (2014) paper, Barrage (2014) considers a version of the model where not all of the assumptions are met. In particular, this version of the model has more standard transitional dynamics (with a calibration in line with the macroeconomic literature). For example, the assumption that the consumption-output ratio is constant will not hold exactly along a transition path, but the departures almost do not change the results at all. Also, at least US data show very minor fluctuations in this ratio so to the extent a model delivers more drastic movements in the consumption-output ratio it will have trouble matching the data. Higher curvature in utility also delivers very minor changes in the tax rate, with the correction that discounting now involves not just β but also the consumption growth rate raised to $1-\sigma$, where $\sigma=1$ gives logarithmic curvature and $\sigma>1$ higher curvature.

Second, when it comes to the positive analysis—eg, the implications for temperature and damages under different policy scenarios—the message is quite different: many of the assumptions can matter greatly for the quantitative results. Perhaps the best example of nonrobustness is the example considered in Golosov et al. (2014): the elasticity of substitution between energy sources was raised by setting $\rho=0.5$, ie, assuming an elasticity of 2 instead of one slightly below one. If the different energy sources are highly substitutable, coal can easily be used instead of oil, making the laissez-faire allocation deliver very high coal use. On the other hand, taxes are now more powerful in affecting the use of different energy sources. This means, in particular, that the difference in outcomes between an optimal tax and laissez-faire is very large compared to the benchmark, where the different energy sources are less substitutable. Hence, the substitutability across energy sources is an example of an area where more work is needed. Relatedly, we expect that the modeling of technical change in this area—energy saving, as in Section 2.3.3 or making new energy resources available—will prove very important.

A number of straightforward extensions to the setting are also possible and, in part, they have been pursued by other researchers.[df] One is the inclusion of damages that involve growth effects; Dell et al. argue that such effects may be present.[dg] It is easy to introduce such damages to the present setting by letting the TFP term read $e^{-\gamma_l S+\gamma_g St}$, where γ_l regulates level effect of carbon concentration S, and γ_g the damages to the growth rate of output; the baseline model admits closed-form solution. As already pointed out, the baseline model can also accommodate time-inconsistent preferences rather easily.[dh]

Finally, the discussion of dynamic integrated assessment models here is based entirely on the simple baseline model in Golosov et al. (2014) not because it is the only model of this sort, or even the most satisfactory one in some overall sense; rather, this model has been chosen, first, because it is the model with the closest links to standard macroeconomic settings (with forward-looking consumers, dynamic competitive equilibrium with taxes, and so on). Second, the baseline model in Golosov et al. admits highly tractable analysis (with closed-form solutions) and hence is very well suited for illustrations; moreover, for the optimal carbon tax it gives a very robust formula that is also quantitatively adequate. The model is also useful for positive analysis but here it is important to point out that many other approaches can offer more realistic settings and, at least from some

[df] For example, Rezai and van der Ploeg (2014).

[dg] See Moyer et al. (2013).

[dh] Such cases have been discussed by Karp (2005) and, in settings closely related to the model here, Gerlagh and Liski (2012) and Iverson (2014) show that it is possible to analyze the case without commitment relatively straightforwardly; lack of commitment and Markov-perfect equilibria are otherwise quite difficult to characterize.

perspectives, do a better job at prediction. It would require a long survey to review the literature and such an endeavor is best left for another paper; perhaps the closest relative among ambitious, quantitative settings is the WITCH model, which also builds on forward-looking and, among other things, has a much more ambitiously specified energy sector.[di]

REFERENCES

Acemoglu, D., 2009. Introduction to Modern Economic Growth. Princeton University Press.

Acemoglu, D., Aghion, P., Bursztyn, L., Hémous, D., 2012. The environment and directed technical change. Am. Econ. Rev. 102 (1), 131–166.

Acemoglu, D., Aghion, P., Hémous, D., 2014. The environment and directed technical change in a north-south model. Oxf. Rev. Econ. Policy 30 (3), 513–530.

Aghion, P., Dechezlepretre, A., Hémous, D., Martin, R., Van Reenen, J., 2014. Carbon taxes, path dependency and directed technical change: evidence from the auto industry. J. Polit. Econ. (forthcoming).

Aghion, P., Howitt, P., 1992. A model of growth through creative destruction. Econometrica 60 (2), 323–351.

Aghion, P., Howitt, P., 2008. The Economics of Growth. MIT Press.

Aiyagari, R., 1994. Uninsured idiosyncratic risk and aggregate saving. Q. J. Econ. 109 (3), 659–684.

Amador, M., 2003. A political model of sovereign debt repayment. Mimeo.

Archer, D., 2005. The fate of fossil fuel CO_2 in geologic time. J. Geophys. Res. 110.

Archer, D., Eby, M., Brovkin, V., Ridgwell, A., Cao, L., Mikolajewicz, U., Tokos, K., 2009. Atmospheric lifetime of fossil fuel carbon dioxide. Annu. Rev. Earth Planet. Sci. 37, 117–134.

Arrhenius, S., 1896. On the influence of carbonic acid in the air upon the temperature of the ground. Philos. Mag. J. Sci. 41 (5), 237–276.

Azzimonti, M., 2011. Barriers to investment in polarized societies. Am. Econ. Rev. 101 (5), 2182–2204.

Bansal, R., Ochoa, M., 2011. Welfare costs of long-run temperature shifts. NBER Working Paper 17574.

Barrage, L., 2014. Sensitivity analysis for Golosov, Hassler, Krusell, and Tsyvinski (2014): 'optimal taxes on fossil fuel in general equilibrium'. Econometrica. 82. http://www.econometricsociety.org/ecta/supmat/10217_extensions.pdf.

Barrage, L., 2015. Optimal dynamic carbon taxes in a climate-economy model with distortionary fiscal policy. Mimeo.

Barro, R.J., 2013. Environmental protection, rare disasters, and discount rates. NBER Working Paper 19258.

Berndt, E.R., Christensen, L.R., 1973. The translog function and the substitution of equipment, structures, and labor in U.S. manufacturing 1929-68. J. Econom. 1 (1), 81–113.

Bosetti, V., Carraro, C., Galeotti, M., Massetti, E., Tavoni, M., 2006. Witch–a world induced technical change hybrid model. Energy J. 27, 13–37.

Bovenberg, L., Smulders, S., 1995. Environmental quality and pollution-augmenting technological change in a two-sector endogenous growth model. J. Public Econ 57 (3), 369–391.

BP, 2010, 2015. BP statistical review of world energy. http://bp.com/statisticalreview.

Brock, W., Engström, G., Xepapadeas, A., 2014. Spatial climate-economic models in the design of optimal climate policies across locations. Eur. Econ. Rev. 69, 78–103.

Burke, M., Miguel, E., Satyanath, S., Dykema, J.A., Lobell, D.B., 2009. Warming increases the risk of civil war in Africa. Proc. Natl. Acad. Sci. 106 (49), 20670–20674.

Burke, M., Hsiang, S.M., Miguel, E., 2015. Climate and conflict. Ann. Rev. Econ. 7, 577–617.

Chris, H., Anderson, J., Wenman, P., 1993. Policy analysis of the greenhouse effect: an application of the PAGE model. Energy Pol. 21, 327–338.

[di] See Bosetti et al. (2006).

Ciscar, J.C., Iglesias, A., Feyen, L., Szabó, L., Van Regemorter, D., Amelung, B., Nicholls, R., Watkiss, P., Christensen, O.B., Dankers, R., Garrote, L., Goodess, C.M., Hunt, A., Moreno, A., Richards, J., Soria, A., 2011. Physical and economic consequences of climate change in Europe. Proc. Natl. Acad. Sci. 108 (7), 2678–2683.

Cline, W.R., 1992. Economics of Global Warming. Institute for International Economics.

Crost, B., Traeger, C.P., 2014. Optimal CO_2 mitigation under damage risk valuation. Nat. Clim. Change 4, 631–636.

Cuddington, J., Nülle, G., 2014. Variable long-term trends in mineral prices: the ongoing tug-of-war between exploration, depletion, and technological change. J. Int. Money Fin. 42 (C), 224–252.

Dasgupta, P., Heal, G., 1974. The optimal depletion of exhaustable resources. Rev. Econ. Stud 41, 3–28.

Dell, M., Jones, B.F., Olken, B.A., 2012. Temperature shocks and economic growth: evidence from the last half century. Am. Econ. J.: Macroecon. 4 (3), 66–95.

Dell, M., Jones, B., Olken, B., 2014. What do we learn from the weather? The new climate-economy literature. J. Econ. Lit. 52 (3), 740–798.

Desmet, K., Rossi-Hansberg, E., 2015. On the spatial economic impact of global warming. J. Urban Econ. 88, 16–37.

Drijfhouta, S., Bathiany, S., Beaulieu, C., Brovkin, V., Claussen, M., Huntingford, C., Scheffer, M., Sgubin, G., Swingedouw, D., 2015. Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models. Proc. Natl. Acad. Sci. USA. 112(43).

Ellerman, A.D., Buchner, B.K., 2007. The European Union emissions trading scheme: origins, allocation, and early results. Rev. Environ. Econ. Policy 1 (1), 66–87.

Erten, B., Ocampo, J.A., 2012. Super-cycles of commodity prices since the mid-nineteenth century. DESA Working Paper No. 110.

Fankhauser, S., 1994. The economic costs of global warming damage: a survey. Glob. Environ. Chang. 4, 301–309.

Feng, S., Krueger, A.B., Oppenheimer, M., 2010. Linkages among climate change, crop yields and Mexico-US cross-border migration. Proc. Natl. Acad. Sci. USA 107 (32), 14257–14262.

Gars, J., 2012. Essays on the macroeconomics of climate change. Ph.D. thesis. IIES Monograph series No. 74, Institute for International Economic Studies, Stockholm University.

Gerlagh, R., Liski, M., 2012. Carbon prices for the next thousand years. CESifo Working Paper Series No. 3855.

Geweke, J., 2001. A note on some limitations of CRRA utility. Econ. Lett. 71, 341–345.

Gollier, C., 2013. Evaluation of long-dated investments under uncertain growth trend, volatility and catastrophes. Toulouse School of Economics, TSE, Working Papers 12-361.

Golosov, M., Hassler, J., Krusell, P., Tsyvinski, A., 2014. Optimal taxes on fossil fuel in equilibrium. Econometrica 82 (1), 41–88.

Grossman, G., Krueger, A., 1991. Environmental impacts of a North American free trade agreement. NBER Working Paper 3914.

Harari, M., La Ferrara, E, 2014. Conflict, climate and cells: a disaggregated analysis. IGIER Working Paper.

Hart, R., 2013. Directed technical change and factor shares. Econ. Lett. 119 (1), 77–80.

Harvey, D., Kellard, N., Madsen, J., Wohar, M., 2010. The Prebisch-Singer hypothesis: four centuries of evidence. Rev. Econ. Stat. 92 (2), 367–377.

Hassler, J., Krusell, P., 2012. Economics and climate change: integrated assessment in a multi-region world. J. Eur. Econ. Assoc. 10 (5), 974–1000.

Hassler, J., Krusell, P., 2014. The climate and the economy. Mistra-SWECIA Nr. 5.

Hassler, J., Krusell, P., Olovsson, C., 2015. Will we need another mad max? Or will energy-saving technical change save us? Working Paper.

Hassler, J., Krusell, P., Nycander, J., 2016. Climate policy. Econ. Policy. (forthcoming).

Hémous, D., 2013. Environmental policy and directed technical change in a global economy: the dynamic impact of unilateral environmental policies? Working Paper.

Hillebrand, E., Hillebrand, M., 2016. Optimal climate policies in a dynamic multi-country equilibrium model. Working Paper.

Hotelling, H., 1931. Economics of exhaustible resources. J. Polit. Econ. 39 (2), 137–175.

IEA (International Energy Agency), 2010. World Energy Outlook. OECD/IEA.

IPCC, 2006. 2006 IPCC guidelines for national greenhouse gas inventories. Vol. 2 energy. In: Eggleston, S., Buendia, L., Miwa, K., Ngara, T., Tanabe, K., (Eds.), IPCC National Greenhouse Inventories Programme.

IPCC, 2007a. Climate change 2007 the physical science basis. In: Solomon, S., Qin, D., Manning, M., Marquis, M., Averyt, K., Tignor, M.M.B., Miller Jr., H.L.R., Chen, Z. (Eds.), Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

IPCC, 2007b. Climate change 2007 impacts, adaptation and vulnerability. In: Parry, M., Canziani, O., Palutikof, J., van der Linden, P., Hanson, C. (Eds.), Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

IPCC, 2013. Climate change 2013 the physical science basis. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M.M.B., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M., (Eds.), Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Iverson, T., 2014. Optimal carbon taxes with non-constant time preference. Working Paper.

Jia, R., 2014. Weather shocks, sweet potatoes and peasant revolts in historical China. Econ. J. 124 (575), 92–118.

Jones, C.I., 2001. Introduction to Economic Growth. W.W. Norton.

Jorgenson, D.W., Cao, J., Ho, M.S., 2013a. The economics of environmental policies in China. In: Clearer Skies Over China. MIT Press.

Jorgenson, D.W., Goettle, R.J., Ho, M.S., Wilcoxen, P.J., 2013b. Double Dividend. MIT Press.

Kam, K., et al., 2011. Partial radiogenic heat model for earth revealed by geoneutrino measurements. Nat. Geosci. 4 (9), 647–651.

Karp, L., 2005. Global warming and hyperbolic discounting. J. Public Econ. 89 (2), 261–282.

Kelly, D., Kolstad, C., 1999. Bayesian learning, growth, and pollution. J. Econ. Dyn. Control 23 (4), 491–518.

Krautkraemer, J.A., 1998. Nonrenewable resource scarcity. J. Econ. Lit. 36 (4), 2065–2107.

Krusell, P., Smith, A., 1998. Income and wealth heterogeneity in the macroeconomy. J. Polit. Econ. 106, 867–896.

Krusell, P., Smith, A., 2014. A global economy-climate model with high regional resolution. Working Paper.

Krusell, P., Smith, A., 2015. Climate change around the world. Working Paper.

Leach, A., 2007. The welfare implications of climate change policy. J. Econ. Dyn. Control. 57.

Lemoine, D., 2015. The climate risk premium: how uncertainty affects the social cost of carbon. Working Paper.

Lenton, T.M., Held, H., Kriegler, E., Hall, J.W., Lucht, W., Rahmstorf, S., Schnellnhuber, H.J., 2008. Tipping elements in the earth's climate system. Proc. Natl. Acad. Sci. USA 105 (6), 1786–1793.

Levitan, D., 2013. Quick-change planet: do global climate tipping points exist? Sci. Am. 25.

Lucas Jr., R.E., 1988. On the mechanics of economic development. J. Monet. Econ. 22, 3–42.

Manne, A., Mendelsohn, R., Richels, R., 1995. MERGE: a model for evaluating regional and global effects of GHG reduction policies. Energy Policy 23 (1), 17–34.

Matthews, H.D., Gillet, N.P., Stott, P.A., Zickfeld, K., 2009. The proportionality of global warming to cumulative carbon emissions. Nature 459, 829–833.

Matthews, H.D., Solomon, S., Pierrehumbert, R., 2012. Cumulative carbon as a policy framework for achieving climate stabilization. Philos. Trans. A Math. Phys. Eng. Sci. 370, 4365–4379.

McGlade, C., Ekins, P., 2015. The geographical distribution of fossil fuel unused when limiting global warming to 2°C. Am. Econ. Rev. 517, 187–190.

Mendelsohn, R., Nordhaus, W., Shaw, D.G., 1994. The impact of global warming on agriculture: a Ricardian approach. Am. Econ. Rev. 84 (4), 753–771.

Miguel, E., Satyanath, S., Sergenti, E., 2004. Economic shocks and civil conflict: an instrumental variables approach. J. Polit. Econ. 112 (4), 725–753.

Moyer, E.J., Woolley, M.D., Glotter, M.J., Weisbach, D.A., 2013. Climate impacts on economic growth as drivers of uncertainty in the social cost of carbon. Working Paper.

Nordhaus, W.D., 1973. World dynamics: measurement without data. Econ. J. 83(332).

Nordhaus, W.D., 1974. Resources as a constraint on growth. Am. Econ. Rev. 64(2).

Nordhaus, W.D., 1977. Economic growth and climate: the carbon dioxide problem. Am. Econ. Rev. Pap. Proc. 67 (1), 341–346.

Nordhaus, W.D., 1991. Economic approaches to greenhouse warming. In: Dornbush, R.D., Poterba, J.M. (Eds.), Global warming: Economic policy approaches. MIT Press, Cambridge, MA, pp. 33–68.

Nordhaus, W.D., 1992. An optimal transition path for controlling greenhouse gases. Science 258, 1315–1319.

Nordhaus, W.D., 1993. Rolling the 'DICE': an optimal transition path for controlling greenhouse gases. Resour. Energy Econ. 15, 27–50.

Nordhaus, W.D., 2006. Geography and macroeconomics: new data and new findings. Proc. Natl. Acad. Sci. USA 103 (10), 3510–3517.

Nordhaus, W.D., 2007. To tax or not to tax: the case for a carbon tax. Rev. Environ. Econ. Policy 1 (1), 26–44.

Nordhaus, W.D., 2009. An analysis of the dismal theorem. Working Paper.

Nordhaus, W.D., Boyer, J., 2000. Warming the World: Economic Modeling of Global Warming. MIT Press.

Nordhaus, W.D., Sztorc, P., 2013. DICE 2013R: introduction and user's manual. Mimeo, Yale University.

Otto, A., Otto, F.E.L., Allen, M.R., Boucher, O., Church, J., Hegerl, G., Forster, P.M., Gillett, N.P., Gregory, J., Johnson, G.C., Knutti, R., Lohmann, U., Lewis, N., Marotzke, J., Stevens, B., Myhre, G., Shindell, D., 2013. Energy budget constraints on climate response. Nat. Geosci. 6 (6), 415–416.

Pigou, A., 1920. Economics of Welfare. MacMillan.

Pindyck, R.S., 1978. The optimal exploration and production of nonrenewable resources. J. Polit. Econ. 86 (5), 841–861.

Poole, W., 1970. Optimal choice of policy instruments in a simple stochastic macro model. Q. J. Econ. 84 (2), 197–216.

Prather, M.J., Holmes, C.D., Hsu, J., 2012. Reactive greenhouse gas scenarios: systematic exploration of uncertainties and the role of atmospheric chemistry. Geophys. Res. Lett. 39, 9.

Prebisch, R., 1962. The economic development of Latin America and its possible problems. Econ. Bull. Latin Am. 7 (1), 1–22. Reprinted from: United Nations Department of Economic Affairs, Lake Success, NY (1950).

Revelle, R., Suess, H., 1957. Carbon dioxide exchange between atmosphere and ocean and the question of an increase of atmospheric CO_2 during past decades. Tellus 9, 18–27.

Rezai, A., van der Ploeg, F., 2014. Robustness of a simple rule for the social cost of carbon. Econ. Lett. 132, 48–55.

Rockström, J., Steffen, W., Noone, K., Persson, A., Chapin, F.S., Lambin, E.F., Lenton, T.M., Scheffer, M., Folke, C., Schellnhuber, H.J., Nykvist, B., de Wit, C.A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P.K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R.W., Fabry, V.J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., Foley, J.A., 2009. A safe operating space for humanity. Nature 461, 472–475.

Roe, G.H., Baker, M.B., 2007. Why is climate sensitivity so unpredictable. Science 318 (5850), 629–632.

Rogner, H.H., 1997. An assessment of world hydrocarbon resources. Ann. Rev. Energy Environ. 22, 217–262.

Romer, P.M., 1990. Endogenous technological change. J. Polit. Econ. 98 (5), S71–S102.

Schmitt, A., 2014. Beyond pigou: climate change mitigation, policy making and distortions. Ph.D. thesis, IIES Monograph series No. 85, Stockholm University.

Schwartz, S.E., Charlson, R.J., Kahn, R., Ogren, J., Rodhe, H., 2010. Why hasn't earth warmed as much as expected? J. Clim. 23.

Schwartz, S.E., Charlson, R.J., Kahn, R., Rodhe, H., 2014. Earth's climate sensitivity: apparent inconsistencies in recent assessments. Earth's Fut. 2.

Singer, H.W., 1950. U.S. foreign investment in underdeveloped areas: the distribution of gains between investing and borrowing countries. Am. Econ. Rev. Pap. Proc. 40, 473–485.

Sinn, H.W., 2008. Public policies against global warming: a supply side approach. Int. Tax Public Finance 15 (4), 360–394.

Solow, R., 1974. Intergenerational equity and exhaustible resources. Rev. Econ. Stud. 41, 29–45.

Spiro, D., 2014. Resource prices and planning horizons. J. Econ. Dyn. Control 48, 159–175.

Stern, D.I., 2012. Interfuel substitution: a meta-analysis. J. Econ. Surv. 26, 307–331.

Stern, N., 2007. The Economics of Climate Change: The Stern Review. Cambridge University Press.

Sterner, T., Persson, M., 2008. An even sterner review: introducing relative prices into the discounting debate. Rev. Env. Econ. Pol. 2 (1), 61–76.

Stiglitz, J., 1974. Growth with exhaustible natural resources: efficient and optimal growth paths. Rev. Econ. Stud. 41, 123–137.

Stokey, N., 1998. Are there limits to growth? Int. Econ. Rev. 39, 1–31.

Titus, J.G., 1992. The costs of climate change to the United States. In: Global Climate Change: Implications, Challenges and Mitigation Measures. Pennsylvania Academy of Science, pp. 384–409.

Tol, R.S.J., 1995. The damage costs of climate change toward more comprehensive calculations. Environ. Resour. Econ. 5, 353–374.

Tol, R.S.J., 2009. The economic effects of climate change. J. Econ. Perspect. 23 (2), 29–51.

Weitzman, M.L., 1974. Prices vs quantities. Rev. Econ. Stud. 41 (4), 477–491.

Weitzman, M.L., 1998. Why the far-distant future should be discounted at its lowest possible rate. J. Environ. Econ. Manage. 36 (3), 201–208.

Weitzman, M.L., 2009. On modeling and interpreting the economics of catastrophic climate change. Rev. Econ. Stat. 91, 1–19.

Weitzman, M.L., 2011. Fat-tailed uncertainty in the economics of catastrophic climate change. Rev. Environ. Econ. Policy 5 (2), 275–292.

CHAPTER 25

The Staying Power of Staggered Wage and Price Setting Models in Macroeconomics

J.B. Taylor
Stanford University, Stanford, CA, United States

Contents

Abstract

After many years, many critiques, and many variations, the staggered wage and price setting model is still the most common method of incorporating nominal rigidities into empirical macroeconomic models used for policy analysis. The aim of this chapter is to examine and reassess the staggered

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.04.008

wage and price setting model. The chapter updates and expands on my chapter in the 1999 *Handbook of Macroeconomics* which reviewed key papers that had already spawned a vast literature. It is meant to be both a survey and user-friendly exposition organized around a simple "canonical" model. It provides a guide to the recent explosion of microeconomic empirical research on wage and price setting, examines central controversies, and reassesses from a longer perspective the advantages and disadvantages of the model as it has been applied in practice. An important question for future research is whether staggered price and wage setting will continue to be the model of choice or whether it needs to be replaced by a new paradigm.

Keywords

Staggered contracts, Time-dependent pricing, State-dependent pricing, Contract multiplier, Calvo contracts, Taylor contracts, Hazard rate, Pass-through, Wage Dynamics Network, Nominal rigidities, New Keynesian economics

JEL Classification Codes

E3, E4, E5

1. INTRODUCTION

The staggered wage and price setting model has had remarkable staying power. Originating in the 1970s before the advent of real business cycle models, it has been the theory of choice in generation after generation of monetary business cycle models. In their review of over 60 macroeconomic models in their chapter for this Handbook, Wieland et al. (2016) define three such generations each with representative models that are based on staggered price or wage setting theories.[a]

This chapter examines the role of staggered wage and price setting as a method of incorporating nominal rigidities in empirical macroeconomic models used for policy analysis. It is both an exposition and a survey. It builds on my earlier *Handbook of Macroeconomics* chapter (Taylor, 1999) which reviewed original research papers that had already spawned a vast literature. It focuses on new research since that *Handbook* chapter, and, though it is largely self-contained, a more complete history of thought in this area requires looking at that chapter too. This chapter considers the explosion of microeconomic empirical research on wage and price setting behavior, the main critiques of the model, such as by Chari et al. (2000), and the complementary work on state-dependent pricing by Dotsey et al. (1999) and Golosov and Lucas (2007). Finally, the chapter reassesses from a longer vantage point the advantages and disadvantages of the model as it has been applied in practice, and it considers possible directions for future research.

[a] See Wieland et al. (2016), table 5.

2. AN UPDATED EMPIRICAL GUIDE TO WAGE AND PRICE SETTING IN MARKET ECONOMIES

I started off my 1999 *Handbook of Macroeconomics* chapter with "an empirical guide to wage and price setting in market economies" noting that "one of the great accomplishments of research on wage and price rigidities in the 1980s and 1990s is the bolstering of case studies and casual impression with the evidence from thousands of observations of price and wage setting collected at the firm, worker, or union level." The same could be said of the new research on microeconomic data during the past two decades except that there is much more of it—a virtual explosion of "Big Data" microeconomic studies, especially in the United States and European countries. These studies have confirmed much of the earlier work, but they have also uncovered new important facts about the timing, frequency, and determinants of price and wage change which are relevant for future research and model building. Accordingly, in this section I give an "*updated* empirical guide to wage and price setting in market economies."

As a starting point, recall that informal observation informed the original theoretical research on staggered wage and price setting models in the 1970s since there was virtually no microeconomic empirical research to guide it.[b] For many firms and organizations, whether in a formal employment contract or not, wages—including fringe benefits—appeared to be adjusted about once per year after a performance review and after consideration of prevailing wages in the market. A large fraction of the wage payment appeared to be a fixed amount, though overtime pay, bonuses, profit sharing, and piece rates were not uncommon, with as many similarities as differences between union and nonunion workers. Indexing of wages was seen to be rare in wage setting arrangements of 1 year or less. And wage adjustments looked to be unsynchronized—occurring at different times for different firms throughout the year—though there were exceptions such as the Shunto (spring wage offensive) in Japan.

Regarding prices, research work by Stigler and Kindahl (1970) had begun to document the extent of price rigidity for a wide variety of products and led people to distinguish informally between "auction markets" where prices changed continuously and "customer markets" where they changed infrequently, a terminology coined by Okun (1981). Though online purchasing has begun to blur this distinction, price changes, like wage changes, appeared to be unsynchronized and firms appeared to take the prevailing price of competing sellers into account.

Fortunately, a huge number of microeconomic studies of wage and price setting over the past few decades have given modelers much more to go on than informal observation. I first consider microeconomic empirical research on wage setting and then on price setting.

[b] I will describe the 1970s modeling research in the next section. Informal observation, of course, guided earlier models of price and wage adjustment, going way back to the time of Hume's (1742) classic essay "On Money" in which he wrote "by degrees the price rises, first of one commodity, then of another."

2.1 Microeconomic Evidence on Wage Setting

To my knowledge, the first empirical study to use actual microeconomic wage data to validate or calibrate the staggered wage setting models of the 1970s was my (1983) study using union wage contracting data in the United States. At the time, the Bureau of Labor Statistics had been calculating detailed data on major collective bargaining agreements for about 10 million workers in the United States and publishing the results in *Current Wage Developments*. The "major" contracts included agreements affecting 1000 or more workers. Although that sector represented only 10% of US employment, it was where the data were, and it was a place to begin.

The data indicated that wage setting was highly nonsynchronized, with agreements spread throughout the year though with relatively more settlements in the second and third quarters. Of these 10 million workers only about 15% had contract adjustments each quarter and only 40% each year. I used these micro data to calibrate a staggered wage setting model with heterogeneous contract lengths and simulated various monetary policies, and in a companion study (Taylor, 1982), I assumed that the remaining workers had shorter contracts. Looking at the union data over a period of time, Cecchetti (1984) found that the average period between wage changes declined with higher inflation, but was still more than 1 year during the high inflation period of the 1970s. There were few international comparisons at that time, though Fregert and Jonung (1986) found that wage setting in Sweden was unsynchronized and that contract length decreased with higher inflation, but it never dropped below 1 year on average.

There was then a lull in research on microeconomic wage setting practices, perhaps due to the increased interest in real business cycles and a corresponding "dark age" of research on wage and price rigidities, as I described in Taylor (2007). In any case, a gap was left between macroeconomic models of wage setting and the microeconomic evidence.

An explosion of research since the early 2000s (just after the completion of the *Handbook of Macroeconomics*, *Volume 1*!) has gone a long way to filling that gap. An important example, which has contributed greatly to our knowledge of micro wage setting, is the research enabled by the data collected from firms in a survey by the Wage Dynamics Network (WDN). The WDN was created after the founding of the European Central Bank; it consists of researchers at the central banks in the Eurosystem. The WDN surveyed wage and price setting practices at 17,000 European firms. The sample was designed to reflect firm employment size and sector distribution in each country. The survey covered both firms with employees in and out of unions. The percentage of employees in unions varies greatly across countries, ranging from over 70% in Scandinavian countries to less than 10% in Central and Eastern European countries, France, Spain, a percentage similar to the United States.

The report by Lamo and Smets (2009) summarizes the research on this survey referring to 81 different WDN papers and publications. They report that about 60% of the 17,000 firms surveyed change wages once a year, while 26% change wages less frequently.

The average duration of wages is about 15 months and is longer than the average duration of prices, which is about 9.5 months according to a parallel price setting survey in European countries.

Lamo and Smets (2009) also report "strong evidence of time dependence in wage-setting" with 55% of firms reporting that their wage changes occur in a particular month.[c] The timing of wage changes is characterized by a mix of staggering and synchronization. Indeed, there is a lot of heterogeneity across countries; the percentage of firms that change wages "more frequently than once a year ranges from 2.6% in Hungary and 4.2% in Italy to 33.9% in Greece, and 42.1% in Lithuania" according to Lamo and Smets.

There is also related time series work for specific European countries. Lünnemann and Wintr (2009), for example, examined monthly micro data from the Luxembourg social security authority. The data are reported by employers about their employees and pertain to the period from January 2001 to December 2006. They report that measurement error biases upwards the frequency of wage change, but adjusting for this measurement error they find a frequency of wage change of 9–14% per month, which is lower than for consumer prices at 17%. They also find a great deal of heterogeneity across forms. There is clear time dependence with many wages set around the month of January.

Le Bihan et al. (2012) examine a time series of French wage data. They use a quarterly panel of 38,000 French establishments with 6.8 million employees. They examine the base wage for 12 employee categories over 1998–2005. They argue that the base wage is a relevant indicator of wages in France because the base wage represents 77.9% of gross earnings. Furthermore, most bonuses (like "13th month" payments or holidays bonuses) constitute a fixed part of the earnings (5.2%) and are linked to the base wage. The frequency of quarterly wage change is around 38%, and in the case of France, there is not much cross-sectoral heterogeneity in wage stickiness.

They estimate a hazard function—the probability of a change in the wage conditional on an unchanged wage spell of a given duration. Their estimates of the hazard function are shown in Fig. 1. The authors state that the hazard function has a "noticeable spike at four quarters but is rather flat otherwise" and note that "such a pattern is consistent with the prevalence of Taylor-like, 1-year contracts."

Le Bihan et al. (2012) also estimate and report the frequency of wage change each quarter and the variation of that frequency over time. Their estimates are shown in Fig. 2 for all wages as well as for wages near the minimum wage. As they argue "there is evidence of a large degree of staggering since the frequency of wage changes is in no quarter lower than 20%." Note that there is some synchronization in the first quarter for all wages and in the third quarter for minimum wages, the later corresponding

[c] Some of the terminology used in this section—such as time dependence, state dependence, Taylor fixed-length contracts, Calvo model—is defined later in the chapter.

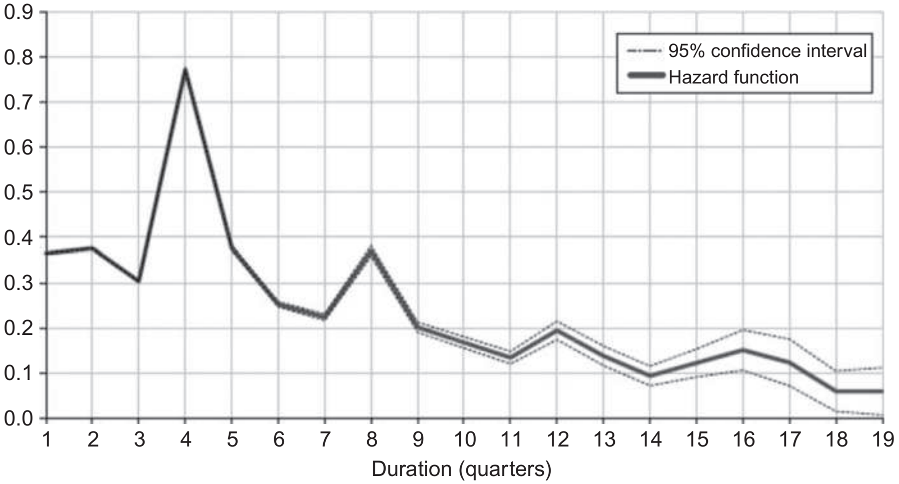


Fig. 1 Estimate of the hazard function of wage change in France. Source: *Le Bihan, H., Montornès, J., Heckel, T., 2012. Sticky wages: evidence from quarterly microeconomic data. Am. Econ. J. Macroecon. 4 (3), 1–32.*

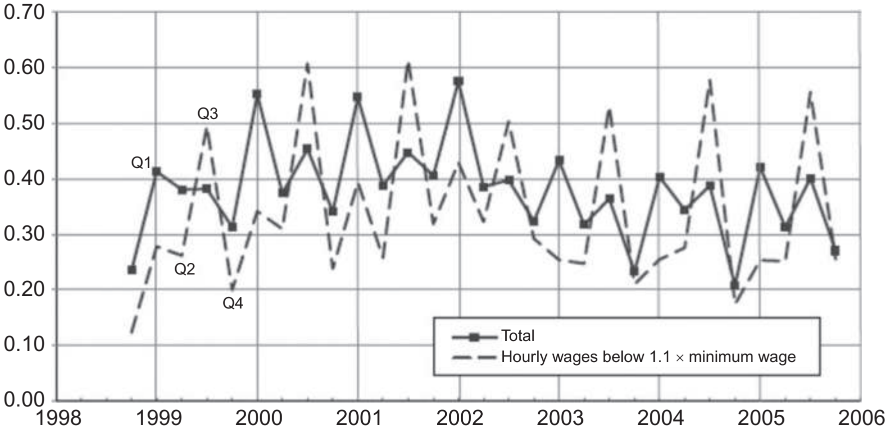


Fig. 2 Time variation in the frequency of wage change by quarter in France. Source: *Le Bihan, H., Montornès, J., Heckel, T., 2012. Sticky wages: evidence from quarterly microeconomic data. Am. Econ. J. Macroecon. 4 (3), 1–32.*

to the national minimum wage update in France each summer. They also report that their "micro-econometric evidence … suggests wage adjustment is mainly time dependent in France." And while wage changes are largely staggered across establishments, the authors report that there is a large degree of synchronization of wage changes within establishments.

Avouyi-Dovi et al. (2013) also examine the wage setting process in France. In contrast to Le Bihan et al. (2012), they collect and examine data on wage bargaining agreements, as Taylor (1983) did for the United States, but with much more detail. Their data pertain to both firms and industries. They find a sharp peak in the distribution of wage contract durations at 12 months. They also find that the "hazard rate shows a peak above 40% at twelve months and remains flat below 10% elsewhere." Indeed, their plots of the hazard function look like much like those in Figure 1 in this chapter with even more pronounced peaks. Finally, they find that the "wage change decisions are staggered over the year" with some evidence of seasonality that also shows up in the aggregate data. In many respects the findings Avouyi-Dovi et al. (2013) and those of Le Bihan et al. (2012) are very similar even though they use completely different data sets.

Another time series study is the paper by Sigurdsson and Sigurdardottir (2011) which examines wage setting behavior in Iceland. They use a micro wage dataset with a monthly frequency for the years 1998–2010. They find that average frequency of wage change is 10.8% per month. They find that "wage setting displays strong features of time dependence: half of all wage changes are synchronized in January, but other adjustments are staggered through the year" though later work by Sigurdsson and Sigurdardottir (2016), which focuses more on the global financial crisis, finds more evidence of state-dependent wage setting. The authors also estimate a hazard function and find that it has a large spike at 12 months. These facts indicate that, as the authors put it, "wage setting is consistent with the Taylor (1980) fixed duration contract model, but there exist contracts with both shorter and longer duration than precisely 1 year."

Recent work by Barattieri et al. (2014) has added important time series information about wage setting in the United States. They use high frequency panel data from the Survey of Income and Program Participation which follows people for a period of from 24 to 48 months with interviews every 4 months. The authors focus on hourly wage data (rather than salaries) which leaves them with a panel of 17,148 people from March 1996 to February 2000. The panel consisted of 49.4% women; ages ranged from 16 to 64 years and the average wage is $10.03 per hour. As with individual data reported by Lünnemann and Wintr (2009), the authors found a great deal of measurement error which adds noise to the wage series and effectively reduces the reported time that a wage is fixed. They corrected for this measurement error using structural break tests commonly used in time series analysis to look for big and persistent changes by filtering out smaller and more temporary changes.

They find that the quarterly frequency of wage adjustment, after correcting for measurement error, ranges from 12% to 27%, which is much lower than the 56% without correction for measurement error. They note that this corrected range is comparable to that found in the European studies reviewed earlier when reported on a common quarterly frequency:

Lünnemann and Wintr (2009)	19–36%
Le Bihan et al. (2012)	35%
Sigurdsson and Sigurdardottir (2011)	13–28%

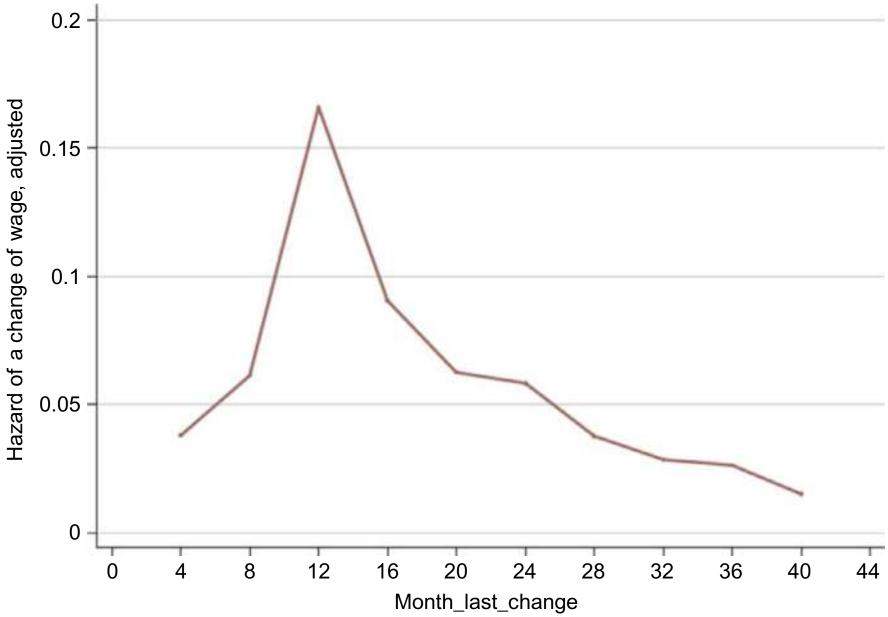


Fig. 3 Estimated hazard function for a within job wage change in the United States. Source: *Barattieri, A., Basu, S., Gottschalk, P., 2014. Some evidence on the importance of sticky wages. Am. Econ. J. Macroecon. 6 (1), 70–101.*

Finally, Barattieri et al. (2014) estimate a hazard function for the United States with their data corrected for measurement error. Their estimates are shown in Fig. 3. There is a sharp peak at 12 months leading the authors to conclude that "Taylor-type fixed-length contracts have stronger empirical support than Calvo-type constant-hazard models." This corresponds with the time series studies on wage setting in France and Iceland reported earlier.

If some structural assumptions about the general form of wage setting are made, it is also possible to extract information about individual wage setting mechanisms indirectly from the autocorrelation functions of aggregate time series data, as I explained in my chapter in the first *Handbook of Macroeconomics* with examples of these indirect methods including Backus (1984), Benabou and Bismut (1987), Levin (1991), and Taylor (1993). In a more recent example, Olivei and Tenreyro (2010) show that the impact of monetary policy shocks depends on the timing of wage changes, suggesting that time-dependent wage setting has important macroeconomic implications. They compare the effect of Japan's Shunto with different wage change timing in the United States and Germany, and they show that the impact of an aggregate monetary shock is larger when it occurs at a time when only a few wages are being adjusted. Estimates of time-varying distributions are also reported in Taylor (1993a) to accommodate the Shunto mechanism in Japan.

2.2 Microeconomic Evidence on Price Setting

Until the recent explosion of microeconomic research on price setting, the evidence on the prices of particular products showed remarkably long periods of set prices. Carlton (1989) found that the time between adjustment of prices ranged from 14 years for steel, cement, and chemicals to 4 years for plywood and nonferrous metals. Cecchetti (1986) found that the average length of time between price changes for magazines was 7 years in the 1950s and about 3 years in the 1970s. Kashyap (1995) found that mail order catalog prices were fixed for as long as 2 years. Blinder et al. (1998) found that about 40% of firms change their prices once per year, 10% change prices more frequently than once per year; and 50% leave their prices unchanged for more than a year. Dutta et al. (2002) found evidence of more frequent price changes for several types of frozen and refrigerated orange juice.

In contrast more recent detailed research by Bils and Klenow (2004), Klenow and Kryvtsov (2008), Nakamura and Steinsson (2008), and the ECB surveys in Europe shows more frequent changes in prices. A very useful review of this research is provided in a chapter in the *Handbook of Monetary Economics* by Klenow and Malin (2011), so there is no need to summarize it again here. They report that the average time between price changes is every 4 months for items in the consumer price index (CPI) and every 6–8 months for items in the producer price index. However, there is a great deal of heterogeneity across items with service prices changing less rapidly than good prices. They also report that price setting is unsynchronized, a finding that also goes back to Lach and Tsiddon (1996) who also noted within-store synchronization. Finally, Klenow and Malin (2011) emphasize that reference prices tend to be changed less frequently than regular prices.

As with wage setting, useful information about price setting in Europe comes from surveys of firms conducted by central banks. Fabiani et al. (2006) investigated the pricing behavior of more than 11,000 firms based on a survey conducted by the Eurosystem of national central banks. They found that "price reviews happen with a low frequency, of about one to three times per year in most countries, but prices are actually changed even less." They also found that "one-third of firms follow mainly time-dependent pricing rules, while two-thirds allow for elements of state dependence." The majority of the firms take into account both past and expected economic developments in their pricing decisions.

2.3 Pertinent Facts About Microeconomic Data on Wage and Price Setting

Though it is difficult to glean key facts from so many empirical studies, I would emphasize the following general features of price and wage setting as relevant to theoretical research on models of staggered wages and prices which I will review in the following sections:

(1) Both wage setting and price setting are staggered or unsynchronized over time. Even in unusual situations when there is a specific time of year for changing wages—such as in the spring in Japan and in January in some European counties, there are many other months where wages are changed. An example of evidence for staggered wage setting is that there was not one quarter where the frequency of wage change fell below 20% in France during the years from 1998 to 2006. Similarly, price changes are also typically not synchronized, as Klenow and Malin (2011) emphasize in their review.

(2) There is considerable evidence that most wages are set for a fixed length of time rather than changed at random intervals. The most common interval for wage changes is four quarters or 12 months. In Europe, the WDN survey shows that 60% of firms adjust wages once per year. Moreover, when it has been estimated, such as in France and the United States, the hazard function has a sharp peak at four quarters or 12 months.

(3) Wages and prices are set at a constant level during the length of time that they are set, rather than predetermined in advance to increase by certain amounts. Although originally clear from informal observation, this fact was confirmed for prices in empirical work by Klenow and Kryvtsov (2008) and Nakamura and Steinsson (2008). An exception in the case of wages occurs in the case of multiyear union contracts where deferred increases in later years are often agreed to in advance.

(4) There is strong evidence of time dependence in wage-setting and slightly less in price setting. Regarding wage setting, 55% of European firms report that wage changes occur in a particular month. In contrast, one-third of European firms follow mainly time-dependent pricing practices and two-thirds allow for elements of state dependence.

(5) Wage adjustment is less frequent than price adjustment, according to the most recent microeconomic empirical research, a finding which reverses the order reported in my 1999 *Handbook of Macroeconomics* chapter. In the European survey, the average duration of wages is greater than the average duration of prices. According to Barattieri et al. (2014), the quarterly frequency of wage adjustment in the United States, when correcting for measurement error, is much less than the CPI data as summarized by Klenow and Malin (2011). Price and wage rigidities are temporary, but prices and wages do not all change instantaneously and simultaneously, as if determined on a spot market with full information. There is no empirical reason—aside from the need for a simplifying assumption or the desire to illustrate a key point—to build an empirical model in which wages are perfectly flexible (determined on a spot market with full information) while prices are temporarily rigid, or vice versa.

(6) The frequency of wage and price changes depends on the average rate of inflation. While this is a robust finding, it should be emphasized that for the range of inflation rates observed in recent years in the developed economies, the average duration of wages and prices remains high. For a given target inflation rate, constant frequency of price adjustment is a good assumption to make in an empirical or policy model.

(7) There is a great deal of heterogeneity in wage and price setting practices across countries, across firms, across products, and across types of workers. Though the data reveal certain tendencies, as describe in the six points above, there is no practice that applies 100%. Wages in some industries change once per year on average, while in others wages change once per quarter or once every 2 years. There is a mixture of state dependence and time dependence in most countries. The price of services changes less frequently than goods. Wages of unskilled workers change more frequently than for skilled workers. One might hope that a model with homogeneous "representative" price or wage setting would be a good approximation to this more complex world, but models with some degree of heterogeneity are needed to describe reality accurately.

3. ORIGINS OF THE STAGGERED WAGE AND PRICE SETTING MODEL

When you look through graduate level textbooks in monetary theory and policy you find that the chapters on modern macro models with nominal rigidities begin with the idea of staggered contracts or staggered wage and price setting that had its origin in the 1970s at about the same time that the idea of rational expectations was being introduced to macroeconomics. Carl Walsh's treatment in his third edition (Walsh, 2010) of "early models of intertemporal nominal adjustment" starts with Taylor's (1979b, 1980) model of staggered nominal adjustment and then goes on to examine the version due to Calvo (1983). David Romer's chapter in his fourth edition (Romer, 2012) starts off with three modeling frameworks from this period: Phelps and Taylor (1977), Taylor (1979b), and Calvo (1983). Likewise, Woodford's (2003) chapter on nominal rigidities is mainly about staggered price or wage setting models that emanate from those days.

It is no coincidence that staggered contract models arose at about the same time as rational expectations were introduced to macroeconomics. Rational expectations meant that one could not rely on slow adjustment of expectations—so-called adaptive expectations—or on ad hoc partial adjustment models as the reason why prices and wages moved sluggishly over time. One had to think more about the economics in modeling the adjustment of prices and wages and the impact of monetary policy.

The earliest work by Fischer (1977), Gray (1976), and Phelps and Taylor (1977) assumed that the price or wage was set in advance of the period it would apply and at a value such that markets would be expected to clear.[d] In other words, prices would be set to bring expected demand into equality with expected supply. In the case of Phelps and Taylor (1977), the price was set one period in advance, and the price could change every period—no matter how short the period—much like in perfectly flexible

[d] These researchers were working largely independently of each other even though the papers were eventually published at the same time (and two in the same issue of the *Journal of Political Economy*). One possible exception was a conversation I had at the time with Stan Fischer who asked me what I was working on. I replied by describing a paper I was working with Phelps on sticky prices and rational expectations. Stan replied that he thought that it was a good topic, but I do not recall that he mentioned that he was working on the topic.

price models. In the case of Fischer (1977) and Gray (1976), the wage could be set more than one period in advance but at a different level each period, so that expected supply could equal expected demand in every period, again not much different empirically from flexible price models.

In all these models the price or the wage would change continuously, period by period. If the model was quarterly, then the price or wage could change every quarter; if the model was monthly, the price or wage could change every month. However, in the real world prices are set at the same level for more than one period; they usually remain at the same level for several weeks, months, or even quarters; and the same is true for wages with the representative period of constancy being about 12 months.

In addition to being inconsistent with the microeconomic data (as later confirmed in formal microeconomic empirical research referred to in the previous section), this type of model was completely inconsistent with the aggregate dynamics of wages, prices, or output. I realized this as soon as I tried to bring models along the lines of Phelps and Taylor (1977) to the data. Such models could not come close to generating the time series persistence or autocorrelation that was in real world data. In effect, the price or wage setting assumption in these models was only slightly different from the assumption that prices and wages were market clearing. I proposed the staggered contract model and its key property—the contract multiplier—as a way to generate needed persistence and solve this problem. The model was explicitly designed to capture the key characteristics of the micro data and at the same time to match the aggregate dynamics.

4. A CANONICAL STAGGERED PRICE AND WAGE SETTING MODEL

The simplest way to see this is to consider the canonical staggered price setting model illustrated in Fig. 4 using a degree of abstraction and simplification similar to expositions of the overlapping generations model. Later in this chapter, I will discuss a range of

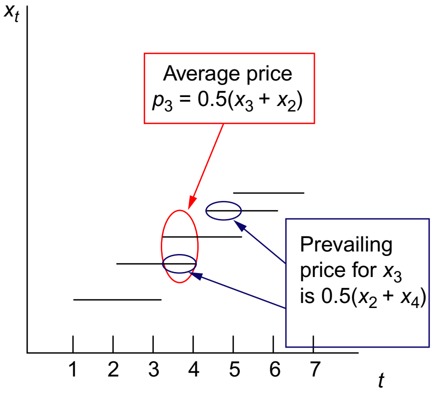


Fig. 4 Illustration of a canonical staggered contract model.

variations and extensions of this simple form. The basic idea of staggered price setting is that firms do not change their prices instantaneously from period to period. Instead there is a period of time during which the firm's price is fixed, and the pricing decisions of other firms are made the same way but at different times. Price setting is thus staggered and unsynchronized.

This "contract" or "set" price x_t is shown in Fig. 4. Note that it is fixed at the same level for two periods. Half the firms set their price each period in the canonical model. In the case where x is a wage rather than a price, it would also be set for two periods. There is no reason for either the price or the wage to be a formal contract or even an implicit contract; rather the price or wage set by the firm could apply to any particular good purchased or any worker of a certain type hired.

4.1 Canonical Assumptions

Two essential assumptions of staggered price setting are clear in Fig. 4. First, the set price lasts for more than an instant, or in this discrete time setup for more than one period. Second, the price setting is unsynchronized or overlapping. When you think about how a market might work in these circumstances, you realize two more important things not in the classic supply and demand framework. First, you realize that some firms' prices will be outstanding when another firm is deciding on a price to set. So firms need to look back at the price decisions of other firms. Second, you realize that the firm's price will be around for a while, so the firm will have to think ahead and forecast the price decisions of other firms.

Fig. 4 also illustrates two important concepts: the average price $p_t = (x_t + x_{t-1})$ and the prevailing price. For period t, the prevailing price is the average of the price in effect in period $t-1$ and the price expected to be in effect in period $t+1$, that is $0.5(x_{t-1} + E_{t-1}x_{t+1})$. This is what is relevant for the price decision of the firm in period t.

Given this setup, a decision rule for the firm setting the price x_t at time t can be written down directly, as I originally did in Taylor (1979a,b,c), as a function of the prevailing price (set by other firms in the market) and a measure of demand pressure in the market during the period the price will be in effect. The intuitive idea is simply that firms increase their price above the prevailing price if they see that demand conditions in the market are strong, and vice versa if demand conditions are weak. There can also be a random shock reflecting mistakes or other factors affecting the pricing decision. The result is shown in Eq. (1). As we will see later in this chapter, this equation can be derived explicitly from a specific profit maximization problem of a firm in monopolistic competition.[e]

[e] Note that (ignoring the expectations operator) the first term on the right-hand side of Eq. (1) can be written as $\frac{1}{2}(p_t + p_{t+1})$ because this equals $\frac{1}{2}\left[\frac{1}{2}(x_t + x_{t-1}) + \frac{1}{2}(x_{t+1} + x_t)\right]$ and thus $x_t = \frac{1}{2}(x_{t-1} + x_{t+1}) + \cdots$.

The term E_{t-1} represents the conditional expectations operator, the term y_t is a measure of demand (which for simplicity I will take to be the percentage deviation of real output from potential output), and ε_t is a serially uncorrelated, zero mean random shock.

$$x_t = \frac{1}{2}(x_{t-1} + E_{t-1}x_{t+1}) + \frac{\gamma}{2}(E_{t-1}y_t + E_{t-1}y_{t+1}) + \varepsilon_t \tag{1}$$

As I explain later, the "demand" variable on the right-hand side of Eq. (1) can also be interpreted as marginal cost in the case of a price decision (Woodford, 2003) or marginal revenue product in the case of a wage decision (Erceg et al., 2000) rather than the output gap.

4.2 Two More Equations and a Dynamic Stochastic General Equilibrium Model

To derive the implications of the staggered contracts assumption for aggregate dynamics and the persistence of shocks, we need to embed the staggered price setting equation into a model of the economy. For this purpose, consider two additional simple equations: An aggregate demand equation based on a money demand function (which could be derived from a money-in-the-utility or cash-in-advance framework) and an equation describing a monetary policy rule in which the money supply is adjusted by the central bank in response to movements in the price level. The two equations are thus:

$$y_t = \alpha(m_t - p_t) + v_t \tag{2}$$

$$m_t = gp_t \quad (g < 1) \tag{3}$$

which can be combined to get

$$y_t = -\beta p_t + v_t \tag{4}$$

where $\beta = \alpha(1-g)$ is the key policy parameter.

Here we define y to be the log of real output (detrended) as in Eq. (1) and m to be the log of the money supply. In the case where $\alpha = 1$, ν is simply the log of velocity, which can be a random variable with zero mean. The policy rule is effectively a price rule with a price level target of 0 for the log of the price level. Now if we insert the staggered contract Eq. (1) into the model we get the following difference equation with lags and leads

$$\begin{aligned} x_t &= \frac{1}{2}(x_{t-1} + E_{t-1}x_{t+1}) + \frac{\gamma}{2}\left[-\beta\left(\frac{E_{t-1}x_t + x_{t-1}}{2}\right) - \beta\left(\frac{E_{t-1}x_{t+1} + E_{t-1}x_t}{2}\right)\right] + \varepsilon_t \\ &= \frac{1}{2}(x_{t-1} + E_{t-1}x_{t+1}) - \frac{\gamma\beta}{4}[E_{t-1}x_{t+1} + 2E_{t-1}x_t + x_{t-1}] + \varepsilon_t \end{aligned}$$

The solution is

$$x_t = ax_{t-1} + \varepsilon_t \tag{5}$$

where $a = c \pm \sqrt{c^2 - 1}$ and where $c = (1 + \beta\gamma/2)/(1 - \beta\gamma/2)$. Clearly $c > 1$, and we can chose stable root for uniqueness. In terms of the aggregate price level, this implies that

$$p_t = ap_{t-1} + 0.5(\varepsilon_t + \varepsilon_{t-1}) \tag{6}$$

an ARMA(1,1) from which steady-state variances can easily be found

$$\sigma_p^2 = 0.5\sigma_\varepsilon^2/(1 - a)$$
$$\sigma_y^2 = \beta^2\sigma_p^2$$

Note that the three equation macro model consists of a staggered price setting Eq. (1), a policy transmission Eq. (2), and a policy rule (3). The model is a combination of sticky prices and rational expectations which is the hallmark of *New Keynesian* models, a term which distinguishes them from *Old Keynesian* models in which expectations are not rational and prices are either fixed or determined in a purely backward-looking manner, unlike Eq. (1). To be sure, the term New Keynesian is used in different ways by different researchers and can be misleading. For example, in some usages the term refers only to models in which the monetary transmission equation is an IS curve—perhaps derived from a Euler equation—relating the policy interest rate to aggregate demand and the policy rule is an interest rate rule like the Taylor rule.

Observe that the persistence of the aggregate price level, which is determined by the parameter a in Eq. (6), and aggregate output depends on the structure of the staggered pricing γ but also on the policy rule g. In other words, persistence is a general equilibrium phenomenon depending on both the price setting mechanism and on policy. This idea that one needs a whole model rather than a single price setting equation to assess the degree of aggregate persistence will come up again in this chapter.

Also note that in this simple model the money supply is stationary so the persistence is in the price level rather than the inflation rate. In a more realistic model, the growth rate of the money rather than the money supply would be stationary.

4.3 The Policy Problem and the Output and Price Stability Tradeoff Curve

An objective function or loss function for monetary policy in this model can be written in terms the variances of y_t and p_t. For example, if the loss function is $\lambda\text{var}(p_t) + (1 - \lambda)\text{var}(y_t)$, then the monetary policy problem is to choose a value of g (which determines β and thus a) to minimize this loss function. As the policy parameter is changed, the variances of p and y move in opposite directions tracing out a variance tradeoff curve. The lower panel of Fig. 5 illustrates this variance tradeoff curve. Inefficient monetary policies would be outside the curve. Points inside the curve are not feasible. Performance could be improved by moving toward the curve.

The upper panel of Fig. 5 is an aggregate demand–aggregate supply diagram which illustrates how the choice of g, and thus β, affects the variance of p and y. Suppose that there is a shock ε to the price setting equation. Then a steep aggregate demand

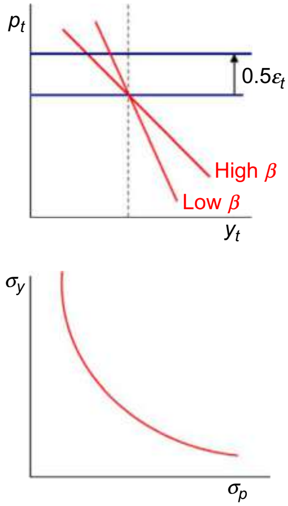


Fig. 5 Output and price stability tradeoff curve with graphical explanation.

curve (a monetary policy choice) makes for smaller fluctuations in y, but also means that a given shock to the price level takes a long time to diminish and thus a larger average fluctuation in p.

4.4 Key Implications

A number of important implications of staggered contracts can be illustrated with the canonical model, and they also hold in more complex models. I summarize these implications here.

(1) The theory centers around a simple equation that can be used and tested. I list this result first because if the theory had not yielded an equation, such as Eq. (1), it would have been difficult to achieve the progress I report in this chapter—including the empirical validation exercises reported in the previous section and the theoretical derivation of the equation using a profit maximization with monopolistic competition framework reported later. A key variable in this equation is the prevailing price (or wage) set by other firms. The prevailing price itself is an average of prices set in the past and prices to be set in the future. In this case the coefficients on past and the future are equal.

(2) Expectations of future prices matter for pricing decisions today. This is shown clearly in Eq. (1). The reason is that with the current price decision expected to last into the future, some prices set in the future will be relevant for today's decision. This is an important result because expectations of *future* inflation now come into play in the theory of inflation. It gives a rationale for central bank credibility and for having an inflation target.

(3) There is inertia or persistence in the price setting process; past prices matter because they are relevant for present price decisions. The coefficients on past prices can be calculated from the staggered price setting assumptions. This implication can be most readily seen in Eq. (5). The contract price is serially correlated. It is persistent and it can be described by an autoregressive process.

(4) The inertia or persistence is longer than the length of the period during which prices are fixed. Price shocks take a long time to run through the market because last period's price decisions depend on price decisions in the period before that and so on into the distant past. I originally called this phenomenon the "*contract multiplier*" because it was analogous to the Keynesian multiplier where a shock to consumption builds up and persists over time as it works its way through the economy from income to consumption to income back again and so on. This is most easily seen in Eq. (5) or the ARMA model in Eq. (6). The first-order autoregression implies an infinite autocorrelation function or an infinite impulse response function. The larger the autoregressive coefficient (that is, a) is, the larger will be the contract multiplier.

This is one of the most important properties of the staggered contract model because it means that very small rigidities at the micro level can generate large persistent effects for the aggregates. Klenow and Malin (2011) explain it well: "Real effects of nominal shocks … last three to five times longer than individual prices. Nominal stickiness appears insufficient to explain why aggregate prices respond so sluggishly to monetary policy shocks. For this reason, nominal price stickiness is usually combined with a 'contract multiplier' (in Taylor's, 1980 phrase)."

(5) The degree of inertia or persistence depends on monetary policy. That is, the autoregressive coefficient a depends on the policy parameter g. The more accommodative the central bank is to price level movements (higher g), the more inertia there will be (higher a).

(6) The theory implies a tradeoff curve between price stability and output stability. This tradeoff curve has provided a framework for discussion and debate about the role of policy in economic performance for many years. Originally put forth in Taylor (1979a) it is referred to as the Taylor curve in various contexts (King, 1999; Bernanke, 2004; Friedman, 2010). Bernanke (2004) used such a tradeoff curve to explain the role of monetary policy during the Great Moderation. His explanation was that monetary policy improved and this brought performance from the upper right-hand part of the diagram down and to the left closer to or even on the curve.

King (1999) made similar arguments. However, when the Great Recession and the slow recovery moved the performance in the direction of higher output instability—the end of the Great Moderation—King (2012) argued that the tradeoff curve itself shifted. As he put it, "A failure to take financial instability into account creates an unduly optimistic view of where the Taylor frontier lies …Relative to a Taylor frontier that reflects only aggregate demand and cost shocks, the addition of financial instability shocks generates what I call the Minsky-Taylor frontier."

Note that the tradeoff implies that there is no "*divine coincidence*" as put forth by Blanchard and Gali (2007). Divine coincidence means that there is no such tradeoff between output stability and price stability, completely contrary to the existence of the tradeoff in Fig. 5. Divine coincidence could occur if there were no shocks to the contract price or wage equation, but that is not the basic assumption of the staggered contract model. Broadbent (2014) suggested that the Great Moderation was due to the sudden appearance of divine coincidence, rather than to an improved monetary policy performance that brought the economy closer to the tradeoff curve as Bernanke (2004) and others argued.

(7) The costs of reducing inflation are less than in a backward-looking expectations augmented Phillips curve. In the staggered contract model, disinflation could be less costly if expectations of inflation were lower because of the forward-looking component of the model, as explained in Taylor (1982) though with reservations from others such as Gordon (1982). The disinflation costs would not normally be zero as in the case of rational expectations models with perfectly flexible prices, but they would be surprisingly small. This prediction proved accurate when people later examined the disinflation of the early 1980s.

5. GENERALIZATIONS AND EXTENSIONS

These results remain robust to variations in the model. An important variant is to allow for a greater variety of time intervals during which prices are fixed. Of course one could have longer contracts as in Taylor (1980) where contracts were of a general length *N*. However, a model with all price and wage setting being the same length is a simplifying assumption, not something that could be used in empirical work. The high degree of heterogeneity described in the microeconomic research reviewed earlier makes this very clear. Not all contracts are *N* periods in length; some are shorter and some are longer. Indeed, there is a whole distribution of contracts and this is what I assumed in early empirical work with these models. For example, a generalized distribution of price–wage setting intervals was used by Taylor (1979c) in an estimated model of the United States.

Eq. (1) was thus modified as follows:

$$x_t = \sum_{i=0}^{N-1} \theta_{it} E_t(p_{t+i} + \gamma y_{t+i} + \varepsilon_{t+i}) \tag{7}$$

$$p_t = \sum_{i=0}^{N-1} \delta_{it} x_{t-i} \tag{8}$$

The weights θ_{it} and δ_{it} were estimated using aggregate wage data in the United States. The estimation of the lag and lead coefficients was only mildly restricted, allowing for a peak somewhere between one and eight quarters. The estimated distribution from Taylor (1979c, table 4) is plotted in Fig. 6. It has a peak at three quarters with 24% of workers; only 7% had one quarter contracts and only 2% had eight quarter contracts.

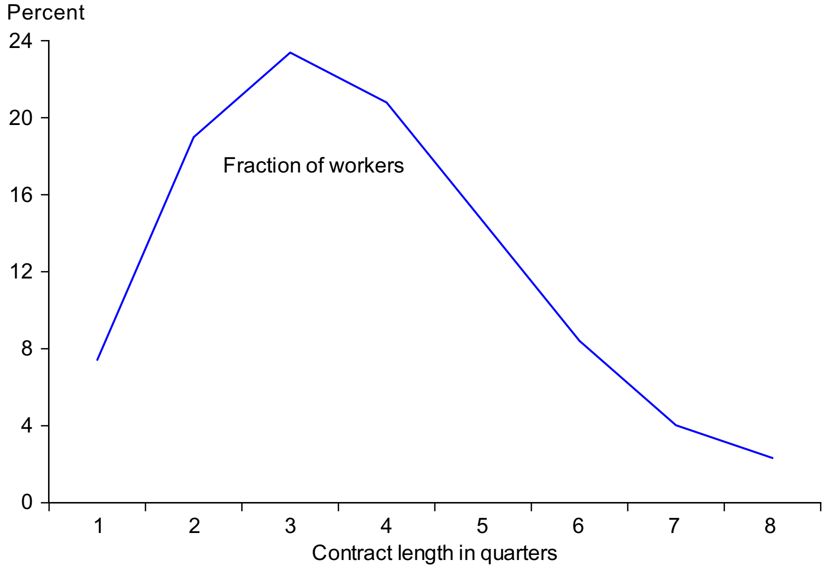


Fig. 6 The estimated distribution of workers by contract length.

The interpretation was that the economy consisted of a whole variety of price and wage setting practices.

Observing this empirical distribution of wage setting intervals in Taylor (1979c) gave my then colleague at Columbia University, Guillermo Calvo, the idea of an important simplification. Why not assume a geometric distribution, which would be considerably simpler? Moreover, such a distribution could be interpreted as being generated probabilistically rather than deterministically if each wage contract expired randomly rather than deterministically. The resulting model came to be called the Calvo model and the random selection process came to be called the Calvo fairy. The equation for the price change is a specific version of Eqs. (7) and (8) and can be written as follows:

$$x_t = (1-\beta\omega)\sum_{i=0}^{\infty}(\beta\omega)^i E_t(p_{t+i} + \gamma y_{t+i} + \varepsilon_t) \tag{9}$$

$$p_t = (1-\omega)\sum_{i=0}^{\infty}\omega^i x_{t-i} \tag{10}$$

After some manipulation, these two equations can be rewritten as

$$x_t = \beta\omega E_t x_{t+1} + (1-\beta\omega)(p_t + \gamma y_t + \varepsilon_t)$$
$$p_t = \omega p_{t-i} + (1-\omega)x_t$$

Once a model for y and the impact of monetary policy is added, you have a well-defined rational expectations model as before.

The two equations can also be rewritten in an interesting form:

$$\pi_t = \beta E_t \pi_{t+1} + \delta \gamma y_t + \delta \varepsilon_t \tag{11}$$

where

$$\delta = \left[\frac{(1-\omega)(1-\beta\omega)}{\omega} \right]$$

Which is very simple and reminiscent of an old expectations augmented Philips curve except that the expected inflation rate next period rather than this period is on the right-hand side. Calvo's modifications helped the staggered contract model grow in use and popularity.

Indeed, the form of the staggered price setting model in Eq. (1) came to be popularly known as the New Keynesian Phillips curve.

6. DERIVATION OF STAGGERED PRICE SETTING WHEN FIRMS HAVE MARKET POWER

Another important development regarding the staggered contract model was its derivation from an optimization problem in which firms face a downward sloping demand curve and decide on an optimal price subject to the staggered contract restriction that they cannot change prices every period. The idea of using market power to derive a price setting equation goes back to Svensson (1986), Blanchard and Kiyotaki (1987), and Akerlof and Yellen (1991) as I reviewed in Taylor (1999). As described below, Chari et al. (2000) used the approach as part of a critique of staggered price setting. For expository purposes here, I focus on a simple derivation used in Taylor (2000) in which firms maximize profits taking the downward sloping demand curve for their products as given.

Consider a firm selling a product that is differentiated from the other goods. The demand curve facing each firm is linear in the difference between the firm's own price for its product and the average price for the other differentiated products. Such a linear demand curve can be derived from models of consumer utility maximization. Suppose that this linear demand curve is written as

$$y_t = \varepsilon_t - \beta(x_t - p_t) \tag{12}$$

where y_t is production, x_t is the price of the good, and p_t is the average price of other (differentiated) goods. The term ε_t is a random shift to demand.

Suppose that the firm sets its price to last for two periods, and that it sets its price every second period. Other firms set their price for two periods, but at different points in time. These timing assumptions correspond to the canonical model in Fig. 1, and the average price is just as in the canonical model $p_t = 0.5(x_t + x_{t-1})$.

Let c_t be the marginal cost of producing the good. Under these assumptions, the firm's expected profit for the two periods to which the price set in period t applies is given by

$$\sum_{i=0}^{1} E_t(x_t y_{t+i} - c_{t+i} y_{t+i}) \tag{13}$$

where x_t applies in period t and period $t+1$. (I have assumed for simplicity that the discount factor is 1.) Firms maximize profits taking marginal cost and average price at other firms as given.

Differentiating with respect to x_t results in the solution for the optimal price

$$x_t = 0.25 \sum_{i=0}^{1} (E_t c_{t+i} + E_t p_{t+i} + E_t \varepsilon_{t+i}/\beta) \tag{14}$$

which is analogous to the canonical staggered contracting equation in Eq. (1) (see also Footnote a). Note however that it is marginal cost that enters the equation rather than the output gap, an issue I will come back to later in this chapter. Note that the coefficient of 0.25 implies that an increase in the price and marginal cost at other firms results in the same increase in the firm's price.

6.1 Pass-Through Implications

Though the derivation generates the same basic staggered price setting equation as assumed in the canonical model, it reveals another important implication of the theory—an "eighth" implication: a more price stability focused monetary policy—say due to inflation targeting—implies a smaller pass-through of price shocks (commodities or exchange rates) to inflation. That this implication might be borne out by reality was noted in Taylor (2000), but has now been documented in empirical studies in many countries. The reason originally given for the empirically observed decline in pass-through was that there was a reduction in the "pricing power" of firms. But another view is that the decline in pass-through is due to the low inflation rate achieved by a change in monetary policy.

To see this note that, according to Eq. (14), the amount by which a firm matches an increase in marginal cost with an increase in its own price depends on how permanent that marginal cost increase is. Similarly, the extent to which an increase in the price at other firms will lead to an increase in the firm's own price will depend on how permanent that increase in other firms' prices is expected to be. However, in neither case does the extent of this pass-through depend on the slope of the demand curve.

To see how the pass-through of an increase in marginal costs depends on the persistence of the increase, suppose that marginal cost follows a simple first-order autoregression:

$$c_t = \rho c_{t-1} + u_t$$

In this case, the pass-through coefficient will be proportional to $(1+\rho)$. Thus, less persistent marginal costs (lower ρ) reduce the pass-through coefficient, even though it might seem like a reduction in pricing power. The general point is that if an increase in costs is expected to last, then the increase will be passed-through to a greater extent. A more stable price level will reduce the persistence.

For firms that import inputs to production, marginal cost will depend on the exchange rate. Currency depreciation will raise the cost of the imports in domestic currency units. According to this model, if the depreciation is viewed as temporary, the firm will pass-through less of the depreciation in the form of a higher price. Hence, less persistent exchange rate fluctuations will lead to smaller exchange rate pass-through coefficients.

6.2 Marginal Cost vs the Output Gap

Note that Eq. (14) has marginal cost driving price movements rather than output as assumed in Eq. (1). To make the connection between Eqs. (14) and (1) (again keeping Footnote a in mind) we need to think of marginal cost as moving proportionately to the movements in the output gap. Gali and Gertler (1999) or Gali et al. (2005) argue that there are plenty of reasons why marginal cost and the output gap might diverge from time to time. So they look at a version of Eq. (11) in which marginal costs appear rather than the gap (they use the geometric distribution assumption of Calvo rather than the canonical form used here). Though the empirical accuracy of this equation was questioned by Mankiw (2001), the paper by Gali et al. (2005) finds that marginal cost is significant and quantitatively important. However, they introduce a modification in that model. They assume that a fraction of firms changes price with a backward looking "rule of thumb" which simply depends on past inflation. They thereby create a hybrid model with the lagged inflation rate on the right-hand side. The modification is ad hoc—especially compared with the theory that goes into deriving the staggered price setting equation.

Another issue noted by Nekarda and Ramey (2013) is that the markup of price over marginal cost needs to move in a countercyclical way if the equation is to explain empirically the effects of a change in demand on prices. They report, however, that markups are either "procyclical or acyclical conditional on demand shocks" and thereby conclude that the "New Keynesian explanation for the effects of government spending or monetary policy is not supported by the behavior of the markup."

Fuhrer (2006) raised further questions about the New Keynesian Phillips curve. He shows that in the New Keynesian Phillips curve inflation it is persistence of the shock rather than the equation itself that is the dominant source of persistence.

6.3 Debate Over the Contract Multiplier

Yet another issue is whether the contract multiplier is capable of explaining the persistence of prices or output. In the canonical model, including its derivation from profit maximization, the contract multiplier can be represented by the size of the autoregressive coefficient in the aggregate price equation. Chari et al. (2000) argued that for the parameters derived from the maximization problem, this coefficient is not large enough to be capable of explaining persistence, at least for contract lengths of one quarter in length and their particular measure of aggregate persistence. Woodford (2003, pp. 193–194) argues that their conclusion "depends on an exaggeration of the size of the contract multiplier that would be needed and an underestimate of the empirically plausible degree of strategic complementarities." He also argues that Chari et al. (2000) setup too high a persistence hurdle for the contract multiplier, in effect asking it to explain persistence that is more reasonably due to other serially correlated variables in the model.

Christiano et al. (2005) argue that assuming that the representative length of contracts is only one quarter is too small. If one uses somewhat longer contracts, say close to the survey summarized by Klenow and Malin (2011), the contract multiplier seems to work fine. Christiano et al. (2005) also question the persistence measure used by Chari et al. (2000).

7. PRICE AND WAGE SETTING TOGETHER

Much of this review has focused thus far on staggered *price* setting, but the original work on staggered contracts was about wages, where the time between wage changes is quite a bit longer according to the recent microeconomic empirical research summarized in this chapter. In Taylor (1980), the staggering of wages was the key part of the model, and this created a persistence of prices through a simple fixed markup of prices over wages. The micro finding summarized by Klenow and Malin (2011) that "price changes are linked to wage changes" supports this idea. Of course the markup need not be literally fixed. In the empirical multicountry model in Taylor (1993), the staggered wage contracting equations were estimated for seven countries and markups of prices over wages were influenced by the price of imports.

Erceg, Henderson, and Levin (2000) brought the focus back on wages, but with an important innovation. Rather than simply marking up prices over wages, they built a model which combined staggered price and wage setting, and, moreover, they derived both equations from profit or utility maximization considerations as in Section 5. Their work in turn helped enable the development of more empirically accurate estimated policy models, such as those due to Christiano et al. (2005), Smets and Wouters (2003), and many others that have become part of Volker Wieland's model database described in Wieland et al. (2012).

The model of Christiano et al. (2005) assumes staggered contracts for prices and wages with Calvo contracts. It was the first medium-sized, estimated example of a New Keynesian model explicitly derived from optimizing behavior of representative households and firms. It stimulated the development of similar optimization-based models for many other countries and has been dubbed the second-generation New Keynesian model along with Smets and Wouters (2003) by Wieland et al (2016).

Smets and Wouters (2003, 2007) also showed how to use Bayesian techniques (Geweke, 1999; Schorfheide, 2000) in estimating such models.

An important question for research is how the overall properties of the models changed as a result of the innovations. The eight implications mentioned earlier still hold in my view but the quantitative sizes of the impacts are important to pin down. Taylor and Wieland (2012) investigated this question using Wieland's database of models designed for this purpose. They considered a first-generation model—the Taylor (1993) multi-country model mentioned in the previous section with staggered contracts. And they compared this with two second-generation models—the Christiano et al. (2005) model and the Smets and Wouters (2007) model. Although the models differ in structure and sample period for estimation, the impacts of unanticipated changes in the federal funds rate are surprisingly similar. In the chapter prepared for this handbook, Wieland et al. (2016) show that these surprising results continue to hold if one adds a third-generation of models in which credit market frictions play a role in the monetary transmission mechanism.

There is a difference between the models in the evaluation of monetary policy rules, however. Model-specific policy rules that include the lagged interest rate, inflation, and current and lagged output gaps are not robust. Policy rules without interest-rate smoothing or with GDP-growth replacing the GDP gap are more robust, but performance in each model is worse with the more robust rule.

8. PERSISTENCE OF INFLATION AND INDEXING

Prior to the work of Chari et al. (2000), Fuhrer and Moore (1995) raised questions about the ability of the staggered contract model to explain the persistence of inflation rather than the persistence of the price level. They proposed a modification of the model to deal with this problem. As I reviewed in Taylor (1999), they transformed the model from price levels into the inflation rate, noting that it was *relative* wages rather than absolute wages that would go into the staggering equations. But the rationale for focusing on relative wages was weak and questions about this issue continued into the 2000s.

In recent years many have argued that the degree of persistence implied by the basic staggered contract model is just fine and consistent with the data. Guerrieri (2006), for example, argued that when the staggered contract model is viewed within the context of a fully specified macro model, inflation persistence and its changes over time could be explained with the regular staggered contract setup. I illustrated this idea with the canonical model I presented earlier in this chapter in which persistence is a general equilibrium phenomenon.

Guerrieri (2006) used a vector autoregression with inflation, the interest rate, and output to represent the facts that a staggered contract model should explain. He found that the basic staggered contract model did as well as the Fuhrer and Moore (1995) relative contract model in generating the actual inflation persistence in the United States through the 1990s. The impulse response functions reported in his paper show the degree to which both specifications can explain the inflation process. The staggered contract models are well within the 95% confidence bands with the exception of the cross-impulse response functions for output and inflation.

Nevertheless, both Christiano et al. (2005) and Smets and Wouters (2003) felt the need to modify the staggered price and wage setting equations in order to get the proper persistence and better match the other cross correlations. They assumed backward-looking indexation in those periods when prices and wages were not allowed to adjust. The Christiano et al. (2005) model assumes wages and prices are indexed to last period's inflation rate during periods between changes. The Smets–Wouters model assumes firms index to a weighted average of lagged and steady-state inflation.

None of these modifications are part of the optimization process; they are akin to simply assuming that wage and price inflation is autoregressive in an ad hoc way rather than deriving the equations: Why bother with a microfounded staggered wage and price setting model if you are just going to add ad hoc lag structure anyway?

According to recent research it appears that the persistence problem is not due the staggered contract model but rather to the special Calvo form it takes in these models.

9. TAYLOR CONTRACTS AND CALVO CONTRACTS

Much has been written comparing "Calvo contracts" described in Section 5 and "Taylor contracts" which appear in the canonical model in the case of two period contracts in Section 4. Walsh (2010, p. 243) notes some of the similarities between equations (his eqs. 6.17 and 6.36) derived from the two staggered price setting models, but others, including Kiley (2002), have emphasized the differences. For example, the persistence of inflation and output appears to be greater in the Calvo contracts for the same average frequency of price change.

There is no question that there is a much longer tail in the Calvo model than for any fixed-length contract, but Dixon and Kara (2006) argue that Kiley's comparison is flawed because it compares "the average age of Calvo contracts with the completed length of Taylor contracts." When Dixon and Kara (2006) compare average age Taylor contracts with the same average age Calvo contracts, the differences become much smaller. They also show that output can be more autocorrelated with Taylor contracts with "age-equivalent" Calvo contracts.

Carvalho and Schwartzman (2015) examine the differences in monetary neutrality in the two types of models by distinguishing between Taylor contracts and Calvo contracts in terms of their "selection effect." At any point in time after a monetary shock, some firms have a lot of old prices and some do not. "Positive" selection is defined as a situation where old prices are overrepresented among adjusting prices. In Taylor contracts, selection favors old prices; in Calvo contracts there is no selection, since prices change completely at random. This selection effect characterizes pricing frictions. Taylor contracts imply smaller nonneutralities of money on output than Calvo contracts because of differences in selection.

Of course there is no reason to focus—as these studies do—on the special case of "Taylor contracts" in which all contracts are the same length as in the simple exposition in the canonical model. The microeconomic evidence and casual observation suggest rather that there is a great deal of heterogeneity of lengths of both wage contracts and price contracts. In a series of papers, Dixon and Kara (2005, 2006, 2011) and Kara (2010) develop models which are built on this heterogeneity. They call these models a generalized Taylor economy (GTE) in which many sectors have staggered contracts with different lengths. When two such economies have the same average length contracts, monetary shocks are more persistent with longer contracts. They also show that when two GTE's have the same distribution of completed contract lengths, the economies behave in a similar manner. See also Huw Dixon's comprehensive web page http://huwdixon.org/GTE.html on the GTE and his paper with Dixon and Le Bihan (2012).

In a more recent paper, Kara (2015) shows that adding the heterogeneity in price stickiness to the Smets and Wouters model deals with criticisms of the staggered contract model including the Chari et al. (2009) criticism that the Smets and Wouters model relies on unrealistically large price mark-up shocks to explain the data on inflation and the Bils et al. (2012) criticism that reset price inflation in the model is more volatile than the data show. Kara (2015) shows that adding heterogeneity in the length of contracts to correspond with the data implies smaller price mark-up shocks and less volatile reset price inflation.

In yet another study comparing the two approaches, Knell (2010) examined survey data on wage setting in 15 European countries from the WDN discussed in Section 2. It is informative to quote from his paper: "There are at least four dimensions along which the data contradict the basic model with Calvo contracts. First, the majority of wage agreements seems to follow a predetermined pattern with given contract lengths. Second, while for most contracts this predetermined length is 1 year (on average 60% in the WDN survey) there exists also some heterogeneity in this context and a nonnegligible share of contracts has longer (26%) or shorter (12%) durations. Third, 54% of the firms asked in the WDN survey have indicated that they carry out wage changes in a particular month (most of them—30%—in January). Fourth, 15% of all firms report to use automatic indexation of wages to the rate of inflation. In order to be able to take these real-world characteristics of wage setting into account one has to move beyond the convenient but restrictive framework of Calvo wage contracts." Knell then presents a model along the lines of Taylor (1980) that allows one to incorporate all of these institutional details.

Musy (2006) and Ben Aissa and Musy (2010) have investigated the differences between the Calvo contracts model and the Taylor contracts model and others. Their analysis shows that criticism of a lack of persistence or an under estimate of the costs of disinflation are due to very special features of the Calvo assumptions. Recall that the "Calvo fairy" is a mechanism for randomly choosing a price to change each period. That probability is a constant, so in effect Calvo contracts are neither time dependent nor state dependent. The work of Musy and Ben Aissa shows that a change in money growth will not be accomplished in a costless manner in the Taylor model even though it is in the Calvo model, and that persistence is greater.

10. STATE-DEPENDENT MODELS AND TIME-DEPENDENT MODELS

Another development has been to relax the simplifying assumption that prices are set for an exogenous interval and allow the firm's price decision to depend on the state of the market, which gave rise to name "state dependent" pricing models and created the need to give the original canonical model a new name, "time dependent" (see Dotsey et al., 1999; Golosov and Lucas, 2007; Gertler and Leahy, 2008). There are some benefits from these improvements as Klenow and Kryvtsov (2008) have shown using new microeconomic data. Many of the key policy implication mentioned earlier hold, but the impact of monetary shocks can be smaller.

Alvarez and Lippi (2014) consider a state-dependent model with multiproduct firms, which is otherwise similar to the state-dependent model of Golosov and Lucas (2007). They find that as they alter the model from one product firm to a multiproduct firm, the impact of monetary shocks becomes larger and more persistent. For a large number of products they show that the economy works as in the staggered contract model: it has the same aggregation and impulse response to a monetary shock. In this sense, the menu cost models with multiproduct firms gives another basis to the staggered contract model.

Woodford (2003, p. 142) questions whether the state-dependent models are really any better than the staggered contract models. Not only are they more complex, he argues, but they may be less realistic and have inferior microfoundations. The idea that firms are constantly evaluating the price misses the point that firms set their prices for a while to reduce "the costs associated with information collection and decision making." Kehoe and Midrigan (2010) have developed a model in which formal considerations of such management costs do indeed increase the impact and persistence of shocks.

Bonomo and Carvalho (2004) develop a model of the microfoundations of the time-dependent model in which the length of time that prices are fixed is endogenous. In their model firms face a joint lump-sum adjustment and information cost rather than a pure adjustment cost, and for this reason optimal pricing is not state dependent. Their model is thus a way to deal with the observation that contract length depends on the rate of inflation and the variability of inflation and other shocks. They not only show that time-dependent models are optimal, they derive the optimal contract length.

They examine the effect of different policies such as a disinflation and examine the difference with invariant time-dependent arrangements. In a subsequent paper, Bonomo and Carvalho (2010) estimate the macroeconomic costs of a lack of credibility of monetary policy. They find that the costs are greater for the endogenous time-dependence model than for an exogenous time-dependent model.

11. WAGE-EMPLOYMENT BARGAINING AND STAGGERED CONTRACTS

In recent years, there has been an increased interest in explaining fluctuations in unemployment as well as output. As explained by Hall (2005), the standard wage-employment bargaining model needs to assume some form of sticky wages if it is to be consistent with the data, and for this reason the idea of nominal rigidities is common to this research. It is not surprising therefore that many of the models built to examine this question have combined staggered contracts with a formal treatment of the wage-employment bargaining. Ravenna and Walsh (2008), Gertler et al. (2008), and Christiano et al. (2013) are examples.

There are some by-products of this research too. The Christiano et al. (2013) model is able to drop the arbitrary indexing assumption in Christiano, Eichenbaum, and Evans and still get the requisite persistence. This works because when a monetary shock increases the demand for output which sticky price firms produce, the firms also purchase more wholesale goods. With this model, the authors argue that "alternating offer bargaining mutes the increase in real wages, thus allowing for a large rise in employment, a substantial decline in unemployment, and a small rise in inflation."

12. STAGGERED CONTRACTS VS INATTENTION MODELS

Mankiw and Reis (2001) have argued that the staggered wage and price setting should be replaced by a model with inattention. They argue in favor of sticky information rather than sticky prices, mainly because such a model would solve the persistence problem alluded to earlier. Recall that the concern is that there may be too little persistence of inflation following monetary shocks in staggered price setting models. Though some would argue that the persistence is fine, the lack of persistence may be more related to the specific form of the Calvo model rather than to the staggered contracts per se.

Why do Mankiw and Reis (2001) find that there is more persistence with inattention than with staggered contracts? Upon examination of their model, it appears that in the sticky information model, the price could be set to increase during the period where it is fixed in the regular model. For example in a staggered contract model of four periods the price would be 1.015, 1.015, 1.015, and 1.015 while in the sticky information it could be set as 1.0, 1.01, 1.02, and 1.03 and not change from that path. In effect, some inflation persistence is built in. Fig. 7 illustrates this and can be compared with Fig. 4.

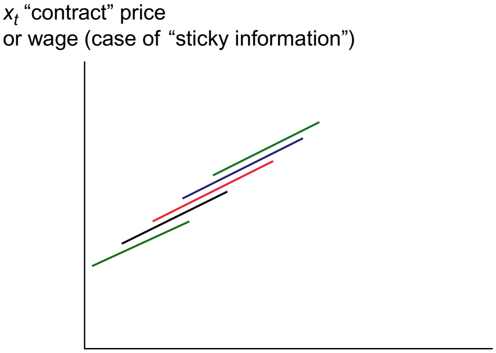


Fig. 7 Price setting with sticky information (for comparison with Fig. 4).

If prices or wages are set in this way, it is clear that there will be more persistence of inflation. It is very rare, however, for prices or wages to be set in this manner except in multiyear union contracts as explained in Taylor (1983) and Avouyi-Dovi (2013).

13. CRITICAL ASSESSMENT AND OUTLOOK

From its origins nearly four decades ago to its applications today, the staggered wage and price setting model continues to be a focus of attention in empirical and theoretical research in macroeconomics, especially in monetary business cycle models and monetary models used for policy analysis. In recent years, "Big Data" style research projects have radically expanded our knowledge of the microeconomics of wage and price setting behavior from a few salient facts about magazine prices or personal salary experiences into complex datasets with thousands or millions of observations. These datasets require new methods of analysis, but they also permit researchers to test and discriminate much more thoroughly between one type of model and another. Criticisms—whether about inadequate microfoundations, inability to explain certain facts, or questionable policy implications—have led to constructive improvements, clarifications, variations, new research lines, and, in some cases, less than fully satisfactory fixes.

In assessing the outlook for future research and applications of these models, one cannot help but be struck by a certain tension in current research. The large-scale surveys and empirical research show a great deal of heterogeneity in wage and price setting behavior, yet most models still employ simplified models clearly at odds with this heterogeneity. Yes, there is evidence that prices are set at a fixed level for 6 months or more, especially if sales and reference prices are accounted for properly. Yes, there is evidence that wages are set a fixed level for longer periods and that there is a peak in the estimated hazard function at 1 year that precludes certain simplifications such as the Calvo model. Yes, there is evidence that both wage and price decisions are staggered or unsynchronized over time, and that this staggering creates a contract multiplier which converts short spells of rigidity at the micro level into longer persistence at the macro level. Yes, there is more evidence of time dependence than state dependence. But in each of these dimensions—length, degree of staggering, shape of the hazard function, degree of state dependence—there is a great deal of heterogeneity across countries, types of product, types of employment, and types of industry structure.

This heterogeneity is not simply a nuisance; it has major implications for aggregate dynamics, and it has been offered as a response to criticism of staggered wage and price setting models. Often that criticism applies to a particular simple staggered contract model that does not capture either the regularities mentioned earlier or the heterogeneity, and that criticism disappears when heterogeneity is taken into account as Kara (2010) and Knell (2010) have emphasized. Rather than "jury-rig" simple staggered contract models with ad hoc add-ons, such as indexing in the models by Christiano et al. (2005) or Smets and Wouters (2003), this research suggests that building the heterogeneity into the model would both better fit the micro data and provide a straight forward explanation of macro persistence.

In other words, future research would likely yield large benefits if it moved on from "representative" staggered wage and price setting models to "heterogeneous" staggered wage and price setting models. The suggestion is similar to the idea of moving from "representative agent models" to "heterogeneous agent models," though the gains from such a move could be much greater.

The challenge is that building in this heterogeneity would complicate existing macro models which are already quite complicated, as I found when I began to build in such heterogeneity in my early research (Taylor, 1979c) including in a multicountry model (Taylor, 1993) with different degrees of staggered wage setting in different countries. Indeed, their complexity is the main object of criticism of the existing models as expressed by Chari et al. (2009) and others.

At the least future research could go beyond continued comparisons of simplest textbook style models—such as the random-length-contract Calvo model and the N-period-length-contract Taylor model—and look at heterogeneous or generalized models with a mix of contract types. But more fundamentally the challenge for future work is to take account of the rich variety of wage and price setting procedures in a way that is tractable and understandable for policy analysis. Indeed, that has been the challenge for all areas of macroeconomic research from the very beginning.

ACKNOWLEDGMENT

I wish to thank Susanto Basu, John Cochrane, Huw Dixon, Robert Hall, Jim Hamilton, Engin Kara, Pete Klenow, Olivier Musy, Carlos Viana de Carvalho, Harald Uhlig, and Carl Walsh for helpful comments.

REFERENCES

Akerlof, G.A., Yellen, J.L., 1991. How large are the losses from rules of thumb behavior in models of the business cycle. In: Brainard, W., Nordhaus, W., Watts, H. (Eds.), Money, Macroeconomics, and Economic Policy: Essays in Honor of James Tobin. MIT Press, Cambridge, MA.

Alvarez, F., Lippi, F., 2014. Price setting with menu cost for multiproduct firms. Econometrica 82 (1), 89–135.

Avouyi-Dovi, S., Fougère, D., Gautier, E., 2013. Wage rigidity, collective bargaining and the minimum wage: evidence from French agreement data. Rev. Econ. Stat. 95 (4), 1337–1351.

Backus, D., 1984. Exchange rate dynamics in a model with staggered wage contracts. Discussion Paper No. 561 (Queen's University).

Barattieri, A., Basu, S., Gottschalk, P., 2014. Some evidence on the importance of sticky wages. Am. Econ. J. Macroecon. 6 (1), 70–101.

Ben Aissa, M.S., Musy, O., 2010. The dynamic properties of alternative assumptions on price adjustment in New Keynesian models. Bull. Econ. Res. 63 (4), 353–384.

Benabou, R., Bismut, C., 1987. Wage bargaining and staggered contracts: theory and estimation. Discussion Paper No. 8810, CEPREMAP, Paris, France.

Bernanke, B.S., 2004. The Great Moderation. Eastern Economic Association, Washington, DC.

Bils, M., Klenow, P.J., 2004. Some evidence on the importance of sticky prices. J. Polit. Econ. 112 (5), 947–985.

Bils, M., Klenow, P.J., Malin, B.A., 2012. Reset price inflation and the impact of monetary policy shocks. Am. Econ. Rev. 102 (6), 2798–2825.

Blanchard, O., Gali, J., 2007. Real wage rigidities and the New Keynesian model. J. Money Credit Bank. 39 (Suppl. 1), 35–64.

Blanchard, O.J., Kiyotaki, N., 1987. Monopolistic competition and the effects of aggregate demand. Am. Econ. Rev. 77, 647–666.

Blinder, A.S., Canetti, E.D., Lebow, D.E., Rudd, J.B., 1998. Asking About Prices: A New Approach to Understanding Price Stickiness. Russell Sage Foundation, New York, NY.

Bonomo, M., Carvalho, C., 2004. Endogenous time-dependent rules and inflation inertia. J. Money Credit Bank. 36 (6), 1015–1041.

Bonomo, M., Carvalho, C., 2010. Imperfectly credible disinflation under endogenous time-dependent pricing. J. Money Credit Bank. 42 (5), 799–831.

Broadbent, B., 2014. Unemployment and the conduct of monetary policy in the UK. In: Federal Reserve Bank of Kansas City Economic Symposium, Jackson Hole, Wyoming, August.

Calvo, G.A., 1983. Staggered contracts in a utility-maximizing framework. J. Monet. Econ. 12, 383–398.

Carlton, D.W., 1989. The theory and the facts of how markets clear: is industrial organization valuable for understanding macroeconomics? In: Schmalensee, R., Willig, R.D. (Eds.), Handbook of lndustrial Organization, vol. 1. North-Holland, Amsterdam, pp. 909–946.

Carvalho, C., Schwartzman, F., 2015. Selection and monetary non-neutrality in time-dependent pricing models. J. Monet. Econ. 76 (C), 141–156.

Cecchetti, S.G., 1986. The frequency of price adjustment: a study of newsstand prices of magazines. Econ. J. 31, 255–274.

Cecchetti, S.G., 1984. Indexation and incomes policy: a study of wage adjustment in unionized manufacturing. J. Labor Econ. 5, 391–412.

Chari, V.V., Kehoe, P., McGrattan, E., 2000. Sticky price models of the business cycle: can the contract multiplier solve the persistence problem? Econometrica 68 (5), 1151–1180.

Chari, V.V., Kehoe, P.J., McGrattan, E.R., 2009. New Keynesian models: not yet useful for policy analysis. Am. Econ. J. Macroecon. 1 (1), 242–266.

Christiano, L., Eichenbaum, M., Evans, C., 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. J. Polit. Econ. 113 (1), 1–45.

Christiano, L.J., Eichenbaum, M.S., Trabandt, M., 2013. Unemployment and Business Cycles. Unpublished working paper, Northwestern University.

Dixon, H., Kara, E., 2005. Persistence and nominal inertia in a generalized Taylor economy: how longer contracts dominate shorter contracts. Working Paper No. 489, European Central Bank.

Dixon, H., Kara, E., 2006. How to compare Taylor and Calvo contracts: a comment on Michael Kiley. J. Money Credit Bank. 38 (4), 1119–1126.

Dixon, H., Kara, E., 2011. Contract length heterogeneity and the persistence of monetary shocks in a dynamic generalized Taylor economy. Eur. Econ. Rev. 55, 280–292.

Dixon, H., le Bihan, H., 2012. Generalised Taylor and generalised Calvo price and wage setting; micro-evidence with macro implications. Econ. J. 122 (May), 532–554.

Dotsey, M., King, R.G., Wolman, A.L., 1999. State-dependent pricing and the general equilibrium dynamics of money and output. Q. J. Econ. 114 (2), 655–690.

Dutta, S., Levy, D., Bergen, M., 2002. Price flexibility in channels of distribution: evidence from scanner data. J. Econ. Dyn. Control. 26, 1845–1900.

Erceg, C., Henderson, D., Levin, A., 2000. Optimal monetary policy with staggered wage and price contracts. J. Monet. Econ. 46 (2), 281–313.

Fabiani, S., Druant, M., Hernando, I., Kwapil, C., Landau, B., Loupias, C., Martins, F., Mathä, T., Sabbatini, R., Stahl, H., Stokman, A., 2006. What firms' surveys tell us about price-setting behavior in the euro area. Int. J. Cent. Bank. 5 (3), 3–47. Special Issue on Staggered Pricing Models Face the Facts.

Fischer, S., 1977. Long-term contracts, rational expectations, and the optimal money supply rule. J. Polit. Econ. 85 (1), 191–205.

Fregert, K., Jonung, L., 1986. Monetary regimes and the length of wage contracts: Sweden 1908-1995. Working Paper 1998-3, University of Lund.

Friedman, M., 2010. Trade-offs in monetary policy. In: David Laidler's Contributions to Economics. Palgrave MacMillan, London.

Fuhrer, J.C., 2006. Intrinsic and inherited inflation persistence. Int. J. Cent. Bank. 5 (3), 49–86. Special Issue on Staggered Pricing Models Face the Facts.

Fuhrer, J.C., Moore, G.R., 1995. Inflation persistence. Q. J. Econ. 110 (1), 127–159.

Gali, J., Gertler, M., 1999. Inflation dynamics: a structural econometric analysis. J. Monet. Econ. 44, 195–222.

Gali, J., Gertler, M., Lopez-Salido, J.D., 2005. Robustness of the estimates of the hybrid New Keynesian Phillips curve. J. Monet. Econ. 52, 1107–1118.

Gertler, M., Leahy, J., 2008. A Phillips curve with an Ss foundation. J. Polit. Econ. 116, 3.

Gertler, M., Sala, L., Trigari, A., 2008. An estimated monetary DSGE model with unemployment and staggered nominal wage bargaining. J. Money Credit Bank. 40 (8), 1713–1764.

Geweke, J., 1999. Using simulation methods for Bayesian econometric models: inference, development and communication. Econ. Rev. 18, 1–126.

Golosov, M., Lucas Jr., R.E., 2007. Menu costs and Phillips curves. J. Polit. Econ. 115 (2), 199–271.

Gordon, R.J., 1982. Discussion. In: Monetary Policy Issues for the 1980s. Federal Reserve Bank of Kansas City, Symposium, Jackson Hole Wyoming.

Gray, J.A., 1976. Wage indexation: a macroeconomic approach. J. Monet. Econ. 2 (2), 221–235.

Guerrieri, L., 2006. Inflation persistence of staggered contracts. J. Money Credit Bank. 38 (2), 483–494.

Hall, R., 2005. Employment fluctuations with equilibrium wage stickiness. Am. Econ. Rev. 95 (1), 50–65.

Hume, D., 1742. On money. Part II, Essay III, paragraph 7 of his In: Essays, Moral, Political, and Literary. Liberty Fund Books. http://www.econlib.org/library/LFBooks/Hume/hmMPL.html.

Kara, E., 2010. Optimal monetary policy in the generalized Taylor economy. J. Econ. Dyn. Control. 34, 2023–2037.

Kara, E., 2015. The reset inflation puzzle and the heterogeneity in price stickiness. J. Monet. Econ. 76, 29–37.

Kashyap, A.K., 1995. Sticky prices: new evidence from retail catalogues. Q. J. Econ. 110, 245–274.

Kehoe, P., Midrigan, V., 2010. Prices are sticky after all. NBER Working Paper No. 16364.

Kiley, M., 2002. Price adjustment and staggered price-setting. J. Money Credit Bank. 34, 283–298.

King, M., 1999. Challenges for monetary policy: new and old. In: New Challenges for Monetary Policy. Federal Reserve Bank of Kansas City, Jackson Hole.

King, M., 2012. Twenty years of inflation targeting. Stamp Memorial Lecture. London School of Economics, London. October 9.

Klenow, P., Kryvtsov, O., 2008. State dependent versus time dependent pricing. Q. J. Econ. 72 (2), 863–904.

Klenow, P., Malin, B., 2011. Microeconomic evidence on price setting. In: Friedman, B., Woodford, M. (Eds.), Handbook of Monetary Economics, vol. 3. Elsevier, Amsterdam.

Knell, M., 2010. Nominal and real wage rigidities in theory and in Europe. Working Paper Series, European Central Bank.

Lach, S., Tsiddon, D., 1996. Staggering and synchronization in price-setting: evidence from multiproduct firms. Am. Econ. Rev. 86, 1175–1196.

Lamo, A., Smets, F., 2009. Wage Dynamics in Europe: Final Report of the Wage Dynamics Network (WDN). https://www.ecb.europa.eu/home/pdf/wdn_finalreport_dec2009.pdf?68e28b96d494632f27900b1c453586c4. December 4.

Le Bihan, H., Montornès, J., Heckel, T., 2012. Sticky wages: evidence from quarterly microeconomic data. Am. Econ. J. Macroecon. 4 (3), 1–32.

Levin, A., 1991. The macroeconomic significance of nominal wage contract duration. Working Paper No. 91-08, February. University of California at San Diego.

Lünnemann, P., Wintr, L., 2009. Wages are flexible, aren't they? Evidence from monthly micro wage data. Working Paper Series, No. 1074, July, Wage Dynamic Network.

Mankiw, N.G., 2001. The inexorable and mysterious tradeoff between inflation and unemployment. Econ. J. 117, 1295–1328.

Mankiw, N.G., Reis, R., 2001. Sticky information versus sticky prices: a proposal to replace the New Keynesian Phillips curve. NBER Working Paper No. 8290.

Musy, O., 2006. Inflation persistence and the real costs of disinflation in staggered prices and partial adjustment models. Econ. Lett. 91, 50–55.

Nakamura, E., Steinsson, J., 2008. Five facts about prices: a reevaluation of menu cost models. Q. J. Econ. 123 (4), 1415–1464.

Nekarda, C.J., Ramey, V.A., 2013. The Cyclical Behavior of the Price–Cost Markup. University of California, San Diego, CA.

Okun, A.M., 1981. Prices and Quantities: A Macroeconomic Analysis. Brookings Institution, Washington, DC.

Olivei, G., Tenreyro, S., 2010. Wage setting patterns and monetary policy: international evidence. J. Monet. Econ. 57, 785–802.

Phelps, E., Taylor, J.B., 1977. Stabilizing powers of monetary policy under rational expectations. J. Polit. Econ. 85 (1), 163–190.

Ravenna, F., Walsh, C., 2008. Vacancies, unemployment, and the Phillips curve. Eur. Econ. Rev. 52, 1494–1521.

Romer, D., 2012. Advanced Macroeconomics, fourth ed. McGraw-Hill, New York, NY.

Schorfheide, F., 2000. Loss function based evaluation of DSGE models. J. Appl. Econ. 15 (6), 645–670.

Sigurdsson, J., Sigurdardottir, R., 2011. Evidence of nominal wage rigidity and wage setting from Icelandic microdata. Working Paper No. 55, Central Bank of Iceland.

Sigurdsson, J., Sigurdardottir, R., 2016. Time-dependent or state-dependent wage-setting? Evidence from periods of macroeconomic instability. J. Monet. Econ. 78, 50–66.

Smets, F., Wouters, R., 2003. An estimated dynamic stochastic general equilibrium model of the euro area. J. Eur. Econ. Assoc. 1 (5), 1123–1175.

Smets, F., Wouters, R., 2007. Shocks and frictions in U.S. Business cycles: Bayesian DSGE approach. Am. Econ. Rev. 97 (3), 506–606.

Stigler, G., Kindahl, J., 1970. The Behavior of Industrial Prices. NBER General Series, No. 90, Columbia University Press, New York, NY.

Svensson, L.E.O., 1986. Sticky goods prices, flexible asset prices, monopolistic competition, and monetary policy. Rev. Econ. Stud. 52, 385–405.

Taylor, J.B., 1979a. Estimation and control of a macroeconomic model with rational expectations. Econometrica 47 (5), 1267–1286. September. Reprinted in Lucas, R.E., Sargent, T.J. (Eds.), 1981. Rational Expectations and Econometric Practice, University of Minnesota Press.

Taylor, J.B., 1979b. Staggered wage setting in a macro model. Am. Econ. Rev. 69 (2), 108–113. May. Reprinted in Gregory Mankiw, N., Romer, D. (Eds.), 1991. New Keynesian Economics. MIT Press, Cambridge.

Taylor, J.B., 1979c. An econometric business cycle model with rational expectations: some estimation results. Working Paper, June, Columbia University. http://web.stanford.edu/~johntayl/Onlinepaperscombinedbyyear/1979/An_Econometric_Business_Cycle_Model_with_Rational_Expectations-Some_Estimations_Results-1979.pdf.

Taylor, J.B., 1980. Aggregate dynamics and staggered contracts. J. Polit. Econ. 88 (1), 1–23.

Taylor, J.B., 1982. The role of expectations in the choice of monetary policy. In: Monetary Policy Issues for the 1980s. Federal Reserve Bank of Kansas City Economic Symposium, Jackson Hole, Wyoming, August.

Taylor, J.B., 1983. Union wage settlements during a disinflation. Am. Econ. Rev. 73 (5), 981–993.

Taylor, J.B., 1993. Macroeconomic Policy in a World Economy: from Econometric Design to Practical Operation. W.W. Norton, New York, NY.

Taylor, J.B., 1999. Staggered price and wage setting in macroeconomics. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics. first ed., part 1 Elsevier, North-Holland, pp. 1009–1050.

Taylor, J.B., 2000. Low inflation, pass-through, and the pricing power of firms. Eur. Econ. Rev. 44 (7), 1389–1408.

Taylor, J.B., 2007. Thirty-five years of model building for monetary policy evaluation: breakthroughs, dark ages, and a renaissance. J. Money Credit Bank. 39 (Suppl. 1), 193–201.

Taylor, J.B., Wieland, V., 2012. Surprising comparative properties of monetary models: results from a new model data base. Rev. Econ. Stat. 94 (3), 800–816.

Walsh, C.E., 2010. Monetary Theory and Policy, third ed. The MIT Press, Cambridge, MA.

Wieland, V., Cwik, T., Müller, G.J., Schmidt, S., Wolters, M., 2012. A new comparative approach to macroeconomic modeling and policy analysis. J. Econ. Behav. Organ. 83, 523–541.

Wieland, V., Afanasyeva, E., Kuete, M., Yoo, J., 2016. New methods for macro-financial model comparison and policy analysis. In: Handbook of Macroeconomics, vol. 2A. Elsevier, Amsterdam, Netherlands, pp. 1241–1319.

Woodford, M., 2003. Interest and Prices. Princeton University Press, Princeton, NJ.

Neoclassical Models in Macroeconomics

G.D. Hansen[*],[†], L.E. Ohanian[*],[†],[‡]
[*]UCLA, Los Angeles, CA, United States
[†]NBER, Cambridge, MA, United States
[‡]Hoover Institution, Stanford University, Stanford, CA, United States

Contents

Handbook of Macroeconomics
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.04.010

Abstract

This chapter develops a toolkit of neoclassical macroeconomic models, and applies these models to the US economy from 1929 to 2014. We first filter macroeconomic time series into business cycle and long-run components, and show that the long-run component is typically much larger than the business cycle component. We argue that this empirical feature is naturally addressed within neoclassical models with long-run changes in technologies and government policies. We construct two classes of models that we compare to raw data, and also to the filtered data: *simple neoclassical models*, which feature standard preferences and technologies, rational expectations, and a unique, Pareto optimal equilibrium, and *extended neoclassical models*, which build in government policies and market imperfections. We focus on models with multiple sources of technological change, and models with distortions arising from regulatory, labor, and fiscal policies. The models account for much of the relatively stable postwar US economy, and also for the Great Depression and World War II. The models presented in this chapter can be extended and applied more broadly to other settings. We close by identifying several avenues for future research in neoclassical macroeconomics.

Keywords

Neoclassical models, Dynamic general equilibrium, Great Depression World War II, Band pass filter, Productivity shocks, Low frequency fluctuations, Business cycles, Economic growth, Great moderation, Great recession

JEL Classification Codes

E13, E2, E6

1. INTRODUCTION

This chapter analyzes the role of *neoclassical models* in the study of economic growth and fluctuations. Our goal is to provide macroeconomists with a toolkit of models that are of interest in their own right, and that easily can be modified to study a broad variety of macroeconomic phenomena, including the impact of economic policies on aggregate economic activity.

Since there is no generally recognized definition of neoclassical macroeconomics within the profession, we organize the development of these models around two principles. One is based on the exogenous factors driving changes in aggregate time series, and the other is based on the classes of model economies that we consider.

The primary sources of changes in macroeconomic variables that we study are long-run changes in technologies and government policies. We focus on these factors because of the observed large changes in productivity and in policies that affect the incentives and opportunities to produce and trade. Policy factors that we consider include changes affecting competition and business regulatory policies, labor policies, and fiscal policies.

We study two classes of intertemporal models that we call *neoclassical macroeconomic models*. The first has standard preferences and technologies, competitive markets, rational expectations, and there is a unique equilibrium that is Pareto optimal. We call these *Simple Neoclassical Models*. This class of models is the foundation of neoclassical macroeconomics, and provides the most transparent description of how competitive market forces operate within a dynamic, general equilibrium environment.

In contrast to common perceptions about neoclassical macroeconomics, we acknowledge that economies are affected by policy distortions and other market imperfections that go beyond the scope of simple models. The second class of models modifies simple models as needed to incorporate changes that require departing from the model assumptions described above. We call the second class of models *Extended Neoclassical Models*, which are constructed by building explicit specifications of government policies or market imperfections and distortions into simple models.

This method nests simple models as special cases of the extended models. Developing complex models in this fashion provides a clear description of how market imperfections and economic policies affect what otherwise would be a *laissez-faire* market economy. We modify the models in very specific ways that are tailored to study episodes in US economic history, and which provide researchers with frameworks that can be applied more broadly. All of the models presented in this chapter explicitly treat fluctuations and growth within the same framework.

Neoclassical frameworks are a powerful tool for analyzing market economies. An important reason is because the US economy has displayed persistent and reasonably stable growth over much its history while undergoing enormous resource reallocation through the competitive market process in response to changes in technologies and government policies. These large reallocations include the shift out of agriculture into manufacturing and services, the shift of economic activity out of the Northern and Mideastern sections of the United States to the Southern and Western states, and large changes in government's share of output, including changes in tax, social insurance, and regulatory labor policies. This also includes the reallocation of women's time from home production to market production, and the increased intensity of employment of highly-skilled labor. Most recently, this has included the reallocation of resources out of the development of mature, mechanical technologies to the development of information processing and communication technologies, including the integrated circuit, fiber optics, microwave technology, laptop computers and tablets, software applications, cellular technology, and the internet.

Our focus on technologies and policies connects with considerable previous research. This ranges from Schumpeter (1927) and Stock and Watson (1988), who argued that changes in entrepreneurship and the development of new ideas are the primary drivers of a market economy, to Kydland and Prescott (1982) and Long Jr and Plosser (1983), who focused on technology shocks and fluctuations. This also includes Lilien (1982), who argued that sectoral shifts significantly affect fluctuations and resource reallocation, Davis and Haltiwanger (1992), who established that resource reallocation across US manufacturing establishments is very large and is continuously evolving, and Greenwood and Yorokoglu (1997) and Manuelli and Seshadri (2014), who analyze the diffusion of new technologies and their long-run economic effects. The analysis also connects with studies of the long-run consequences of government policies, including research by Ljungqvist and Sargent (1998), Prescott (2004), and Rogerson (2008), who analyze how public policies such as tax rate changes, and changes in social insurance programs, have affected long-run labor market outcomes.

Our principle of focusing on long-run movements in data requires a quantitative approach that differs from standard practice in macroeconomics that involves both the selection of the data frequencies that are analyzed, and how the model is compared to data. The standard approach removes a trend from the data that is constructed using the Hodrick–Prescott (HP) filter (1997), hereafter referred to as HP filter, with a smoothing parameter of 1600, and then typically compares either model moments to moments from the HP-filtered data, or compares model impulse response functions to those from an empirical vector autoregression (VAR).This analysis uses a band pass filter to quantify movements not only at the HP-business cycle frequency, but also at the lower frequencies. Our quantitative-theoretic analysis evaluates model economies by conducting equilibrium path analyses, in which model-generated variables that are driven by identified shocks are compared to actual raw data and to filtered data at different frequencies.

We report two sets of findings. We first document the empirical importance of very long-run movements in aggregate variables relative to traditional business cycle fluctuations using post-Korean War quarterly US data, long-run annual US data, and postwar European data. We find that low frequency movements in aggregate time series are quantitatively large, and that in some periods, they are much larger than the traditional business cycle component. Specifically, we analyze movements in periodicities ranging from 2 to 50 years, and we find that as much as 80% of the fluctuations in economic activity at these frequencies is due to the lower frequency component from 8 to 50 years.

The dominant low frequency nature of these data indicates that the business cycle literature has missed quantitatively important movements in aggregate activity. Moreover, the fact that much of the movement in aggregate data is occurring at low frequencies suggests that models that generate fluctuations from transient impediments to trade, such as temporarily inflexible prices and/or wages, may be of limited interest in understanding US time series.

The importance of low frequency movements also has significant implications for the two dominant episodes of the last 35 years, the *Great Moderation* and the *Great Recession*.

The Great Moderation, the period of stable economic activity that occurred between 1984 and 2008, features a sharp decline in volatility at the traditional business cycle frequency, but little volatility change at low frequencies. Similarly, the Great Recession and its aftermath feature a large, low frequency component. These data suggest that the Great Recession was not just a recession per se. Instead, much of this event appears to be a persistent decline in aggregate economic activity.

Following the decomposition of data into low and high frequency components, we report the results of quantitative-theoretic analyses that evaluate how well neoclassical models account for the US historical macroeconomic record from 1929 to 2014.

Our main finding is that neoclassical models can account for much of the movement in aggregate economic activity in the US economic historical record. Neoclassical models plausibly account for major economic episodes that previously were considered to be far beyond their reach, including the Great Depression and World War II. We also find that neoclassical models account for much of the post-Korean War history of the United States.

The chapter is organized as follows. Section 2 presents the United States and European data that we use in this study, and provides a decomposition of the data into low frequency and business cycle frequency components. Section 3 introduces the basic neoclassical macroeconomic model that serves as the foundation for all other models developed in the chapter. Section 4 presents one-, two-, and three-sector *simple neoclassical model* analyses of the post-Korean War US economy. Section 5 presents *extended neoclassical models* to study Depressions. Section 6 presents *extended neoclassical models* with fiscal policies with a focus on the US economy during World War II. Given the importance of productivity shocks in neoclassical models, Section 7 discusses different frameworks for understanding and interpreting TFP changes. Given the recent interest in economic inequality, Section 8 discusses neoclassical models of wage inequality. Section 9 presents a critical assessment of neoclassical models, and suggests future research avenues for neoclassical macroeconomic analysis. Section 10 presents our conclusions.

2. THE IMPORTANCE OF LOW FREQUENCY COMPONENTS IN MACROECONOMIC DATA

It is common practice in applied macroeconomics to decompose time series data into specific components that economists often refer to as *cyclical components, trend components*, and *seasonal components*, with the latter component being relevant in the event that data are not seasonally adjusted. These decompositions are performed to highlight particular features of data for analysis. The most common decomposition is to extract the cyclical component from data for the purpose of business cycle analysis, and the HP filter is the most common filtering method that is used.

Band-pass filters, which feature a number of desirable properties, and which resolve some challenges involved with applying the HP filter, are increasingly being used to filter data.[a] Band-pass filtering allows researchers to choose components that correspond to periodicities over a specific data frequency. An exact band pass filter requires an infinite length of data, so Baxter and King (1999) and Christiano and Fitzgerald (2003) have constructed approximate band pass filters. These two approaches are fairly similar. The main difference is that the Baxter–King filter is symmetric, and the Christiano–Fitzgerald filter is asymmetric.

This section presents decompositions of aggregate data into different frequency components for (i) US post-Korean War quarterly data, (ii) US annual data that extends back to 1890, and (iii) post-World War II annual European data. We use the Baxter–King filter, given its wide use in the literature. The band pass filter isolates cyclical components in data by smoothing the data using long moving averages of the data. Baxter and King develop an approximate band pass filter that produces stationary data when applied to typical economic time series.[b] Since the exact band pass filter is an infinite order process, Baxter and King construct a symmetric approximate band pass filter. They show that the optimal approximating filter for a given maximum lag length truncates the filter weight at lag K as follows:

$$y_t^* = \sum_{k=-K}^{K} a_k y_{t-k} \tag{1}$$

In (1), y^* is the filtered data, y is the unfiltered data, and the a_k denote coefficients that produce the smoothed time series. The values of the a_k coefficients depend on the filtering frequency (see Baxter and King, 1999).

Following early work on business cycles by Burns and Mitchell (1946), Baxter and King study business cycles, which they define as corresponding to periodicities associated with 6–32 quarters. In contrast, we use the band-pass filter to consider a much broader range of frequencies up to 200 quarters. Our choice to extend the frequency of analysis to 200 quarters is motivated by Comin and Gertler (2006), who studied these lower frequencies in a model with research and development spending.

We consider much lower frequencies than in the business cycle literature since changes in technologies and government policies may have a quantitatively important effect on low frequency movements in aggregate data. Relatively little is known about the nature and size of these low frequency fluctuations, however, or how these low frequency fluctuations compare to business cycle fluctuations. We therefore band-pass filter data between 2 and 200 quarters, and we split these filtered data into two components:

[a] In terms of the challenges with the HP filter, it is not clear how to adjust the HP smoothing parameter to assess data outside of the cyclical window originally studied by Hodrick and Prescott (1997). Moreover, HP-filtered data may be difficult to interpret at data endpoints.

[b] The Baxter–King filter yields stationary time series for a variable that is integrated of up to order two. We are unaware of any macroeconomic time series that is integrated of order three or higher.

a 2–32 quarters component, which approximates the business cycle results from the standard parameterization of the HP filter ($\lambda = 1600$), and a 32–200 quarters component. This allows us to assess the relative size and characteristics of these fluctuations. To our knowledge, these comparative decompositions have not been constructed in the literature.

2.1 Band-Pass Filtered Quarterly US Data

This section analyzes US quarterly post-Korean war data from 1954 to 2014, which facilitates comparison with much of the business cycle literature. We then analyze annual US data extending back to 1890, followed by an analysis of postwar European data.[c]

Figs. 1–6 show filtered real GDP, consumption of nondurables and services, gross private domestic investment, hours worked, total factor productivity (TFP), and the relative price of capital equipment. Real GDP, consumption, and investment are from the NIPA.

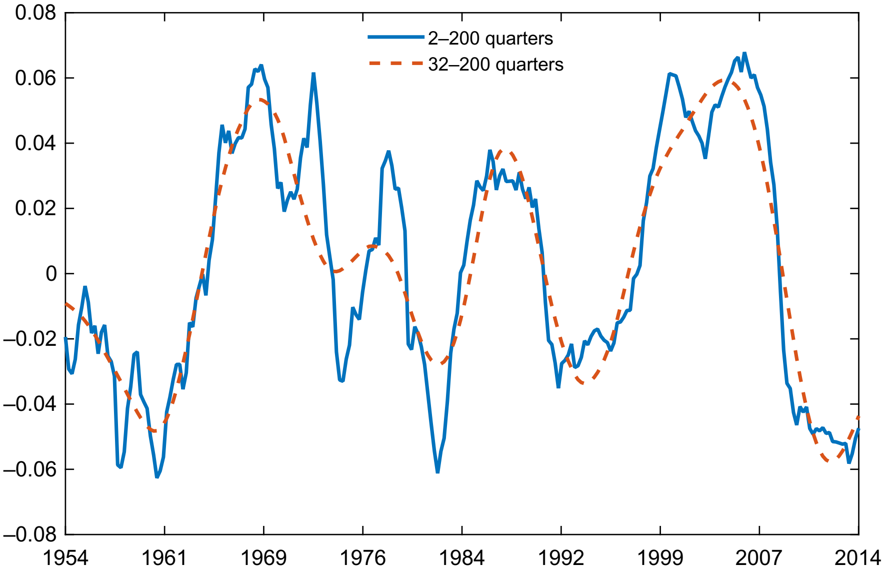


Fig. 1 Log of real GDP.

[c] The Baxter–King filter loses data at the beginning and the end of a dataset. We therefore padded all the data series at both the starting and ending dates by simulating data from ARMA models fit to each series. These simulated data extend the series before the starting date and after the end date, which allows us to construct filtered data for the entire period length. We conducted a Monte Carlo analysis of this padding procedure by generating extremely long artificial time series, and comparing band-pass filtered series using the padded data, to filtered data that doesn't use padding. The length of the data padding is equal to the number of moving average coefficients, k. We use $k = 50$ for the quarterly data, and $k = 12$ for the annual data. The results were insensitive to choosing higher values of k.

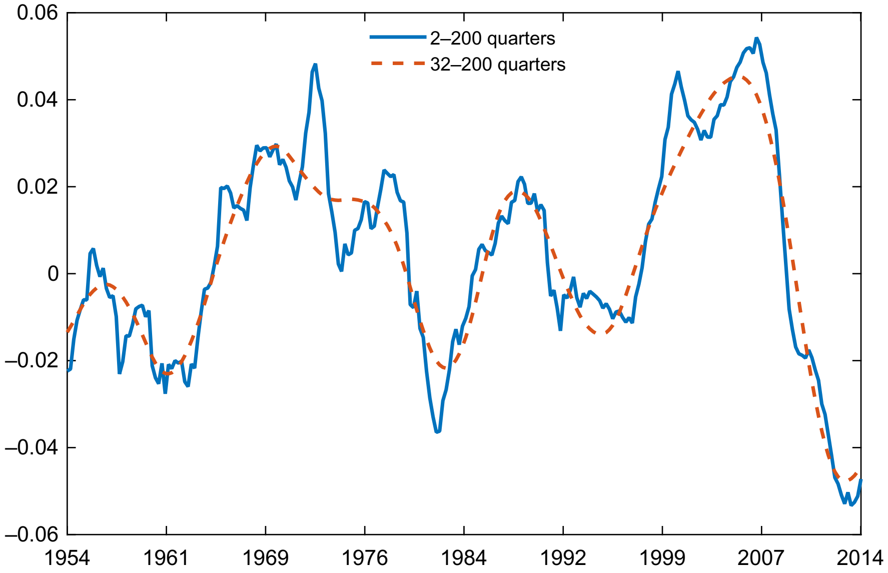


Fig. 2 Log of consumption of nondurables and services.

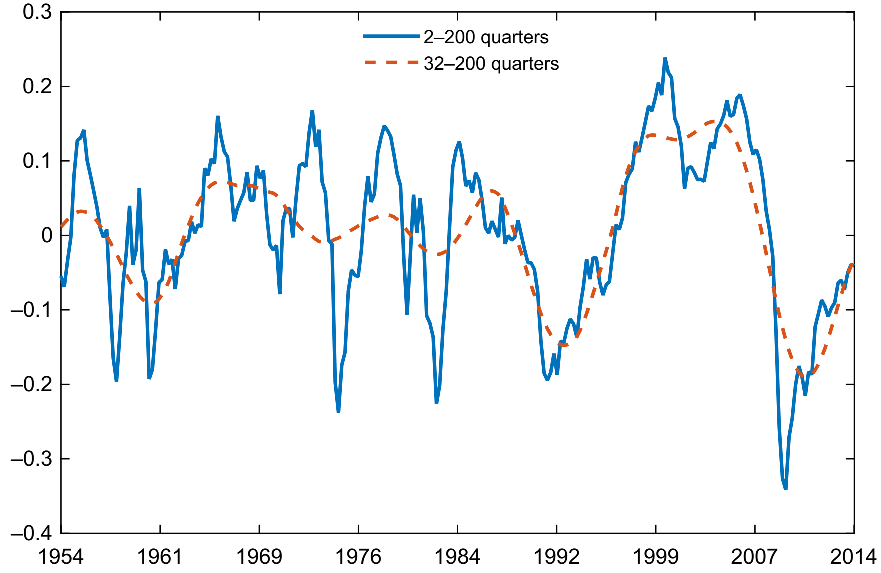


Fig. 3 Log of fixed investment.

Hours worked is constructed by updating the hours worked data of Cociuba et al. (2012), who use hours from the Current Population Survey. TFP is constructed by dividing real GDP by a Cobb–Douglas aggregate of capital, which is the sum of private and public capital stocks, and which has a share of 0.4, and hours worked, which has a share of 0.6.

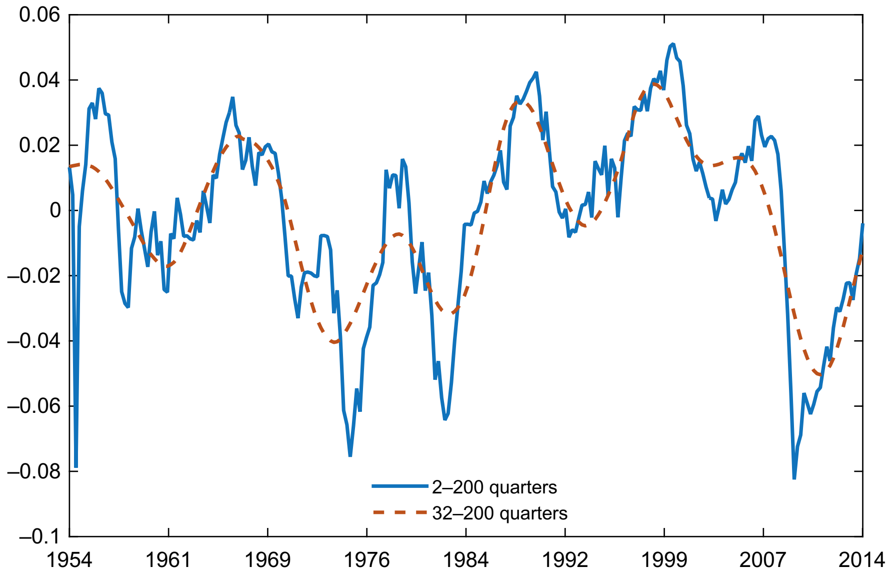


Fig. 4 Log of total hours worked.

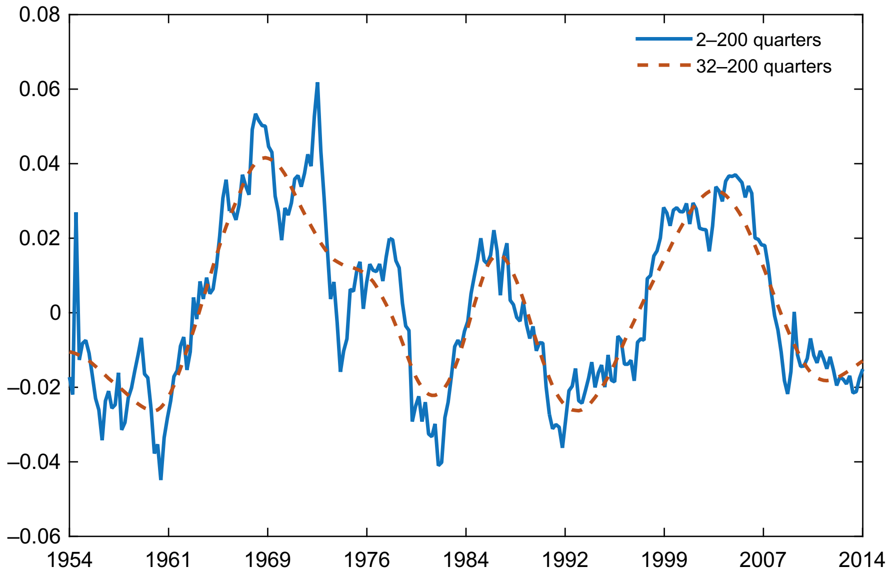


Fig. 5 Log of total factor productivity.

We include the relative price of capital equipment in this analysis because there is a large change in this relative price over time, and because the inverse of this relative price is a measure of equipment-specific technological change in some classes of models, including Greenwood et al. (1997) and Krusell et al. (2000). We construct the relative price of

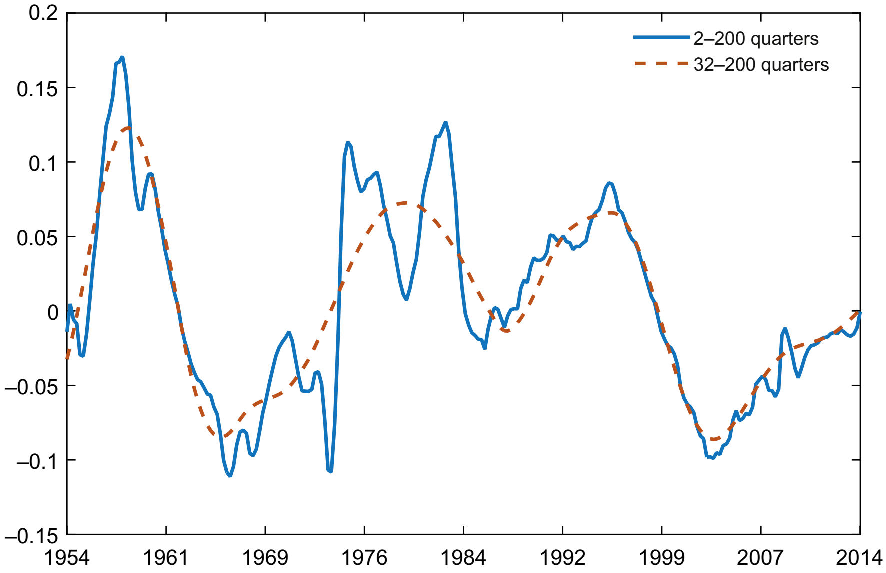


Fig. 6 Log of relative price of equipment.

equipment as the ratio of the quality-adjusted deflator for producer durable equipment, to the NIPA nondurable consumption deflator. Gordon (1990) initially constructed the quality-adjusted equipment deflator, and this time series has been continued in Cummins and Violante (2002) and in DiCecio (2009).[d]

The figures show the 2–200 component and the 32–200 component. Since the band pass filter is a linear filter, the difference between these two lines is the 2–32 component. The most striking feature of all of these filtered data is that much of the movement in the 2–200 component is due to the 32–200 component. These filtered data indicate that business cycle variability, as typically measured, accounts for a relatively small fraction of the overall post-Korean war history of US economic variability. The graphs do show that there are some periods in which the traditional business cycle component is sizeable. This occurs during part of the 1950s, which could be interpreted as the economy readjusting to peacetime policies following World War II and the Korean War. There is also a significant 2–32 component from the 1970s until the early 1980s.

[d] We do not use the NIPA equipment deflator because of Gordon's (1990) argument that the NIPA equipment price deflator does not adequately capture quality improvements in capital equipment. We use DiCecio's (2009) updating of the Gordon–Cummins–Violante data. This data is updated by DiCecio on a real time basis in the Federal Reserve Bank of St. Louis's FRED database (https://research.stlouisfed.org/fred2/series/PERIC). The mnemonic for this series is PERIC.

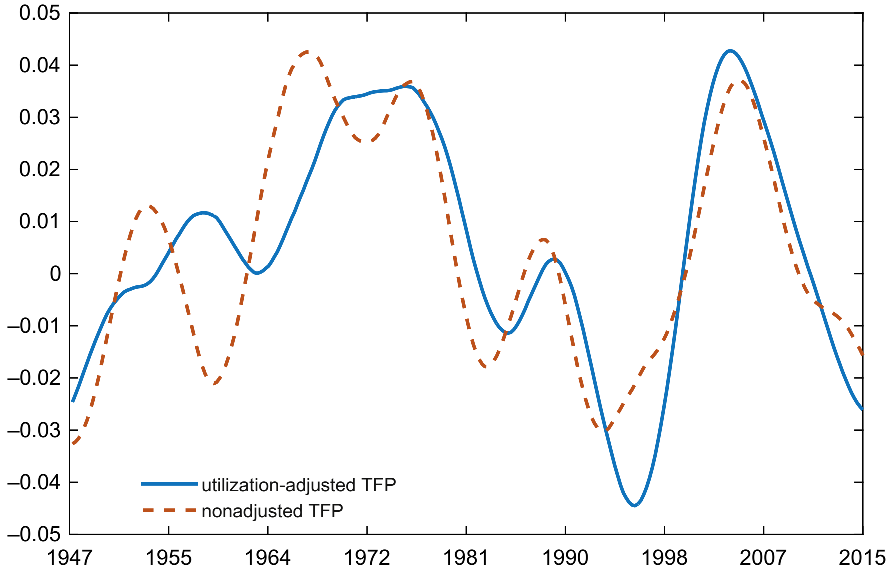


Fig. 7 Fernald TFP (filtered 32–200 quarters).

The 32–200 component of TFP has important implications for the common critique that TFP fluctuations at the standard HP frequency are affected by unmeasured cyclical factor utilization. Fernald's (2014) TFP series is a widely used measure of TFP that is adjusted for unmeasured factor utilization. Fig. 7 shows the 32–200 component of Fernald's adjusted and unadjusted measures of business sector TFP. The long-run components of the adjusted and unadjusted series are very similar, particularly over the last 40 years. This indicates that unmeasured factor utilization is not an issue for measuring TFP at these lower frequencies.

To quantify the relative contribution of the 32–200 component for these variables, we construct the following ratio, which we denote as z_i, in which x_i is the 32–200 filtered component of variable i, and y_i is the 2–200 filtered component of variable i :

$$z_i = \sum_t \frac{(x_{it})^2}{(y_{it})^2} \tag{2}$$

On average, the 32–200 component accounts for about 80% of the fluctuations in output, consumption, TFP, and the relative price of equipment and about 64% of hours. It accounts for about 56% of fluctuations in gross private domestic investment, which includes the highly volatile category of inventory change.

The 32–200 component is also large during the Great Moderation. Specifically, the well-known volatility decline of the Great Moderation, which is typically dated from

1984 to 2007, is primarily due to lower volatility of the 2–32 component. The figures show that the volatility of the 32–200 component remains quantitatively large during the Great Moderation. This latter finding may reflect the large and persistent technological advances in information processing and communications that occurred throughout this period.

This finding regarding the nature of these frequency components in the Great Moderation is consistent with the conclusions of Arias et al. (2007) and Stock and Watson (2003), who report that the traditional business cycles frequency shocks that affected the economy during this period were smaller than before the Great Moderation. This finding about the Great Moderation may also reflect more stable government policies that reduced short-run variability. Taylor (2010) has argued that more stable monetary policy is important for understanding the Great Moderation.

The 32–200 component is also important for the Great Recession and its aftermath. This largely reflects the fact that there has been limited economic recovery relative to long-run trend since the Great Recession.

2.2 Band-Pass Filtered Annual US and European Data

This section presents band-pass filtered annual long-run US data and annual European data. The output data were constructed by splicing the annual Kuznets–Kendrick data (Kendrick, 1961) beginning in 1890, with the annual NIPA data that begins in 1929. The annual Kendrick hours data, which also begins in 1890, is spliced with our update of the hours worked data from Cociuba et al. (2012). These constructions provide long annual time series that are particularly useful in measuring the low frequency components.

Figs. 8 and 9 show the filtered annual US data. The low frequency component, which is measured using the band pass filter from 8 to 50 years for these annual data, is also very large. Extending the data back to 1890 allows us to assess the importance of these different components around several major events, including the Panic of 1907 and World War I. The data show that both the Depression and World War II were dominated by lower frequency components, while the traditional business cycle component was significant during World War I and the Panic of 1907.

The large low frequency component of World War II stands in contrast to World War I, and also stands in contrast to standard theoretical models of wartime economies. These models typically specify wars as a highly transient shock to government purchases. The low frequency component is also large for the Great Depression. Sections 5 and 6 develop neoclassical models of Depressions and of wartime economies, in which both of these events are driven by persistent changes in government policies.

The decomposition ratio presented in (2), and that was used to construct the share of variation in the 2–200 quarter component due to the 32–200 quarter component, is used in a similar way to construct the share of variation in the 2–50 year component due to the

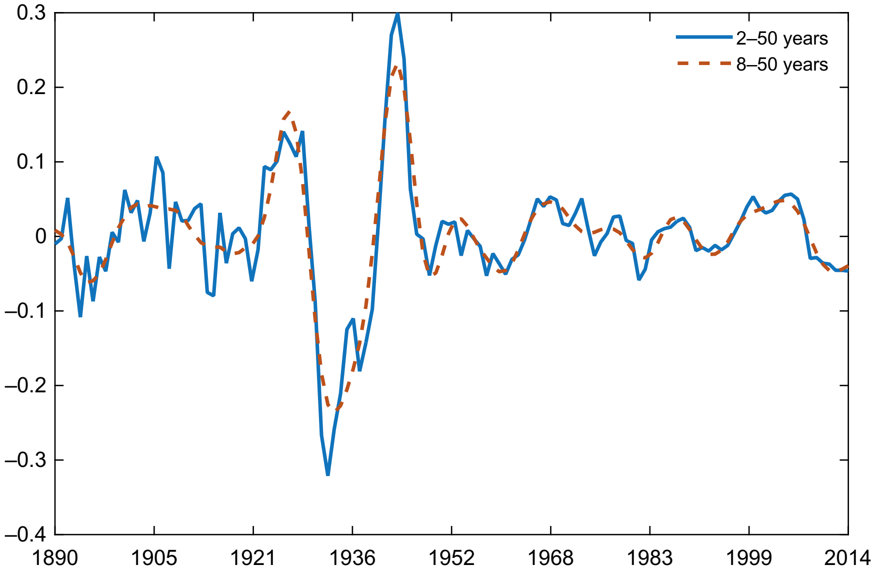


Fig. 8 Annual log of real GDP.

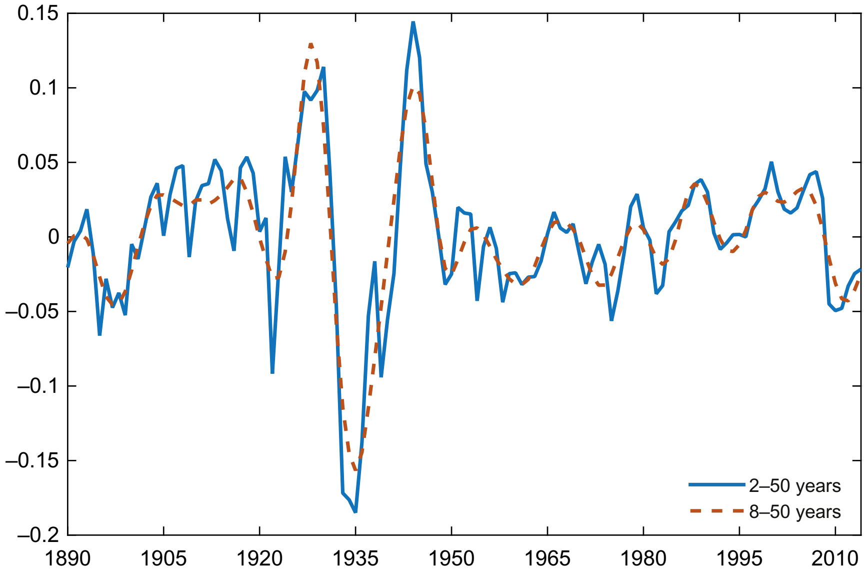


Fig. 9 Annual log of hours worked.

8–50 year component. This low frequency component share is also large in the annual data, ranging between 80% and 85% for real GNP and hours worked.

We also construct the decomposition using annual postwar logged real output data from several European economies: Germany, France, Italy, Spain, and Sweden.

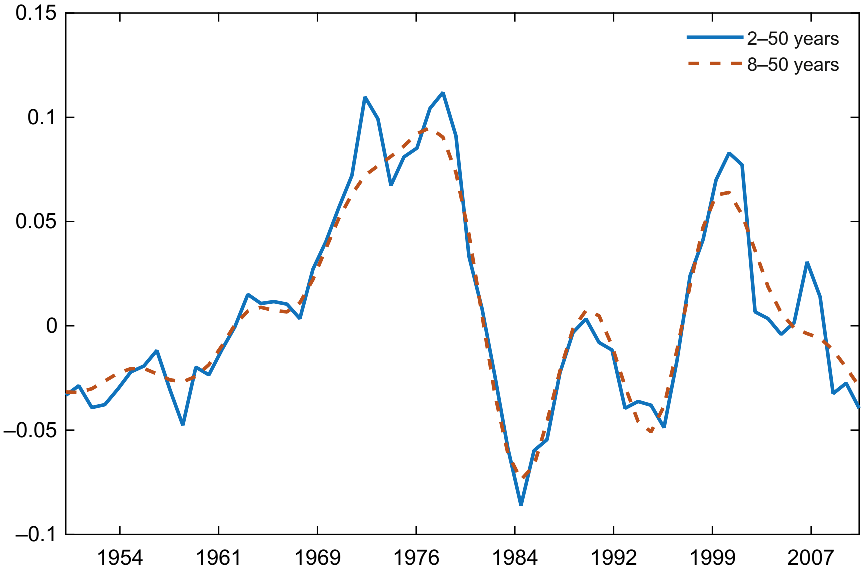


Fig. 10 Log of real GDP—France.

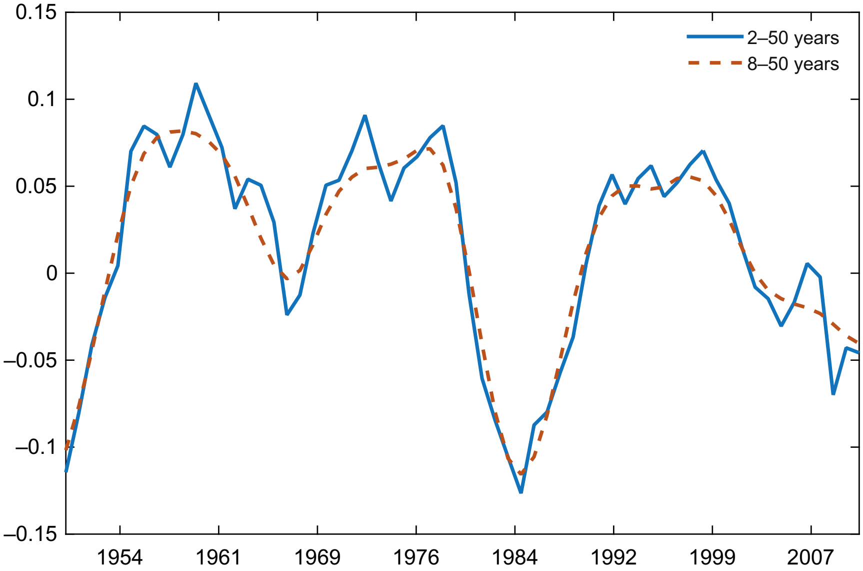


Fig. 11 Log of real GDP—Germany.

These data are from the Penn World Tables (Feenstra et al., 2015). Figs. 10–14 present the filtered data. Most of the variation in the European output data in the 2–50 year component also is accounted for by the low frequency (8–50) component. The long-run European components reflect clear patterns in these data. All of the European economies

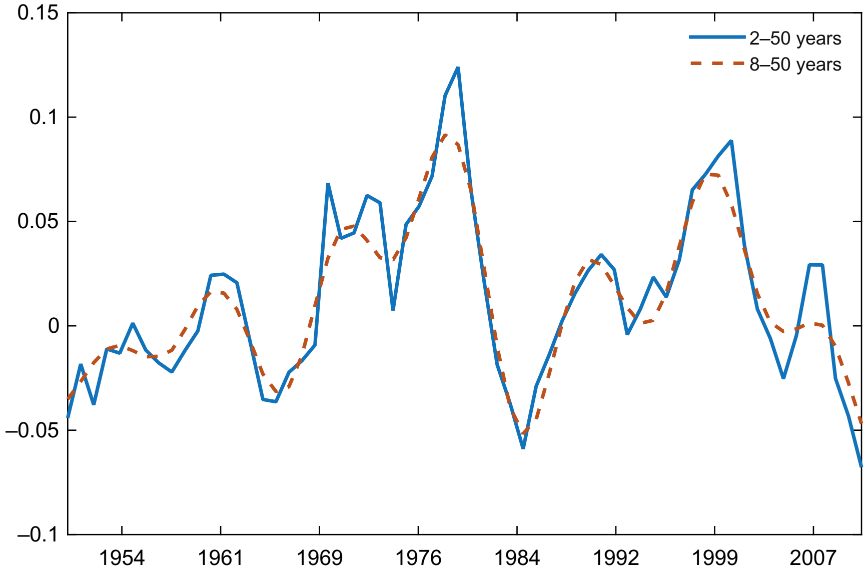


Fig. 12 Log of real GDP—Italy.

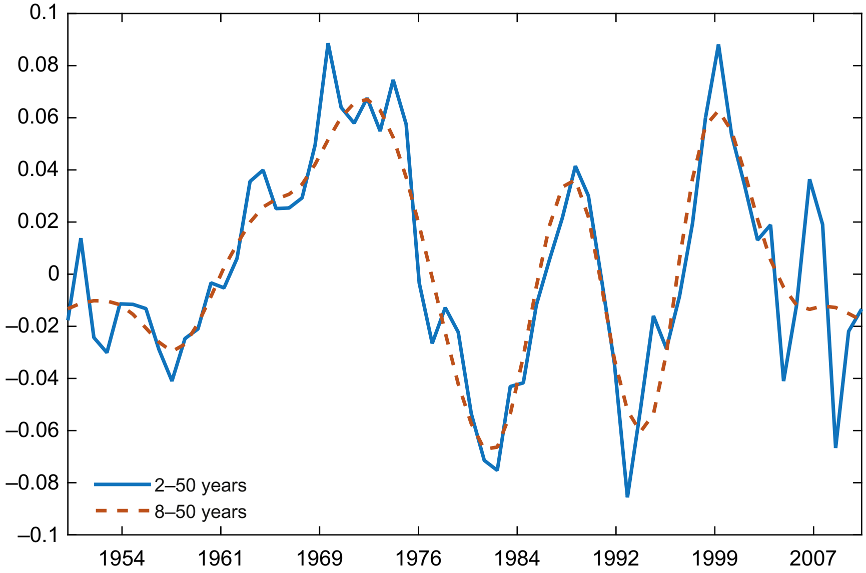


Fig. 13 Log of real GDP—Spain.

grow more rapidly than the US during the 1950s and 1960s. All of these economies then experience large declines relative to trend that begin in the early 1970s and continue to the mid-1980s. The share of the 2–50 component that is accounted for by the 8–50 component is about 80% for Germany, France, Spain, and Sweden, and is about 71% for Italy.

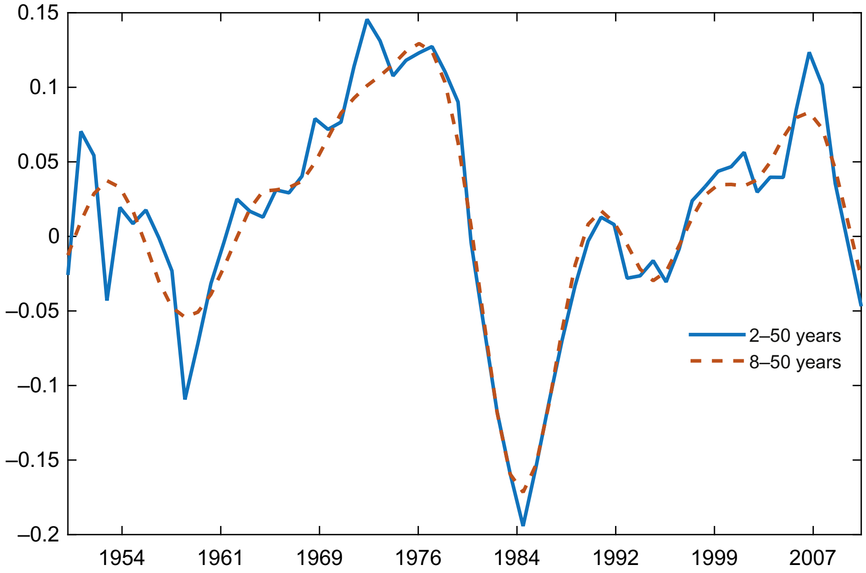


Fig. 14 Log of real GDP—Sweden.

2.3 Alternative to Band-Pass Filtering: Stochastic Trend Decomposition

This section presents an alternative decomposition method, known as stochastic trend decomposition, for assessing the relative importance of low frequency components. One approach to stochastic trend decompositions was developed by Beveridge and Nelson (1981), and is known as the Beveridge–Nelson decomposition. Watson (1986) describes an alternative approach, which is known as unobserved components model decomposition. In both frameworks, a time series is decomposed into two latent objects, a stochastic trend component, and a stationary component, which is often called the cyclical component.

Decomposing the time series into these latent components requires an identifying restriction. The Beveridge–Nelson identifying restriction is that the two components are perfectly correlated. This identifying assumption is thematically consistent with our view that permanent changes in technologies and policies generate both stationary and permanent responses in macroeconomic variables.[e]

[e] The unobserved components models have traditionally achieved identification of the two latent components by imposing that the trend and stationary components are orthogonal. More recently, Morley et al. (2003) show how to achieve identification in unobserved components models with a nonzero correlation between the two components. Morley et al. find that the decomposition for real GDP for their unobserved components model is very similar to the Beveridge–Nelson decomposition. They also present evidence that the zero correlation identifying restriction that traditionally has been used in unobserved components models is empirically rejected.

The Beveridge–Nelson decomposition, which is simple and widely used, is applied in this chapter. The Beveridge–Nelson statistical model begins with a variable that is assumed to have a stochastic trend component. The variable may also have a drift term, which drives secular growth in the variable. The Beveridge–Nelson decomposition removes the drift term, and then decomposes the variable, which we denote as y_t, into a stochastic trend component, x_t and a stationary stochastic component, s_t. The stochastic trend is a random walk, and the innovation term, which is denoted as ε_t, is a white noise process:

$$y_t = x_t + s_t \tag{3}$$

$$x_t = x_{t-1} + \varepsilon_t, E(\varepsilon) = 0, E(\varepsilon^2) = \sigma_\varepsilon^2 \tag{4}$$

This decomposition is applied to the log of US real GDP. The decomposition first requires specifying an ARIMA model for the data. We selected an ARIMA (0,1,1) model for the log of real GDP, given that the first three autocorrelations of the first difference of the logged data are 0.34, 0.19, and 0.06. Stock and Watson (1988) also use this ARIMA specification for the log of real output. The estimated statistical model for the log of real GDP using quarterly data between 1954:1 and 2013:4 is given by:

$$\Delta \ln(GDP_t) = 0.0077 + \varepsilon_t + 0.40\varepsilon_{t-1}. \tag{5}$$

These estimated coefficients are similar to the Stock and Watson estimates that were based on a shorter dataset. Stock and Watson estimated a slightly higher drift term of about 0.008, and a somewhat smaller moving average coefficient of 0.30 rather than 0.40.

Using the Wold decomposition, Beveridge and Nelson show that the permanent component for this estimated statistical model is given by:

$$1.4 * \sum_{j=1}^{t} \varepsilon_j \tag{6}$$

Fig. 15 plots the detrended log of real GDP, which is constructed as the log of real GDP less its accumulated drift component, and the Beveridge–Nelson permanent component of these detrended data. The figure shows that almost all of the movement in detrended real GDP is due to the permanent component, rather than the transitory component. This finding is consistent with the band-pass filtered results regarding the large size of the long-run component.

The results presented in this section show that the bulk of observed fluctuations in aggregate time series are from longer-run changes than those associated with traditional business cycle frequencies. This finding motivates our focus on neoclassical models that are driven by long-run changes in technologies and policies, as opposed to models that are driven by very transient shocks, such as monetary shocks that operate in models with temporarily inflexible prices and/or wages.

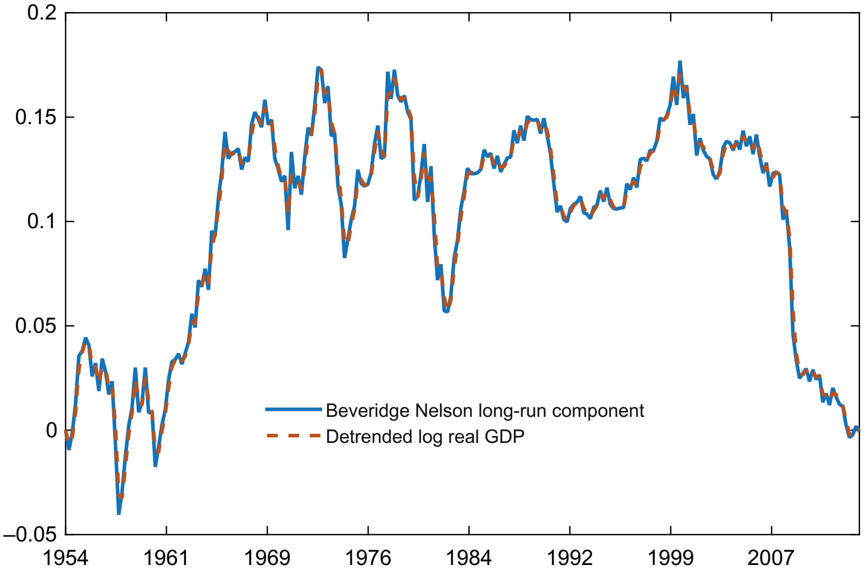


Fig. 15 Beveridge–Nelson decomposition of real GDP.

3. CASS-KOOPMANS: THE FOUNDATION OF SIMPLE MODELS

This section summarizes the one-sector Cass-Koopmans optimal growth model with elastically supplied leisure, as it serves as the foundation for the other models that are developed in this chapter. This model features (1) standard utility maximization problems for households, and standard profit maximization problems for firms, both of whom behave competitively and who have rational expectations, (2) complete markets, (3) a unique and Pareto optimal equilibrium, and (4) constant returns to scale technology.

Since the welfare theorems hold in this economy, we express this model as a social planning problem. For heuristic purposes, we assume perfect foresight. The planner's maximization problem is given by:

$$\max \beta^t \sum_{t=0}^{\infty} u(c_t, l_t). \tag{7}$$

Maximization is subject to the economy's resource constraint, a household time constraint, a transition equation for the capital stock, and nonnegativity constraints on consumption, hours, and capital:

$$f(k_t, h_t) \geq c_t + i_t \tag{8}$$

$$1 \geq h_t + l_t \tag{9}$$

$$k_{t+1} = (1-\delta)k_t + i_t \tag{10}$$

$$c_t \geq 0, h_t \geq 0, k_t \geq 0, k_0 \text{ given.} \tag{11}$$

It is also necessary to impose the transversality condition to rule out explosive paths for the capital stock:

$$\lim_{t \to \infty} \beta^t u_1(c_t, l_t) f_1(k_t, h_t) k_t = 0 \tag{12}$$

The utility function satisfies the usual restrictions: it is concave in its arguments and twice continuously differentiable. The technology, *f*, is constant returns to scale in the two inputs capital, *k*, and labor, *h*, and is also twice continuously differentiable.

We will tailor the construction of different neoclassical models to focus on policies and technological change that we highlight for specific historical episodes. This should not be confused with the idea that fundamentally different models are needed to address different time periods in the history of the US economy. Rather this means that the relative importance of different policies and different types of technological change has varied over time. Specifically, this includes the importance of biased technological change for understanding the post-Korean War US history, cartelization and unionization government policies for understanding the 1930s, and changes in government fiscal policies for understanding the 1940s.

4. NEOCLASSICAL MODELS OF THE US POST-KOREAN WAR ECONOMY

In this section we present a series of neoclassical models, driven by permanent changes in technologies to study the post-Korean War US economy. Our approach, which we describe in detail below, compares the equilibrium paths of the model economies in response to identified shocks, to the actual time series data. We will compare model results to unfiltered data, and also to the three different filtering frequencies described in Section 2. In addition to evaluating the fit of the model for the raw data, this will allow us to assess how well the model matches data at the traditional business cycle frequencies (2–32 quarters), and also at low frequencies (32–200) quarters.

4.1 Quantitative Methodology

Neutral technological change that affects all sectors identically is the standard specification of technology in neoclassical macroeconomic models. However, there is a growing body of evidence that technological change is advancing much more quickly in the information processing sectors of the economy, particularly in capital equipment. This includes the areas of computer hardware, computer peripherals, photocopying equipment and telecommunications equipment, among others.

As described earlier in this chapter, Gordon (1990), Cummins and Violante (2002), and DiCecio (2009) construct capital equipment price data that they argue captures much more of the quality change that has occurred in these goods than is present in the NIPA equipment price data. Fig. 16 shows the relationship between real GDP

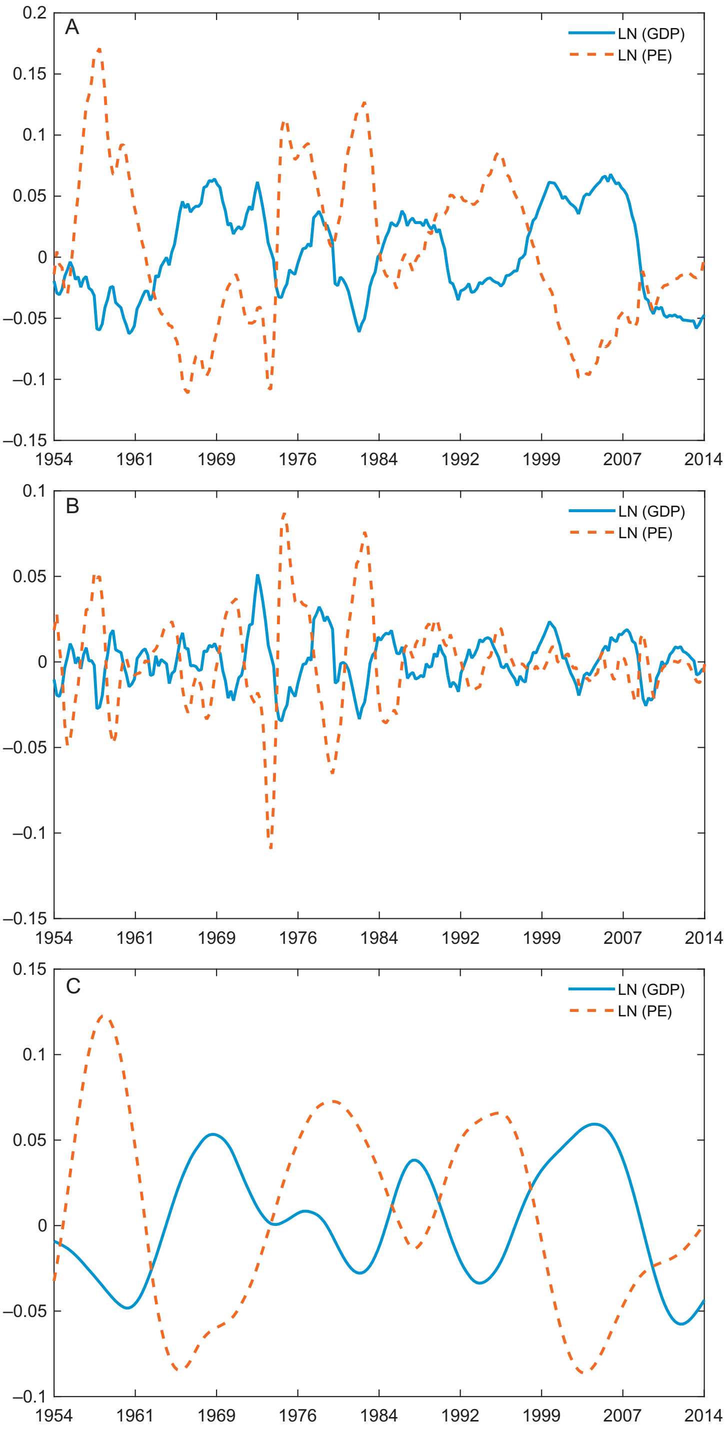


Fig. 16 Filtered GDP and the relative price of equipment. (A) 2–200 quarters. (B) 2–32 quarters. (C) 32–200 quarters.

and the relative price of equipment at the three sets of frequencies that we consider. These figures show that the relative price of equipment is strongly countercyclical at all frequencies.

These strong countercyclical patterns are interesting as a growing number of neoclassical studies are using these data to identify capital-equipment specific technological change. The following sections develop multisector growth models that include both neutral and equipment-specific technological change to study the evolution of the post-Korean War US economy. This is a particularly interesting period for applying multisector models with biased technological change since this period features a number of major advances in information processing and telecommunications technologies, including the integrated circuit, personal computers and tablet technologies, fiber optics, software applications, cellular technologies, and the internet.

Focusing on this period also allows us to connect this analysis with the large business cycle literature, including Kydland and Prescott (1982), Hansen (1985), and the studies in Cooley (1995), which have analyzed the post-Korean War US economy. Note that the post-Korean War period also includes a number of interesting subperiods: the Vietnam War (1957–71), the oil shock years (1974–81), the Great Moderation (1984–2007), and the Great Recession and its aftermath (2008–present).

Our quantitative approach differs from the standard approach used in the real business cycle literature. The real business cycle approach specifies a dynamic stochastic general equilibrium model, which includes a specification of the stochastic process for the exogenous shocks that generate fluctuations in the model economy. The equilibrium decision rules and laws of motion are computed using numerical methods, and these equations plus a random number generator are used to simulate time series for the artificial economy. Summary statistics are then computed and compared with the same summary statistics computed from actual US time series.

The approach we follow is similar to that employed in Hansen and Prescott (1993). We begin with a two-sector growth model in which movements in aggregate time series are the result of two factors we identify from US data that we take to be the exogenous forcing processes in the model. These include technology shocks that are identified with total factor productivity and equipment specific technological change, which we identify from the relative price of equipment. We then calibrate and solve the model in a manner consistent with the real business cycle literature. But, rather than drawing random realizations of the exogenous shock processes, we identify time paths for our two technology shocks from US time series data. We then compute the equilibrium time paths for the endogenous variables (output, consumption, investment and hours worked) using the actual time path of the exogenous shocks. As noted above, we compare model variables to quarterly real variables for the unfiltered data over 1954–2014, as well as for frequency bands corresponding to 2–200, 2–32, and 32–200 quarters.

After comparing the time paths from the two-sector model with the corresponding time paths from US data, we then compare these time paths with those of a standard one-sector neoclassical model in which neutral technology shocks are the only exogenous process hitting the economy. We then consider a three-sector model that adds a nonmarket home production sector to our baseline two-sector model. This extension allows us to study how equipment biased technological change may have induced movements in labor from the home production sector to the market sector.

We omit the details of numerically solving these models. Instead, we focus on the specifics of the model economies, the construction of US data counterparts to the model variables, and the calibration that we use in our computational analyses.

In terms of assessing model fit, our approach differs considerably from the recent approach that is used in the New Keynesian literature. In New Keynesian models, such as Smets and Wouters (2007), as many shocks are added to the model as needed so that the model fits all of the data very closely. While this approach delivers a very good model fit, some of the shocks in the model are often difficult to interpret. Our approach to model fit follows from our theme that permanent changes in technologies are key drivers of the economy. The models analyzed in the following sections have very few shocks, which allows us to transparently evaluate the models' successes and deviations.

4.2 A Two-Sector Model with Aggregate and Investment-Biased Technological Change

This section develops a model with investment-specific technological change, as well as aggregate technological change that impacts all sectors equally. This approach was first developed in Greenwood et al. (1997), who document and discuss investment-specific technological change and its impact on long-run growth. Biased technological change has also been used to study wage inequality (Krusell et al., 2000) and business cycles (Fisher, 2006; Justiniano et al., 2010).

The two-sector stochastic growth model we study consists of a primary sector, $i=1$, producing C_{Mt}, which is the sum of consumer services, nondurable consumption and government consumption, and I_{st}, which is investment in structures.[f] The second sector, $i=2$, produces equipment I_{et} and consumer durables I_{dt}. The technologies associated with each sector are as follows:

$$C_{Mt}+I_{st}=Y_{1t}=z_tAK_{e1t}^{\theta_1}K_{s1t}^{\theta_2}H_{1t}^{1-\theta_1-\theta_2} \tag{13}$$

$$I_{dt}+I_{et}=Y_{2t}=q_tz_tAK_{e2t}^{\theta_1}K_{s2t}^{\theta_2}H_{2t}^{1-\theta_1-\theta_2} \tag{14}$$

All variables are measured in per capita terms with a population growth factor η. Here, K_{eit},K_{sit} and H_{it} are equipment, structures and hours worked, each in sector i.

[f] We will also lump investment in intellectual property with investment in structures.

The variables z_t and q_t are technology shocks that impact these sectors. The laws of motion for the stocks of equipment, structures, and durables is given by the following, where $K_{e,t} = K_{e1t} + K_{e2t}$ and $K_{s,t} = K_{s1t} + K_{s2t}$:

$$\eta K_{e,t+1} = (1-\delta_e)K_{et} + I_{et} \tag{15}$$

$$\eta D_{t+1} = (1-\delta_d)D_t + I_{dt} \tag{16}$$

$$\eta K_{s,t+1} = (1-\delta_s)K_{st} + I_{st} \tag{17}$$

The logarithms of the two shocks, z and q, follow random walks with drift.

$$\log z_{t+1} = \log z_t + \varepsilon_{1,t+1}\ ,\ \varepsilon_1 \sim N(\mu_1, \sigma_1^2) \tag{18}$$

$$\log q_{t+1} = \log q_t + \varepsilon_{2,t+1}\ ,\ \ \varepsilon_2 \sim N(\mu_2, \sigma_2^2) \tag{19}$$

The random variables ε_1 and ε_2 are i.i.d. across time and are contemporaneously uncorrelated.

There is a stand-in household who maximizes the expected discounted sum of utility defined over consumption of nondurables and services, the stock of durables, and leisure:

$$\max E_0\left\{\sum_{t=0}^{\infty}(\beta\eta)^t[\alpha\log C_{Mt} + (1-\alpha)\log D_t + \phi\log(1 - H_{1t} - H_{2t})]\right\} \tag{20}$$

Optimality implies that the value marginal product of each input will be equalized across sectors. Given that identical Cobb–Douglas production functions are assumed, this implies the fraction of the total quantity of each input assigned to each sector is the same across inputs. Letting $H_{Mt} = H_{1t} + H_{2t}$, this implies that $\frac{K_{eit}}{K_{et}} = \frac{K_{sit}}{K_{st}} = \frac{H_{it}}{H_{Mt}}$ for $i = 1{,}2$. Given this result, and the fact that the technology is constant returns to scale, it is possible to aggregate over sectors to obtain the aggregate resource constraint:

$$C_{Mt} + I_{st} + \frac{1}{q_t}(I_{dt} + I_{et}) = z_t A K_{et}^{\theta_1} K_{st}^{\theta_2} H_{Mt}^{1-\theta_1-\theta_2} \equiv Y_t \tag{21}$$

Note that in this aggregate resource constraint, the outputs I_d and I_e are divided by q. In the decentralized version of this economy, $\frac{1}{q}$ is the price of equipment goods relative to output from sector 1. This result shows that data on the relative price of equipment can be used to measure equipment-specific technological change.

Given values for K_{e0}, K_{s0} and D_0, the equilibrium stochastic process for this economy can be found by solving the planner's problem maximizing (20) subject to (15)–(19) and (21).

4.2.1 Balanced Growth Path

Due to the positive drift in the random walks (18) and (19), this model exhibits stochastic growth. In a certainty version of the model in which $\sigma_1 = \sigma_2 = 0$, there is a balanced growth path where the asymptotic growth factors are given by $g_c = \frac{Y_{t+1}}{Y_t} = \frac{C_{M,t+1}}{C_{Mt}} = \frac{I_{s,t+1}}{I_{st}} = \frac{K_{s,t+1}}{Kst} = e^{\frac{\mu_1 + \theta_1 \mu_2}{1-\theta_1-\theta_2}}$ and $g_e = \frac{I_{e,t+1}}{I_{et}} = \frac{I_{d,t+1}}{I_{dt}} = \frac{K_{e,t+1}}{Ket} = \frac{D_{t+1}}{D_t} = g_c e^{\mu_2}$. Given these growth factors, the asymptotic growth path can be written $Y_t = g_c^t \bar{Y}$, $H_{Mt} = \bar{H}_M$, $C_{Mt} = g_c^t \bar{C}_M$, $I_{st} = g_c^t \bar{I}_s$, $K_{st} = g_c^t \bar{K}_s$, $I_{et} = g_e^t \bar{I}_e$. $I_{dt} = g_e^t \bar{I}_d$, $K_{et} = g_e^t \bar{K}_e$ and $D_t = g_e^t \bar{D}$, where the steady state values are the solutions to the following equations (given $\bar{q}$ and $\bar{z}$):

$$\frac{g_c}{\beta} = \theta_2 \frac{\bar{Y}}{\bar{K}_s} + 1 - \delta_s \tag{22}$$

$$\frac{g_e}{\beta} = \theta_1 \frac{\bar{Y} \ \bar{q}}{\bar{K}_e} + 1 - \delta_e \tag{23}$$

$$\frac{g_e}{\beta} = \frac{(1-\alpha)\bar{C}_M \ \bar{q}}{\alpha \bar{D}} + 1 - \delta_d \tag{24}$$

$$\frac{\phi}{1 - \bar{H}_M} = \alpha(1 - \theta_1 - \theta_2) \frac{\bar{Y}}{\bar{H}_M \bar{C}_M} \tag{25}$$

$$\bar{Y} = A \bar{K}_e^{\theta_1} \bar{K}_s^{\theta_2} \bar{H}_M^{1-\theta_1-\theta_2} \tag{26}$$

$$\bar{C}_M = \bar{Y} - \bar{I}_s - \frac{1}{\bar{q}}[\bar{I}_e + \bar{I}_d] \tag{27}$$

$$\bar{I}_s = (\delta_s + \eta g_c - 1)\bar{K}_s \tag{28}$$

$$\bar{I}_e = (\delta_e + \eta g_e - 1)\bar{K}_e \tag{29}$$

$$\bar{I}_d = (\delta_d + \eta g_e - 1)\bar{D} \tag{30}$$

We use this nonstochastic asymptotic growth path to help us calibrate the model and to construct capital stock series that are consistent with the model's balanced growth properties.

4.2.2 Calibrating the Model with US Data

We proceed by connecting each endogenous variable of this model with a counterpart taken from the US National Income and Product Accounts. The data we use runs from

1954Q1 to 2014Q4. On the product side, the model has one nondurable consumption good (C_{Mt}) which we take to be the sum of nondurable consumption, services and government consumption. There are three forms of investment: I_e is the sum of private and government investment in equipment; I_s is the sum of private investment in structures, intellectual property, residential structures, and government investment in structures and intellectual property; and I_d is purchases of consumer durables. Given that we have not allocated every component of Gross Domestic Product to one of these expenditure categories, we take total output to be $Y_t = C_M + I_s + \frac{1}{q}(I_d + I_e)$. The relative price of equipment in our model is equal to $\frac{1}{q_t}$, so we identify q_t from the relative price of equipment calculated by Riccardo DiCecio (see DiCecio, 2009).[g]

The capital stocks, which are the sum of both private and government fixed assets, are computed from annual quantity indexes of fixed assets obtained from the Bureau of Economic Analysis and is the stock associated with each investment series. In particular, K_s is nonresidential and residential structures along with intellectual property, K_e is the stock of equipment, and D is the stock of consumer durables. To obtain quarterly real stocks of capital, the annual quantity indexes are multiplied by the corresponding 2009 nominal value and quarterly series are obtained by iterating on the laws of motion (15)–(17) using the corresponding quarterly investment series.[h] Per capita capital stocks and output are obtained by dividing by the civilian population (16–64) plus military personnel. Finally, the hours series we use is average weekly hours per person (including military hours) based on data from the Current Population Survey. In particular, we have updated the series created by Cociuba et al. (2012).

Given these empirical counterparts, the growth factor for population is $\eta = 1.003$ and the growth factor for per capita output is $g_c = 1.0036$. The parameter $\mu_2 = 0.0104$, which is the average of $\log q_{t+1} - \log q_t$. This implies that $g_e = g_c e^{\mu_2} = 1.014$.

[g] This data series is available on the FRED database maintained by the Federal Reserve Bank of St. Louis.

[h] Given that the model assumes constant depreciation rates, which does not hold in our data sample, we allow the depreciation rate to vary across 10 year periods when constructing the quarterly capital stock series. That is, an initial value for the annual series in year t and a terminal value in year t + 10, we find the depreciation rate such that iterations on the law of motion of the capital stock hits the terminal value in 40 quarters using the corresponding quarterly investment series.

In particular, we find the depreciation rate δ_i for decade i such that $K_{i+10} = (1-\delta_i)^{40} K_i + \sum_{j=1}^{40} (1-\delta_i)^{40-j} I_j$, where K_i is the capital stock at the beginning of year i, K_{i+10} is capital at the beginning of year i + 10, and $\{I_j\}_{j=1}^{40}$ is investment for each quarter between those dates. Once we know δ_i for each subperiod in our sample, it is straightforward to construct quarterly capital stocks for each quarter of year i.

The capital stock obtained, however, is inconsistent with the trend introduced by our empirical measure of q, which is based on different price deflators than those used in producing the NIPA capital stocks. As a result, we also adjust the trend growth of the capital stocks so that these stocks are consistent with long-run growth properties of the model. That is, a trend is added to our quarterly series for K_s so that it has an average growth rate equal to g_c and D and K_e are similarly adjusted to have an average growth factor g_e.

We calibrate the model by setting $\beta = 0.99$, labor's share, $1 - \theta_1 - \theta_2$, equal to 0.6 and the depreciation rates equal to the average of the depreciation rates obtained when forming the quarterly capital stock series. This gives us $\delta_e = 0.021$, $\delta_s = 0.008$, and $\delta_d = 0.05$. The individual capital shares are based on estimates in Valentinyi and Herrendorf (2008) renormalized so they sum to 0.4. In particular, we set $\theta_1 = 0.21$ and $\theta_2 = 0.19$. The parameter α is computed from a version of equation (24) where the term $\frac{\bar{C}_M \ \bar{q}}{\bar{D}}$ is replaced with the average value of $\frac{C_{M,t} \ q_t}{D_t}$ from the empirical counterparts to these variables. This gives $\alpha = 0.817$.

Next, we set $\bar{Y}$, $\bar{H}_M$, and $\bar{q}$ equal to the initial observation in the time series for each of these variables. The seven remaining steady states ($\bar{K}_s$, $\bar{K}_e$, $\bar{D}$, $\bar{I}_s$, $\bar{I}_e$, $\bar{I}_d$, and $\bar{C}_M$) are obtained by solving seven equations (22)–(24) and (27)–(30). So that the steady state capital stocks are equal to the first observations for these variables, we multiply all observations of K_s by $\frac{\bar{K}_s}{K_{s,0}}$, all observations of K_e by $\frac{\bar{K}_e}{K_{e,0}}$ and all observations of D by $\frac{\bar{D}}{D_0}$. These are the capital stocks used to construct the empirical counterpart to z_t.

We construct a quarterly time series for the exogenous shock, z_t, from 1954Q1 to 2014Q4 by setting $z_t = \frac{Y_t}{A K_{et}^{\theta_1} K_{st}^{\theta_2} H_{Mt}^{1-\theta_1-\theta_2}}$ where the parameter A is chosen so that the first observation of z is equal to one. This implies $A = 6.21$. Somewhat surprisingly, the growth rate of z_t when computed in this way turns out to be zero ($\mu_1 = 0$). That is, when measured through the lens of this model, the average rate of growth in per capita income during the postwar period is accounted for entirely by equipment specific technological improvement.

We summarize the calibration of the model in Table 1 in the column labeled "Two sector." This table reports the calibrated parameter values for all models considered, so we will refer back to this table as we discuss these alternatives.

4.2.3 Comparison of Model with Data

Given our time series for z_t and q_t, times series for the endogenous variables of the model are computed for the sample period 1954Q1–2014Q4. This is done using log-linear approximations of the decision rules that solve the planner's problem obtained using standard numerical methods (see, for example, Uhlig, 1999). Fig. 17 shows our measures of output and hours from US data along with the time series for these variables implied by our model.

Output from the data and model are quite close to each other until the mid-1980s when model output becomes lower than in the data. By 2002, however, model output has recovered. Model hours tend to be higher than in the data during the 1960s and 1970s, and lower from the mid-1980s until the Great Recession. Following the Great Recession, the data shows some recovery in hours worked that the model does not.

Table 1 Calibrated parameter values

Parameter description		Two sector	One sector	Three sector (1)	Three sector (2)
Equipment share	θ_1	0.21		0.21	0.21
Structures share	θ_2	0.19		0.19	0.19
Capital share	θ		0.4		
Depreciation rate—Equipment	δ_E	0.021		0.021	0.021
Depreciation rate—Structures	δ_S	0.008		0.008	0.008
Depreciation rate—Durables	δ_D	0.05		0.05	0.05
Depreciation rate—Capital	δ		0.013		
Growth rate—z	μ_1	0		0	0
Growth rate—q	μ_2	0.0104		0.0104	0.0104
Growth rate—z	μ		0.0021		
Population growth factor	η	1.003	1.003	1.003	1.003
Discount factor	β	0.99	0.99	0.99	0.99
Utility share for mkt. consumption	α	0.82		0.33	0.53
Utility parameter for leisure	ϕ	2.37	2.37	1.19	1.19
Scale parameter—Market production	A	6.21	2.7	6.21	6.21
Elasticity parameter—Home production	σ			0	0.4
Elasticity parameter—Mkt./non-mkt. cons.	ω			0.6	0
Durable share—Home production	φ			0.25	0.13
Scale parameter—Home production	A_N			4.19	4.87

Three sector (1)—Standard home production
Three sector (2)—Calibration inspired by Greenwood et al. (2005)

Fig. 18 consists of four panels showing output, hours, consumption and investment—from both the model and the data—that has been filtered to show only fluctuations between 2 and 32 quarters. The real business cycle literature has demonstrated that neoclassical models of this sort generate fluctuations similar to those in postwar US data at this frequency. As the figure illustrates, this is particularly true for output and investment.

Less studied, however, are the low frequency fluctuations exhibited by models of this sort. Fig. 19 is a plot of model and US data for the same four variables that has been filtered to show fluctuations between 32 and 200 quarters. The model seems to do a pretty good job in tracking fluctuations in output, consumption and investment in this frequency band. For hours worked, the model captures some of the low frequency movements, but not others. In the late 1950s, the model shows hours falling sooner than it does in the data, while the model and data track pretty closely during the 1960s and early 1970s. In the late 1970s, the data shows an increase in hours worked that the model does not capture, but the model and data follow each other throughout the 1980s and 1990s. At the time of the Great Recession, the decline in hours—as well as other macro aggregates—is less in the model than in the data.

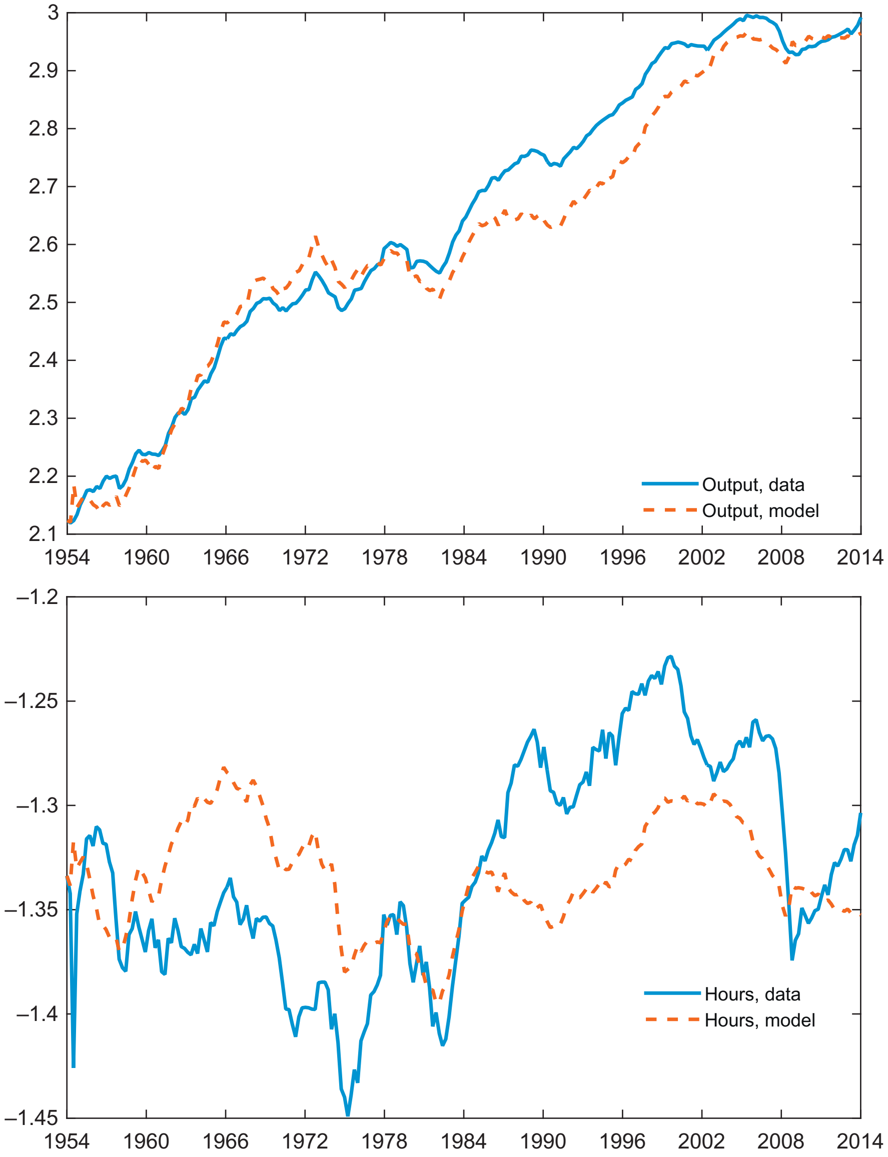


Fig. 17 Output and hours worked, data and two-sector model.

Fig. 20 plots the same data as the previous figure for filtered output and hours for both the 2–32 quarter frequency and the 32–200 quarter frequency. The difference is that we have included a third time series in each plot that shows simulated data under the assumption that there were no fluctuations in z_t and only fluctuations in q_t. That is, when

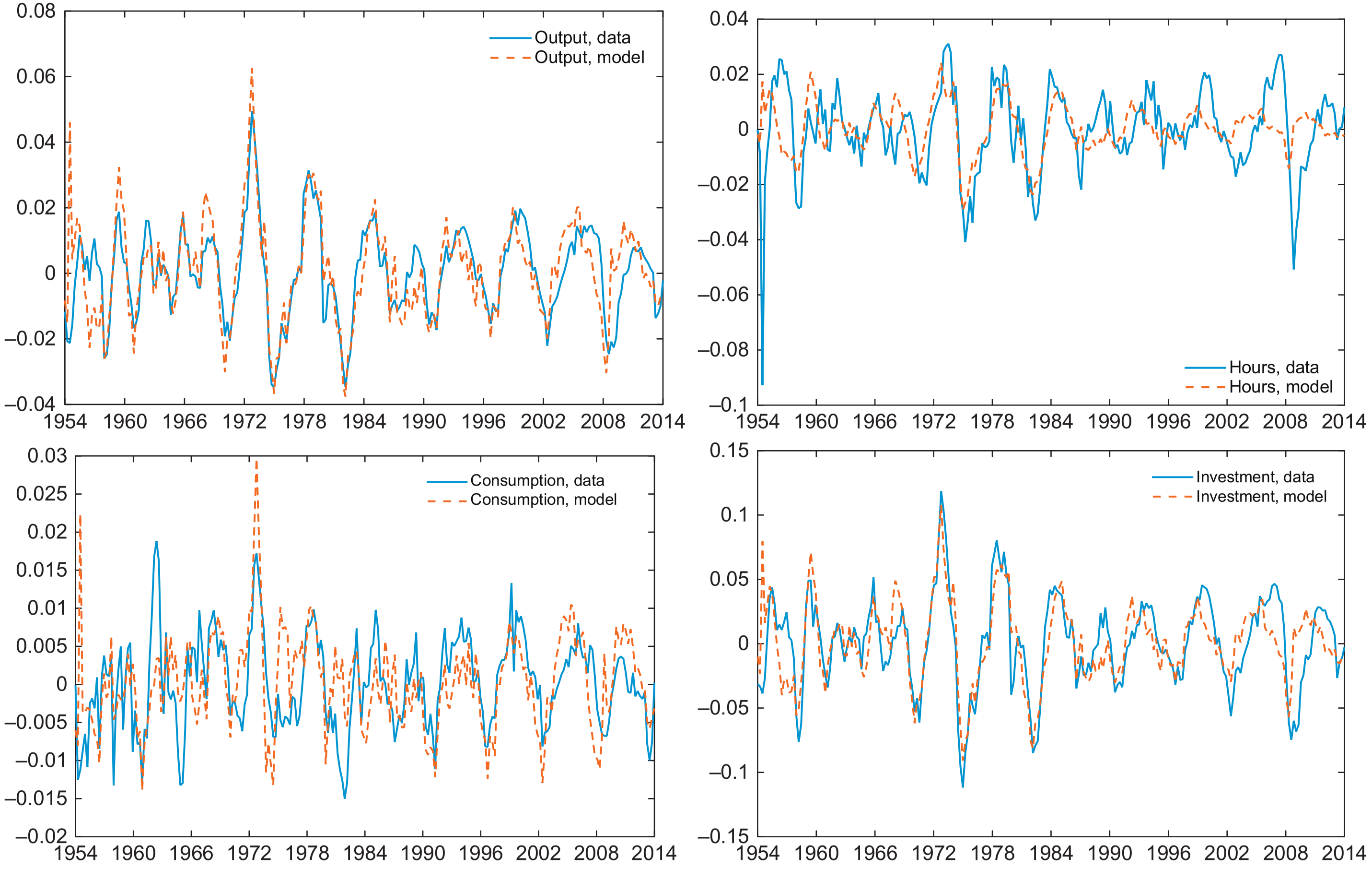


Fig. 18 Filtered actual and two-sector model data (2–32 quarters).

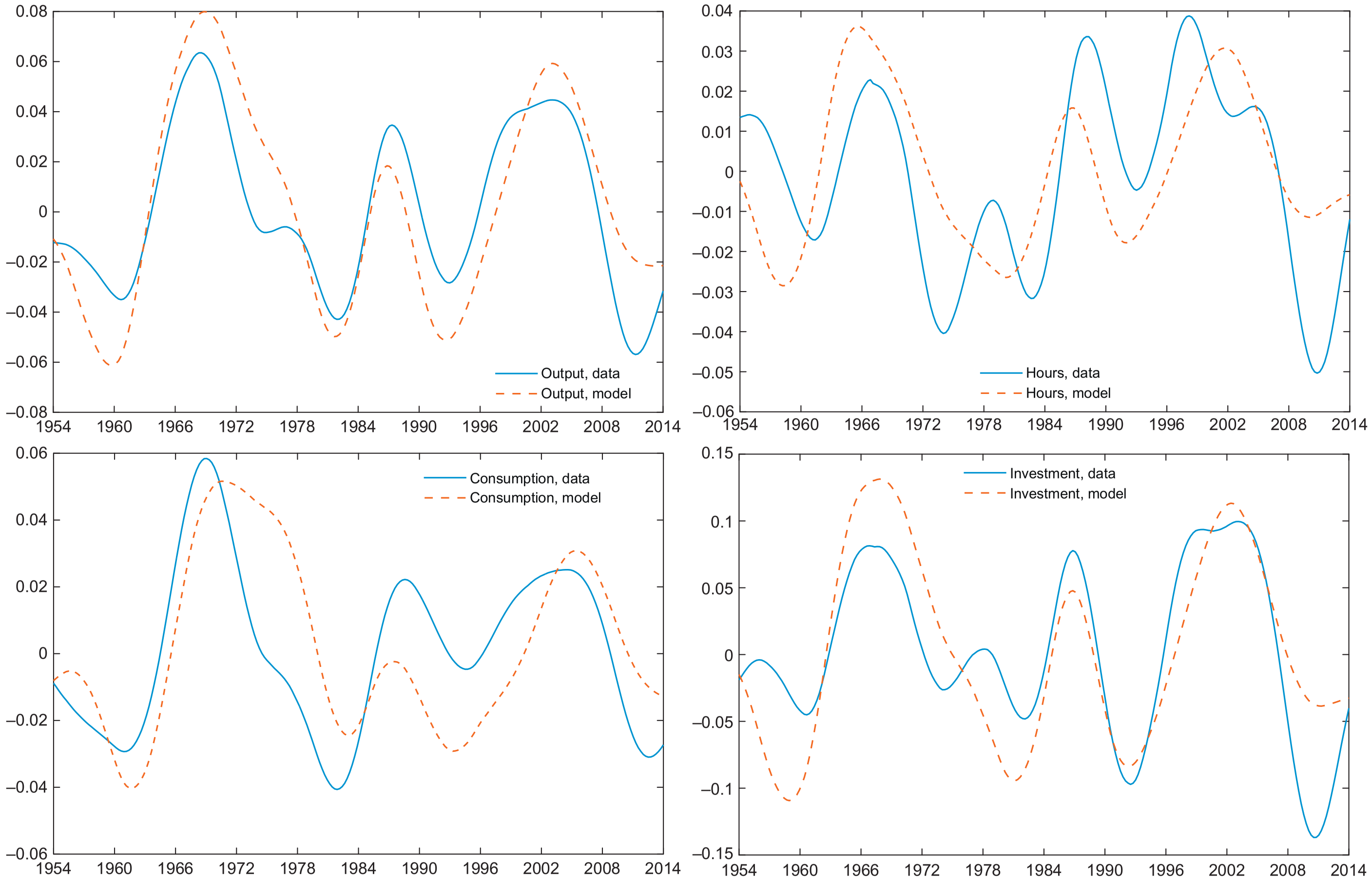


Fig. 19 Filtered actual and two-sector model data (32–200 quarters).

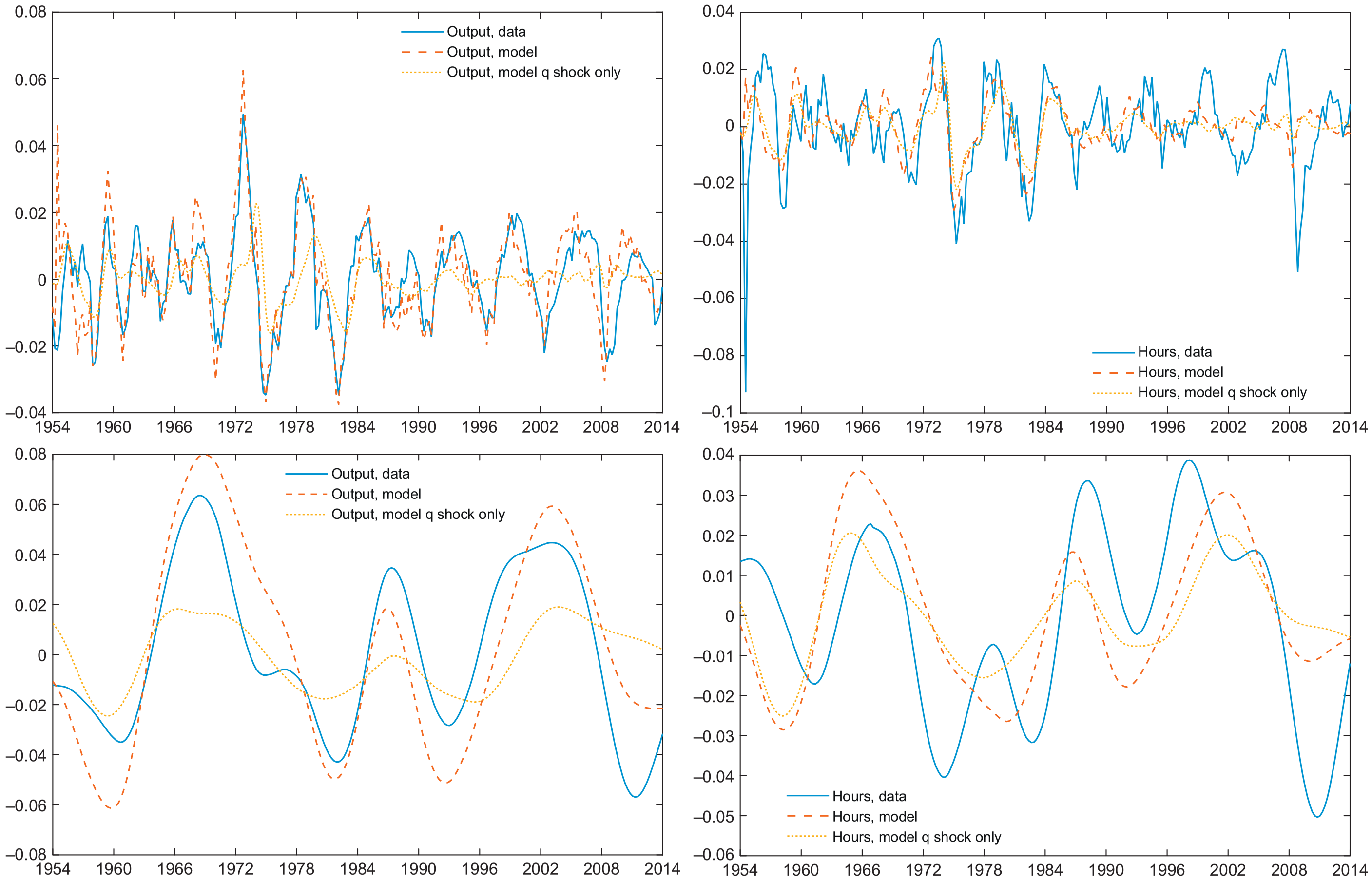


Fig. 20 Contribution of equipment specific technology fluctuations.

computing the simulation, the time series for z_t is replaced by the nonstochastic growth path for z. That is, $z_t = e^{t\mu_1}$ for all t.

This figure shows that much of the high and low frequency fluctuations in hours worked are due to movements in q_t, but this is not as true for fluctuations in output. It is also less true for business cycle fluctuations in hours worked in more recent decades.

4.3 One-Sector Model

We now proceed to compare the fluctuations exhibited by the two-sector model with a standard one-sector neoclassical stochastic growth model. This one-sector economy consists of a single production sector that produces output from capital and labor that can be consumed or invested. It differs from the two-sector model in that there is only one type of capital stock, no separate role for consumer durables, and one type of technology shock. In particular, the resource constraint, which replaces equation (21), is

$$C_t + I_t = Y_t = z_t A K_t^{\theta} H_t^{1-\theta}. \tag{31}$$

The law of motion for capital next period is given by

$$\eta K_{t+1} = (1-\delta)K_t + I_t \tag{32}$$

where the depreciation rate is $0 < \delta \leq 1$ and $1 \leq \eta \leq \frac{1}{\beta}$ is the population growth factor. The logarithm of the technology shock, z_t, is assumed to follow a random walk with drift ($\mu \geq 0$). We assume that the period t realization of z is observed at the beginning of the period.

$$\log z_{t+1} = \log z_t + \varepsilon_{t+1}\ ,\ \varepsilon \sim N(\mu, \sigma^2) \tag{33}$$

The preferences of the representative infinitely-lived household are given by

$$E\sum_{t=0}^{\infty} (\beta\eta)^t [\log C_t + \phi \log L_t] \tag{34}$$

where $0 < \beta < 1$ and $\phi > 0$. The variable L_t is leisure, where

$$L_t + H_t = 1. \tag{35}$$

Given K_0, we compute an equilibrium sequence for $\{C_t, I_t, Y_t, H_t, L_t, K_{t+1}\}$ by maximizing (34) subject to (31)–(33) and (35).

4.3.1 Calibrating the One-Sector Model with US Data

For comparison purposes, we begin by keeping the definition of output the same as in the two-sector model, $Y = C + I_s + \frac{1}{q}(I_d + I_e)$. Given that there is no separate role for consumer durables in this model, we define investment in the one-sector model to be

$I = I_s + \frac{I_e}{q}$ and consumption to be the sum of nondurable consumption plus services and $\frac{I_d}{q}$. That is, $C_t = C_{Mt} + \frac{I_d}{q}$, where C_M is consumption from the two-sector model. The capital stock is the sum $K = K_e + K_s$. The quarterly capital stock series for this sum is formed using the same method as for the two-sector model and the quarterly depreciation rate turns out to be $\delta = 0.013$. As in the two-sector model, $\beta = 0.99$ and labor's share is taken to be 0.6, so $\theta = 0.4$. Given this, a quarterly time series for the exogenous shock z_t, from 1954Q1 to 2014Q4, is constructed by setting $z_t = \frac{Y_t}{AK_t^{\theta} H_t^{1-\theta}}$, where the parameter A is set so that $z_0 = 1$. This implies that $A = 2.7$. In addition, the drift parameter, μ, turns out to be 0.0021.

As in the two-sector model, we set the steady state values for K, H and Y equal to the first observation in our data sample (for 1954Q1). Steady state consumption is then obtained from the steady state version of the resource constraint (31). We can then calibrate the parameter ϕ from the steady state condition for hours worked. That is, $\phi = \frac{(1-\theta)\bar{Y}(1-\bar{H})}{\bar{C} \;\; \bar{H}} = 2.37$.

To facilitate comparison across models, the parameter values are also reported in Table 1.

4.3.2 Comparing the One- and Two-Sector Models with US Data

Table 2 provides two metrics for comparing the closeness of the one- and two-sector model simulations with filtered data. These measures include the ratio of the standard deviations of the model series with the standard deviation of the data series. This provides a measure of how well the model is capturing the volatility in the data. The second measure is the correlation between the model simulations and the data. We report these measures for data filtered to extract fluctuations of 2–32 quarters, 32–200 quarters and 2–200 quarters. In all cases, a number closer to one implies a better fit.[i]

The table shows that the correlation between model and data for business cycle fluctuations is higher for the two-sector model, with the exception of consumption. For low frequency fluctuations, the one-sector model does slightly better, although the correlation between hours worked from the model and data is slightly higher for the two-sector model. The volatility of the various series is generally better accounted for by the two-sector model. Hence, the main conclusion we draw from this table is that the two-sector model fits the data better than the one-sector model, with the exception of consumption fluctuations. We find it interesting that the two-sector model is able to account for

[i] In this table and subsequent tables, we only use data starting from 1955Q1. The reason is that there is an unusual hours observation in 1954 that can be seen in Fig. 17, and we don't want that observation distorting the statistics reported in these tables.

Table 2 Comparing models with data (1955Q1–2014Q4)

	One-sector model		Two-sector model	
	Standard deviation model/data	Correlation model and data	Standard deviation model/data	Correlation model and data
2–32 Quarters				
Y	0.86	0.80	1.09	0.84
C	0.73	0.82	1.00	0.56
I	0.71	0.64	0.86	0.79
H	0.30	0.18	0.63	0.48
32–200 Quarters				
Y	0.85	0.88	1.21	0.86
C	0.70	0.78	1.07	0.64
I	0.81	0.82	1.08	0.81
H	0.35	0.51	0.81	0.53
2–200 Quarters				
Y	0.86	0.86	1.21	0.84
C	0.72	0.77	1.09	0.62
I	0.80	0.77	1.05	0.79
H	0.33	0.40	0.74	0.50

volatility in spite of the fact that we have assumed random walk technology shocks and divisible labor. These are both assumptions that tend to reduce the size of fluctuations.[j]

Fig. 21 provides the same information as Fig. 20 except that the comparison is now with the one-sector simulation for output and hours rather the "q-shock" only simulation. The figure illustrates that much of the low frequency movements in output can be accounted for by the one-sector model almost as well as the two sector. The low frequency volatility of hours, however, is better explained by the two-sector model than the one sector.

4.4 A Three-Sector Model

This section studies a model constructed by adding a nonmarket home production sector to the two-sector model. We develop the three-sector model with two alternative home production specifications. One is the standard home production specification of Benhabib et al. (1991) and much of the literature that follows from this. This formulation

[j] See Hansen (1985) concerning the impact of divisible labor on fluctuations and Hansen (1997) for the impact of random walk technology shocks.

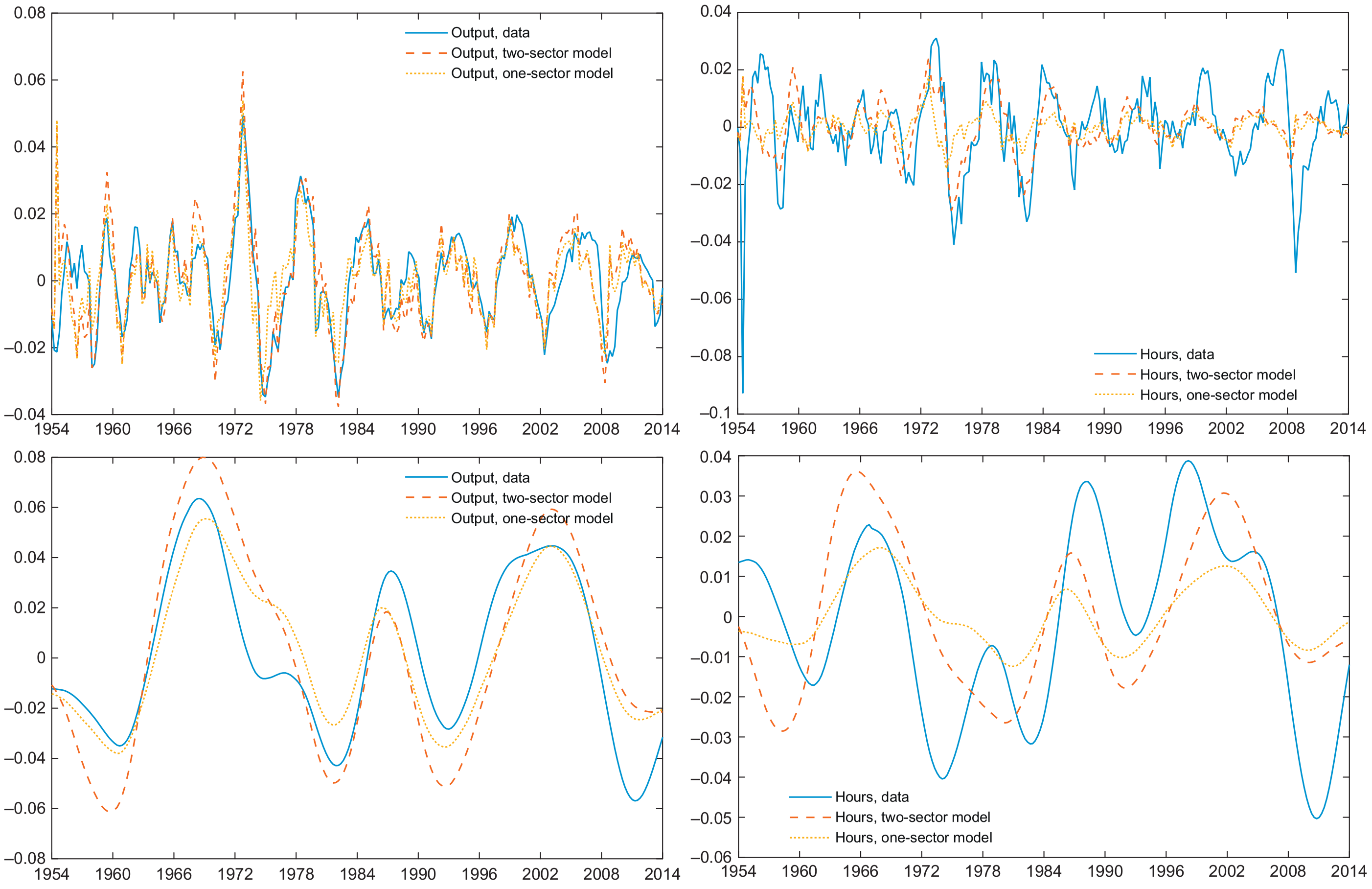


Fig. 21 Comparison of two-sector and one-sector models.

provides an additional margin of substitution for the household in which time can be allocated to market production, home production, or leisure. In the Benhabib, Rogerson and Wright model, there is a relatively high substitution elasticity between home-produced goods and market-produced goods, and this high elasticity generates significant movement of labor between the home sector and market sector in response to shocks. Home goods are produced using a Cobb–Douglas technology with labor and consumer durables.

The alternative home production formulation is motivated by Greenwood et al. (2005), which argues that rapid technological change in labor-saving consumer durables has secularly reallocated time from home production to market production, mainly by women moving into the labor force. In this specification, consumer durables are more substitutable with labor than in the Benhabib et al. (1991) specification that assumes a Cobb–Douglas technology for the home sector.

The model presented here nests both of these specifications. In particular, we assume that a nonmarket consumption good, C_{Nt}, is produced using labor (H_{Nt}) and the stock of consumer durables. As in Greenwood et al. (2005), we allow for the possibility that durables and labor are more substitutable than implied by the standard Cobb–Douglas production function. In particular, we assume the following functional form for the home production function with $\sigma > 0$:

$$C_{Nt} = A_N\left[\varphi\left(\frac{D_t}{e^{\mu_2 t}}\right)^{\sigma} + (1-\varphi)(g_c^t H_{Nt})^{\sigma}\right]^{\frac{1}{\sigma}} \tag{36}$$

The standard version of the model can be recovered by making σ close to zero. Note that the terms $e^{\mu_2 t}$ and g_c^t are included here to guarantee that C_{Nt} grows at the same rate as total output along the balanced growth path.

The second modification relative to the two-sector model is to replace the objective function (20) with the following:

$$\max E_0\left\{\sum_{t=0}^{\infty}(\beta\eta)^t[\log C_t + \phi\log(1 - H_{Mt} - H_{Nt})]\right\}, \tag{37}$$

where consumption, C_t, is a composite consumption good, standard in the home production literature, derived from market and nonmarket consumption goods

$$C_t = \left[\alpha C_{Mt}^{\omega} + (1-\alpha)C_{Nt}^{\omega}\right]^{\frac{1}{\omega}} \tag{38}$$

Given values for K_{e0}, K_{s0} and D_0, the equilibrium stochastic process for this economy can be found by solving the planner's problem maximizing (37) subject to (15)–(19), (21), (36), and (38).

4.4.1 Calibrating the Three-Sector Model to US Data

The calibration strategy is exactly the same as for the two-sector case, although the model introduces four new parameters (A_N, φ, ω, and σ) and two other parameters (α and ϕ) have different interpretations in this model. In addition, two new variables are introduced that are not directly observable in the US data. These are nonmarket consumption (C_N) and nonmarket hours worked (H_N). In the absence of measured counterparts to these variables, we assume that in steady state $\frac{\bar{C}_N}{\bar{C}_M} = 0.25$ and $\bar{H}_N = \frac{1}{6}$, which are values consistent with the home production literature. The mapping between all other model variables and US time series is the same as in the two-sector model.

The steady state values for $\bar{K}_s$, $\bar{K}_e$, $\bar{Y}$, $\bar{C}_M$, $\bar{I}_s$, $\bar{I}_e$, $\bar{I}_d$, $\bar{D}$, $\bar{H}_M$, $\bar{H}_N$, $\bar{C}_N$, and $\bar{C}$ are determined by Eqs. (22), (23), (26)–(30), and the following five equations:

$$\frac{g_E}{\beta} = \frac{(1-\alpha)A_N^{\sigma}\varphi\bar{q}\bar{C}_M^{1-\omega}}{\alpha\bar{C}_N^{\sigma-\omega}\bar{D}^{1-\sigma}} + 1 - \delta_D \tag{39}$$

$$\frac{\phi}{1-\bar{H}_M-\bar{H}_N} = \alpha(1-\theta_1-\theta_2)\frac{\bar{Y}}{\bar{H}_M\bar{C}^{\omega}\bar{C}_M^{1-\omega}} \tag{40}$$

$$\frac{\phi}{1-\bar{H}_M-\bar{H}_N} = \frac{(1-\alpha)A_N^{\sigma}(1-\varphi)}{\bar{H}_N^{1-\sigma}\bar{C}^{\omega}\bar{C}_N^{\sigma-\omega}} \tag{41}$$

$$\bar{C}_N = A_N\left[\varphi\bar{D}^{\sigma} + (1-\varphi)\bar{H}_N^{\sigma}\right]^{\frac{1}{\sigma}} \tag{42}$$

$$\bar{C} \;= \left[\alpha\bar{C}_M^{\omega} + (1-\alpha)\bar{C}_N^{\omega}\right]^{\frac{1}{\omega}}. \tag{43}$$

We experiment with two different sets of values for the parameters σ and ω to differentiate between our two home production specifications. Given values for these parameters, values for α, ϕ, φ and A_N can be obtained from equations (39) to (42) subject to $\frac{\bar{C}_N}{\bar{C}_M} = 0.25$, $\bar{H}_N = \frac{1}{6}$ and $\bar{C}$ is given by equation (43).[k]

The first calibration we consider is referred to as the "standard home production" model. In this case, $\omega = 0.6$ and $\sigma = 0$, which corresponds to values common in the home production literature (see Chang and Schorfheide, 2003). In this case, the utility function (38) allows for more substitutability between home consumption and market consumption than implied by a Cobb–Douglas specification while the home production

[k] We also use the fact that, as in the two-sector case, we choose parameters so that $\bar{q}$, $\bar{H}_M$ and $\bar{Y}$ are the first observation in our data sample.

function (36) is assumed to be Cobb–Douglas. The second calibration, which we refer to as the "alternative home production" model, is motivated by Greenwood et al. (2005) and sets $\omega = 0$ and $\sigma = 0.4$. Here, (38) is assumed to be Cobb–Douglas and we allow for an elasticity of substitution between durables and hours that is greater than 1 in the home production function (36). The parameter values associated with both calibrations are given in Table 1.

4.4.2 Fluctuations in the Three-Sector Model

We begin by comparing the simulations produced by the two versions of the three-sector model that we consider. Fig. 22 shows unfiltered output and hours from the two models as well as from the US time series. Both models account for output movements quite well, although the alternative calibration does a somewhat better job in the 1960s and 1970s while the standard home production calibration fits the data better in the 1980s and 1990s. Both models imply similar paths during the Great Recession period. The same is also true for hours worked—the alternative calibration does better during the early periods and less well during the 1980s and 1990s. Both calibrations give essentially identical results during the 2000s.

An interesting difference between hours worked from the two models can be seen from examining the period from about 1982 to 2000. The rise in hours worked predicted by the alternative calibration during this period is significantly larger than that predicted by the standard home production model. In the spirit of Greenwood et al. (2005), this calibration does a better job of capturing the secular increase in hours worked that occurs over this period, mainly due to women entering the labor force. As one can see from Fig. 23, this difference does not appear in the low frequency fluctuations that we report.

The two calibrations, however, give essentially the same results once the data is filtered. Fig. 23 illustrates this by plotting filtered data for output and hours from the two versions of the model. The data for both business cycle fluctuations as well as low frequency fluctuations essentially lay on top of each other. In particular, the alternative home production model does not exhibit the significantly larger increase in hours worked relative to the standard home production model during the 1980s and 1990s as was observed in Fig. 22.

The closeness of the filtered data from these models with filtered data from US time series is illustrated in Fig. 24 and Table 3. Fig. 24 shows filtered data from the standard home production calibration and the US economy for output and hours. When one compares the panels in Fig. 24 with the corresponding panels in Figs. 18 and 19, the results from the home production model appear very similar to the two-sector model with slightly more volatility in hours worked at both sets of frequencies.

The same sorts of conclusions that can be drawn from Fig. 24 are also apparent in Table 3. This table provides the same set of statistics as in Table 2 for comparing model data with actual data. Here, we compare both calibrations of our three-sector model with the US time series.

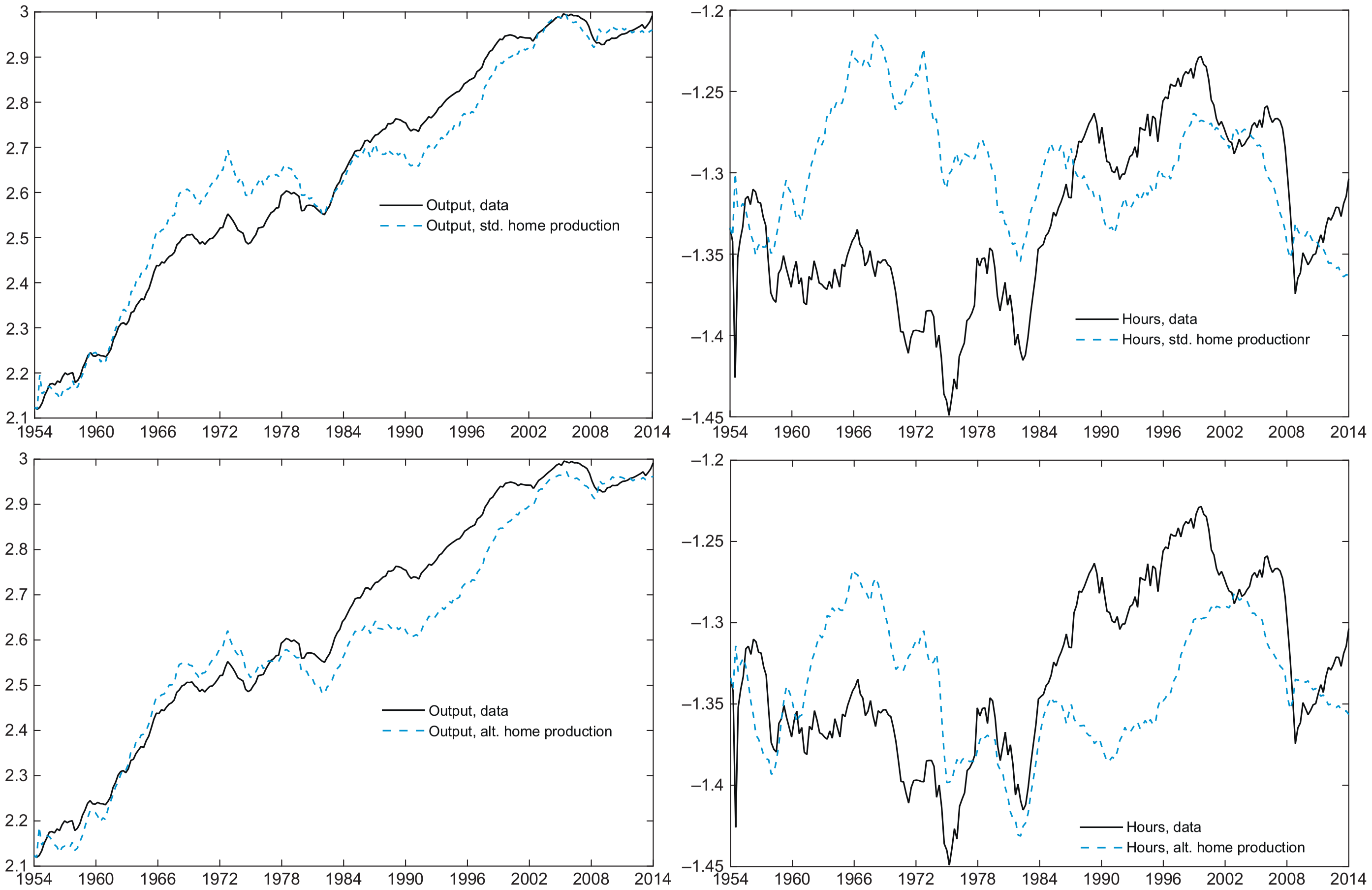


Fig. 22 Standard home production and alternative—output and hours.

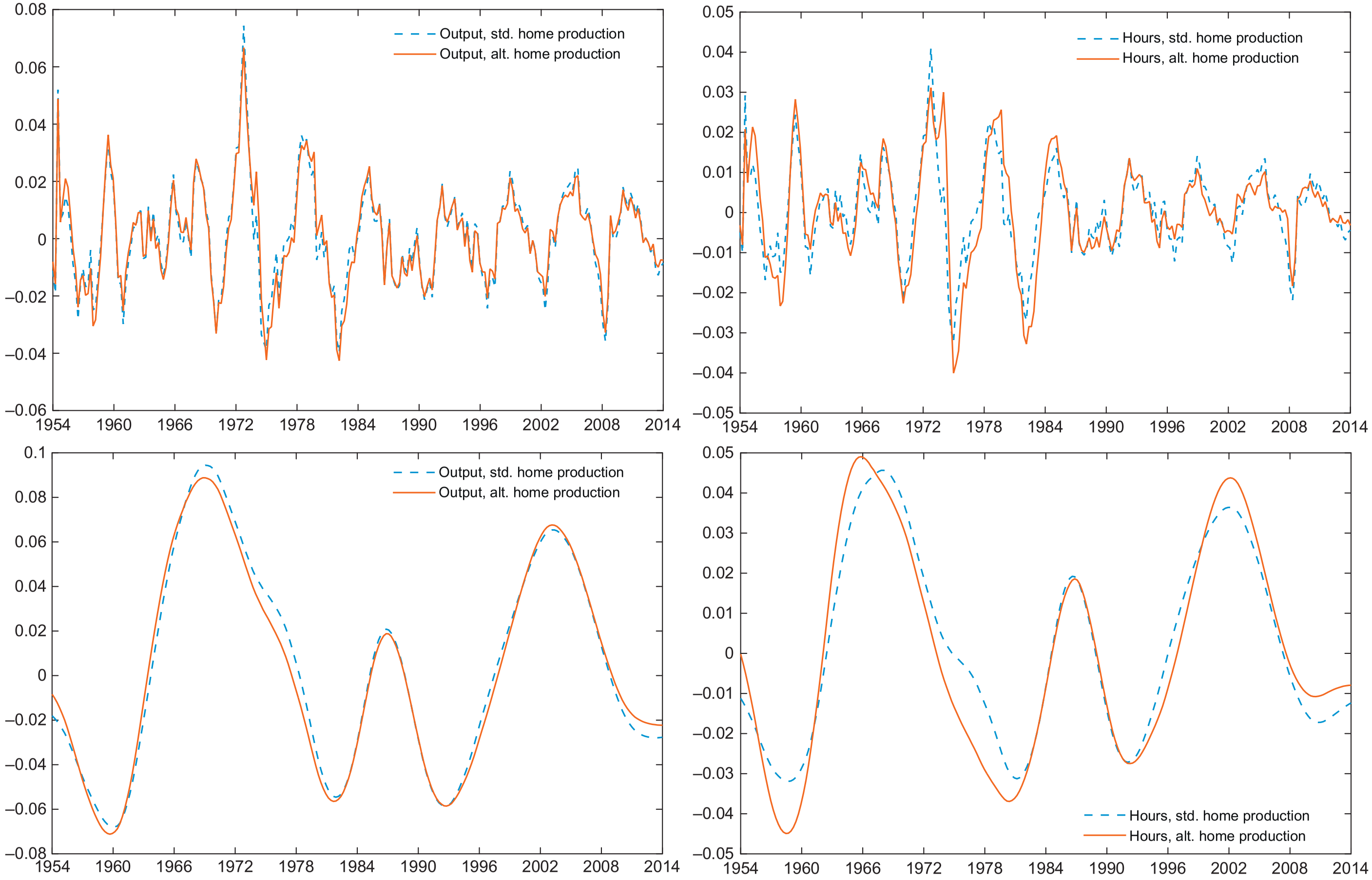


Fig. 23 Standard home production and alternative—filtered output and hours.

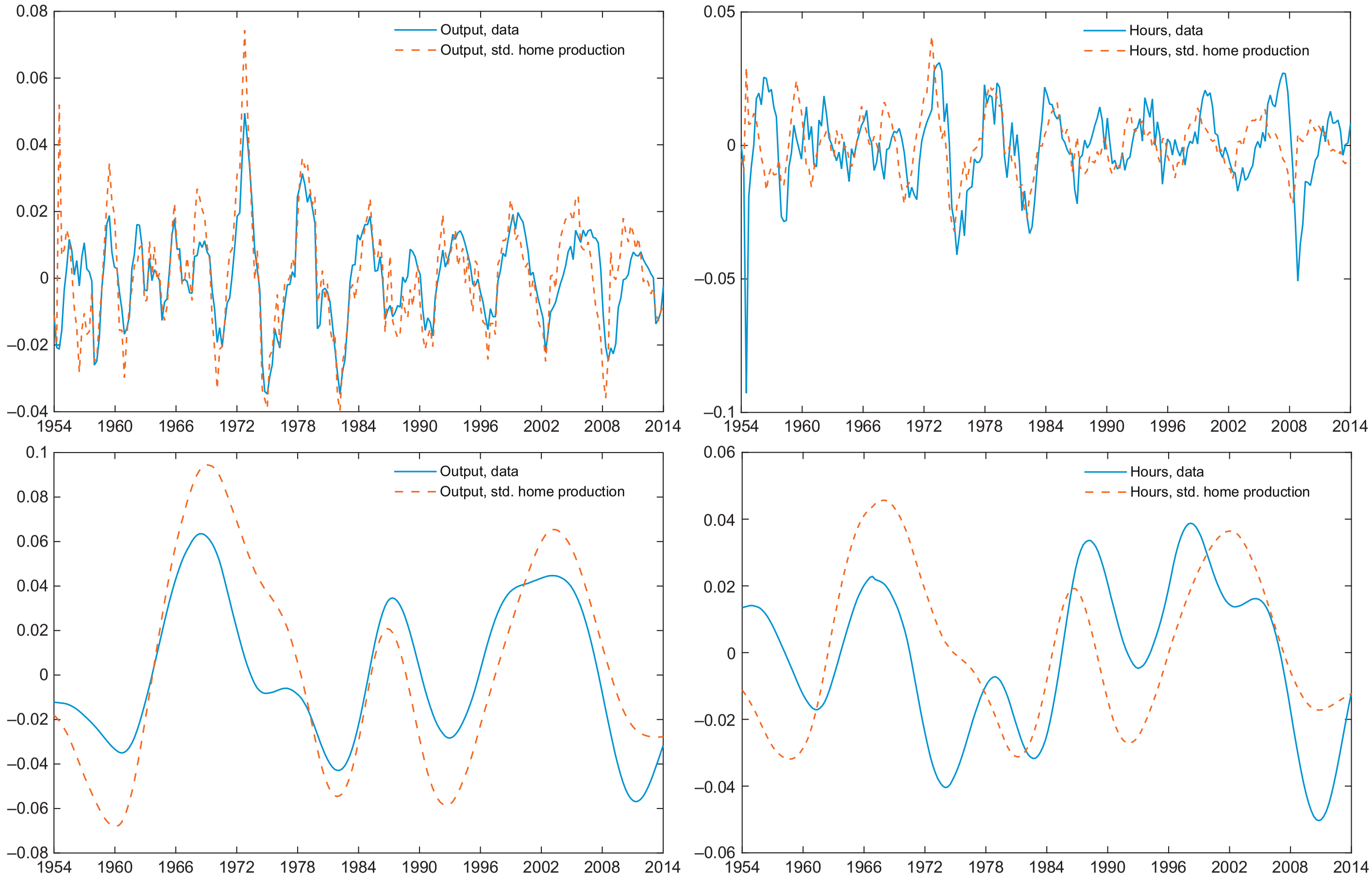


Fig. 24 Standard home production and data—filtered output and hours.

Table 3 Comparing models with data (1955Q1–2014Q4)

	Standard home production (ω = 0.6 and σ = 0)		Alternative (ω = 0 and σ = 0.4)	
	Standard deviation model/data	**Correlation model and data**	**Standard deviation model/data**	**Correlation model and data**
2–32 Quarters				
Y	1.23	0.84	1.23	0.84
C	1.52	0.50	1.02	0.39
I	0.95	0.80	1.09	0.78
H	0.76	0.39	0.89	0.50
32–200 Quarters				
Y	1.43	0.84	1.41	0.84
C	1.42	0.58	1.03	0.51
I	1.20	0.80	1.38	0.77
H	1.02	0.50	1.16	0.48
2–200 Quarters				
Y	1.43	0.86	1.41	0.83
C	1.45	0.56	1.05	0.49
I	1.15	0.78	1.32	0.75
H	0.95	0.44	1.07	0.45

The final set of tables we present in this section report the statistics for comparing model simulation and actual data for three subperiods of the postwar period. Table 4 looks only at the early postwar period from 1955Q1 to 1983Q4 and Table 5 reports statistics for the Great Moderation period from 1984Q1 to 2007Q3. Finally, statistics for the Great Recession and after are reported in Table 6.

Which model best explains postwar fluctuations in output, consumption, investment and hours worked? These tables show that it depends on the sample period and the frequency band of interest.

In the early postwar period (Table 4), all three models do a similar job fitting the data, but different models are better at accounting for fluctuations in different frequency bands. Hours is explained the least well by all of the models, but the correlation between model and data hours is highest for the two-sector model at business cycle frequencies and the home production model for lower frequencies. Output fluctuations are best explained by the two-sector model in all frequency bands considered. Consumption fluctuations are best explained by the one-sector model and investment fluctuations are almost equally well explained by the two- and three-sector models.

A feature seen in all three of these tables is that the volatility of model data relative to actual data rises as the number of sectors is increased. This is due to the increased substitution opportunities offered by multisector economies.

Table 4 Comparing Models with data (1955Q1–1983Q4)

	One-sector model		Two-sector model		Standard home production	
	Standard deviation model/data	Correlation model and data	Standard deviation model/data	Correlation model and data	Standard deviation model/data	Correlation model and data
2–32 Quarters						
Y	0.88	0.83	1.13	0.91	1.25	0.90
C	0.74	0.84	0.92	0.55	1.46	0.45
I	0.73	0.68	0.93	0.87	1.02	0.88
H	0.33	0.24	0.74	0.66	0.86	0.53
32–200 Quarters						
Y	0.97	0.91	1.47	0.95	1.69	0.92
C	0.70	0.80	1.10	0.74	1.44	0.67
I	1.24	0.76	1.87	0.92	2.14	0.90
H	0.46	0.41	1.09	0.44	1.45	0.45
2–200 Quarters						
Y	0.96	0.89	1.42	0.94	1.63	0.91
C	0.72	0.79	1.10	0.72	1.45	0.66
I	1.09	0.72	1.52	0.87	1.66	0.84
H	0.41	0.33	0.93	0.49	1.22	0.44

Table 5 Comparing models with data (1984Q1–2007Q3)

	One-sector model		Two-sector model		Standard home production	
	Standard deviation model/data	Correlation model and data	Standard deviation model/data	Correlation model and data	Standard deviation model/data	Correlation model and data
2–32 Quarters						
Y	0.88	0.84	1.06	0.79	1.23	0.81
C	0.71	0.81	1.10	0.70	1.55	0.68
I	0.74	0.76	0.80	0.71	0.88	0.73
H	0.33	0.24	0.53	0.20	0.73	0.26
32–200 Quarters						
Y	1.02	0.92	1.43	0.93	1.60	0.94
C	0.98	0.81	1.41	0.74	1.73	0.73
I	0.77	0.95	0.96	0.95	1.04	0.96
H	0.46	0.43	0.97	0.47	1.29	0.49
2–200 Quarters						
Y	1.09	0.91	1.52	0.91	1.71	0.92
C	1.05	0.79	1.55	0.74	1.94	0.73
I	0.79	0.91	0.98	0.91	1.06	0.92
H	0.49	0.26	0.98	0.22	1.33	0.28

Table 6 Comparing models with data (2007Q4–2014Q4)

	One-sector model		Two-sector model		Standard home production	
	Standard deviation model/data	Correlation model and data	Standard deviation model/data	Correlation model and data	Standard deviation model/data	Correlation model and data
2–32 Quarters						
Y	0.77	0.42	0.99	0.43	1.20	0.40
C	0.77	0.64	1.42	0.43	2.03	0.40
I	0.52	0.14	0.57	0.30	0.63	0.26
H	0.17	−0.34	0.26	−0.21	0.41	−0.24
32–200 Quarters						
Y	0.63	0.97	0.72	0.95	0.89	0.91
C	0.73	0.99	0.79	0.99	1.11	0.99
I	0.40	0.95	0.52	0.80	0.47	0.80
H	0.14	0.82	0.22	0.90	0.36	0.87
2–200 Quarters						
Y	0.55	0.75	0.66	0.66	0.76	0.55
C	0.67	0.93	0.68	0.91	0.94	0.88
I	0.28	0.33	0.42	0.28	0.37	0.22
H	0.10	0.02	0.16	0.10	0.23	−0.01

During the Great Moderation (Table 5), the one-sector model provides the highest correlations between model and actual data for output, consumption and investment, which is different from what is observed in the earlier period. Hours, however, are slightly better explained by the three-sector model. At lower frequencies, the three-sector model shows the highest correlation for all variables except consumption.

In the most recent period (Table 6), which covers the Great Recession and aftermath, a striking finding emerges regarding hours fluctuations. All three models show negative correlations between model and data hours worked at business cycle frequencies. However, this correlation is quite high, especially for the two- and three-sector models, at lower frequencies. At business cycle frequencies, all three models do a similarly poor job in accounting for fluctuations in output and investment. Again, the one-sector model does best in explaining consumption. But, at lower frequencies, all three neoclassical models show high correlations between model and data for these three variables as well as hours worked.

It is interesting and important that the fit of the two- and three-sector models for the 32–200 component is no different during the Great Moderation than during the 1955–1983 period. This is important because some economists have argued that neoclassical models cannot fit data from this specific period because the business cycle correlation between labor productivity and hours worked becomes negative during the Great Moderation (see Gali and van Rens, 2014). We find that the change in this higher frequency statistic has no bearing on the ability of these models to fit the large, longer-run component in the data. We also note that these models also fit the 32–200 component of the data well during the Great Recession and its aftermath. However, it should be noted that this is a short data interval for measuring the long-run component.

5. NEOCLASSICAL MODELS OF DEPRESSIONS

This section describes neoclassical models of *depressions*, which are prolonged periods in which aggregate economic activity is far below trend. Kehoe and Prescott (2007) define a Great Depression as an event in which per capita real output is at least 20% below trend, in which trend is constructed using a 2% annual growth rate. They also require that real output is at least 15% below this trend within a decade, and that real output always grows at less than 2% per year during the episode.

Neoclassical modeling of depressions has become a very active research field in the last 15 years and is providing new insights into several episodes that have long been considered economic pathologies.[1] Some of the models presented here are tailored to capture features of specific episodes, but all of these models can be modified to study other episodes of depressed economic activity.

This section focuses on the US Great Depression, which is the most widely-studied depression in the literature, and is perhaps the most striking and anomalous period of macroeconomic activity in the economic history of the US. The Great Depression began in the Fall of 1929, and the economy did not recover to its predepression trend until the early 1940s.

Lucas and Rapping (1969) developed the first modern model of the US Great Depression. This model represented a breakthrough by analyzing the Depression within an equilibrium framework. Previous studies of the Depression noted the coincidence of deflation and depression in the early 1930s, and viewed deflation as causing the Depression. The Lucas–Rapping model provided a very different interpretation of this relationship. In the Lucas–Rapping model, deflation depresses output through imperfect information about nominal price changes. Specifically, workers misinterpret falling

[1] Recent models of the Great Depression analyze a number of policies and mechanisms in order to understand this episode. This includes the wage fixing and work-sharing policies of Herbert Hoover (Ohanian, 2009; Ebell and Ritschl, 2008; and Amaral and MacGee, 2015), the worker-industry cartels of the National Industrial Recovery Act and the National Labor Relations Act (Cole and Ohanian, 1999, 2004), changes in capital income tax rates (McGrattan, 2012), the cartel policies of Mussolini in Italy, and Hitler in Germany (Cole and Ohanian, 2016), the impact of tariffs on resource allocation and productivity (Bond et al., 2013), the impact of financial market imperfections and misallocation in the Depression (Ziebarth, 2014), and the impact of contractionary monetary policy on labor markets (Bordo et al., 2000).

nominal wages as reflecting a lower relative price for their labor services. This mistaken perception of the real wage leads to lower employment and lower output. This change in employment and production reflects intertemporal substitution, in which employment and output expand during periods in which workers perceive high real wages and contract during periods of perceived low real wages. The mechanism of imperfect information and nominal price changes was developed further in Lucas's 1972 seminal contribution that rationalized Phillips Curve type relationships within an optimizing model.

Lucas and Rapping's study spawned a large neoclassical literature on fluctuations that focused on intertemporal substitution as the principal channel for understanding business cycle fluctuations. This literature includes contributions by Barro (1981), Barro and King (1984), Lucas (1973a), Sargent (1973), Sargent and Wallace (1975), among others.

But many economists were skeptical of these early neoclassical interpretations of fluctuations, particularly for deep and prolonged crises such as the US Great Depression. Modigliani (1977) argued that neoclassical models of the Depression implausibly portrayed individuals as exhibiting a "a severe attack of contagious laziness" (p. 24). Modigliani, Rees (1970) and many other economists interpreted the substantial job loss of the Depression as involuntary unemployment, which stands in sharp contrast to the market-clearing equilibrium interpretation of Lucas and Rapping. The Modigliani quip has been repeated frequently over time, and is viewed widely as a fundamental critique of neoclassical macroeconomic modeling. This section presents neoclassical models of the Depression that directly confront Modigliani's criticism. The analysis shows how simple neoclassical models can be extended to assess economies with market distortions that create substantial and persistent involuntary job loss.

5.1 The Depth, Duration, and Sectoral Differences of the US Great Depression

The depth, duration, and sectoral differences in severity of the Depression represent a significant challenge for neoclassical models, or for any quantitative theoretic model. Tables 7–9 summarize these features by presenting data on output, consumption, investment, hours worked, and productivity. The data in these tables are divided by the population. In addition, all of the data except for hours worked are detrended at 2% per year. Thus, the value of 100 means that a variable is equal to its steady state growth path value.

Table 7 shows that real GDP declines by more than 35% between 1929 and the Depression's trough in 1933, and remains far below trend after that. Consumption also falls considerably, and remains near its trough level after 1933. Investment declines by about 75%, and remains at 50% below trend by the late 1930s. Hours worked decline about 27% between 1929 and 1933, and remain more than 20% below trend after that.

Total factor productivity (TFP) declines by about 14% below trend by 1933. Such a large drop in productivity raises questions about measurement, and whether this decline reflects factors other than changes in efficiency. Ohanian (2001) found that this TFP

Table 7 US Great Depression levels of real output and its components (index, 1929 = 100)

Year	Real output	Consumption: Nondurables and services	Consumption: Consumer durables	Business investment	Government purchases	Foreign trade: Exports	Foreign trade: Imports
1930	87.4	90.9	76.2	79.2	105.1	85.3	84.9
1931	78.1	85.4	63.4	49.4	105.4	70.6	72.4
1932	65.2	76.0	46.7	27.9	97.3	54.5	58.1
1933	61.9	72.2	44.4	24.6	91.7	52.8	60.8
1934	64.6	72.1	49.0	28.4	101.1	52.8	58.3
1935	68.1	73.1	58.9	34.4	100.1	53.8	69.3
1936	74.9	77.0	70.8	45.9	113.9	55.1	71.9
1937	76.0	77.2	72.2	53.6	106.3	64.3	78.3
1938	70.6	74.3	56.3	37.8	112.0	62.8	58.6
1939	73.5	75.0	64.3	40.5	112.9	61.7	61.6

Data are measured in per capita terms and detrended.

Table 8 Five measures of labor input during US Great Depression (index, 1929 = 100)

Year	Aggregate measures: Total employment	Aggregate measures: Total hours	Aggregate measures: Private hours	Sectoral measures: Farm hours	Sectoral measures: Manufacturing hours
1930	93.8	92.0	91.5	99.0	83.5
1931	86.7	83.6	82.8	101.6	67.2
1932	78.9	73.5	72.4	98.6	53.0
1933	78.6	72.7	70.8	98.8	56.1
1934	83.7	71.8	68.7	89.1	58.4
1935	85.4	74.8	71.4	93.1	64.8
1936	89.8	80.7	75.8	90.9	74.2
1937	90.8	83.1	79.5	98.8	79.3
1938	86.1	76.4	71.7	92.4	62.3
1939	87.5	78.8	74.4	93.2	71.2

Data are measured in per capita terms.

decline was not easily reconciled with capacity utilization, labor hoarding, or compositional shifts in inputs, which suggests significant efficiency loss during this period. TFP recovers quickly and ultimately rises above trend by the late 1930s. This rapid productivity growth after 1932 led Field (2003) to describe the 1930s as "the most technologically progressive decade of the 20th century."

The severity of the Depression differed considerably across sectors. Table 8 shows that manufacturing hours declined enormously, but agricultural hours remained close to trend through the mid-1930s. These two sectors account for roughly 50% of employment at that time.

Table 9 Productivity and real wage rates during US Great Depression (index, 1929 = 100)

		Total factor productivity		Real wage rates		
Year	Labor productivity[a]	Private domestic	Private nonfarm	Total	Manufacturing	Nonmanufacturing
1930	95.3	94.8	94.8	99.3	101.9	98.2
1931	95.2	93.4	92.0	98.9	106.0	96.1
1932	89.4	87.6	85.8	95.8	105.3	92.3
1933	84.8	85.7	82.7	91.3	102.5	87.2
1934	90.3	93.1	92.7	95.7	108.8	91.1
1935	94.8	96.3	95.3	95.1	108.3	90.4
1936	93.7	99.5	99.5	97.6	107.2	94.1
1937	95.1	100.1	99.3	97.8	113.0	92.5
1938	94.6	99.9	98.1	99.1	117.4	92.8
1939	95.2	102.6	100.1	100.1	116.4	94.3

Data are detrended.
[a] Labor productivity is defined as output per hour.

The data summarized here challenge long-standing views of the Depression. Traditional studies omit productivity, and focus instead on monetary contraction and banking crises as the key determinants of the Depression (see Friedman and Schwartz, 1963 and Bernanke, 1983).

However, these factors cannot account for the early stages of the Depression, nor can they account for the post-1933 continuation of the Depression. In terms of the early stages of the Depression, industrial production declined by about 35% between the Fall of 1929 through November of 1930, but there were neither banking crises nor significant monetary contraction during this time.[m]

After 1933, the money stock expanded rapidly and banking crises were quickly eliminated by the introduction of bank deposit insurance. The Lucas–Rapping model and New Keynesian models, such as Eggertsson (2012), counterfactually predict a very rapid recovery to trend as a consequence of rapid monetary expansion and the end of banking crises. In the Lucas–Rapping model, monetary expansion stops deflation, and employment expands as workers perceive that the relative price of their labor services has recovered. In New Keynesian models, such as Eggertsson (2012), inflation moves the economy away from the zero lower interest rate bound, and hours worked increase substantially. These models cannot account for the failure of hours to remain significantly depressed after 1933. Rees (1970) and Lucas and Rapping (1972) discuss the failure of the Lucas and Rapping model to account for hours worked after 1933, and Ohanian (2011) discusses the failure of the Eggertsson model to account for hours worked after 1933.

[m] Ohanian (2010) discusses the immediate severity of the Great Depression that occurred before monetary contraction and before banking crises.

Moreover, the traditional view of the Depression counterfactually implies that the agricultural sector and the manufacturing sector were identically depressed. The large differences between these two sectors mean that any successful model of the Depression must account for the enormous manufacturing depression, but only a modest agricultural decline.

5.2 Diagnosing Depressions with Simple Neoclassical Models

Cole and Ohanian (1999) advocate using simple neoclassical models to *diagnose depressions*. Their idea is that both the successes and the deviations between model and data are informative for developing theories of specific episodes. Cole and Ohanian (1999) focused on the contribution of TFP for the Depression within a standard one-sector stochastic growth model for the 1930s.[n] They fed TFP shocks from 1930 to 1939 into the model and found that the TFP drop accounts for about 60% of the drop in output between 1929 and 1933, and about half of the drop in labor. However, the model generates a completely counterfactual path for the economy after 1933. The rapid recovery of TFP generates a rapid recovery in the model, with labor input recovering to trend by the mid-1930s. In contrast, the actual economy appears to have shifted onto a lower steady state growth path after 1933, with consumption and hours worked remaining near their 1932 trend-adjusted levels.

The post-1933 deviation between model and data provide valuable information about this episode. The results indicate that understanding the post-1933 data requires a large and persistent change in a state variable that substantially depressed and/or restricted the opportunities to produce and trade. The impact of the missing factor must be sufficiently large, such that it prevents recovery in hours worked, despite rapid productivity recovery and despite the low capital stock.

Business cycle accounting (BCA) is another neoclassical diagnostic tool, and its application provides insight regarding this state variable. Cole and Ohanian (1999, 2002), Mulligan (2005), Brinca et al. (2016), and Chari et al. (2007) use a standard one-sector neoclassical model to measure which of the decision margins in that model deviate from theory when actual data is substituted into the first order conditions of the model. For the Great Depression, the condition that equates the marginal rate of substitution between consumption and leisure to the marginal product of labor is significantly distorted. Specifically, the marginal product of labor is higher than the marginal rate of substitution throughout the decade. The deviation in this condition, which is typically called a labor wedge, grows further after 1933, and suggests a major factor that distorted the opportunities and/or the incentives to trade labor services.

[n] The idea of large productivity declines during depressions was initially met with skepticism by some economists. This skepticism is based on the narrow interpretation that lower TFP implies that society lost substantial knowledge over a short period of time. More recently, however, economists are interpreting aggregate productivity changes from alternative perspectives. Section 7 discusses this in detail.

Ohanian (2009) identified economic policies that significantly distorted the opportunities to trade labor services by depressing labor market competition and by preventing wages from adjusting. Simon (2001) analyzed "situation wanted" advertisements from the late 1920s and the early 1930s. These situation wanted advertisements are analogous to help wanted advertisements, but from the supply side of the labor market. In these ads, workers would describe their experience and qualifications, and the wage that they were seeking. Simon shows that the supply price of labor—the desired wage posted in the situation wanted ads—was much lower than the wages that were actually paid in the 1930s. This large gap between the supply price of labor and the wage was not present in the late 1920s, however, when the supply price and actual wages paid were very similar. This evidence suggests that wages were above their market–clearing level, which in turn created an excess supply of labor.

Table 9 provides further evidence of a significantly distorted labor market. The table presents wages from manufacturing and from the farm sector. These data are measured relative to trend, which is the average growth rate of productivity in these sectors (see Cole and Ohanian, 1999). These data show that wages in manufacturing are well above trend, which suggests that they are also above their market–clearing level. In contrast, real wages in the farm sector are well below trend.

Given this backdrop, a new neoclassical literature on the Depression has emerged that studies how government policy changes distorted labor markets. Ohanian (2009) studied the downturn phase of the US Great Depression, and Cole and Ohanian (2004) studied the delayed recovery from the Depression. Both papers use neoclassical frameworks that build on the facts described above. Given the large differences in hours worked and wages in the manufacturing and agricultural sectors, these models begin by modifying the standard one-sector growth model to incorporate multiple sectors, and then build in government policies.

5.3 A Neoclassical Model with Wage Fixing and Work-Sharing Policies

There were large shifts in government policies throughout the 1930s that distorted labor and product markets by significantly restricting competition in industrial labor and product markets, but not in agricultural markets. Ohanian (2009) describes how these policies began in November 1929, following the October stock market decline. President Herbert Hoover met with the leaders of the largest industrial firms, including General Motors, Ford, General Electric, US Steel, and Dupont. Hoover lobbied these firms to either raise wages, or at a minimum, to keep wages at their current levels. He also asked industry to share work among employees, rather than follow the typical practice of laying off workers and keeping retained workers on a full-time shift.

In return for maintaining nominal wages and sharing work, organized labor pledged to maintain industrial peace by not striking or engaging in any efforts that would disrupt production. The Hoover bargain was perceived by firms to be in their interest. Specifically, it is widely acknowledged that the major manufacturing firms had substantial market power at this time, with considerable industry rents. Kovacic and Shapiro (2000) note that this period represents the zenith of collusion and cartels among major industry, and capital's share of income was at an all-time high. Industry agreed to keep wages fixed, and Ford Motor in fact raised wages following the meeting with Hoover. However, as the price level declined, and as productivity declined, these fixed nominal industrial wages led to rising real wages and rising unit labor costs. Ohanian (2010) documents that industry asked Hoover several times for permission to reduce nominal wages, but Hoover declined these requests. Nominal wages among the biggest employers did not begin to fall until late 1931, after hours worked in industry had declined by almost 50%.

Ohanian (2009) develops a neoclassical model with a policy of nominal wage fixing and work-sharing that affected the industrial sector. This requires a model with multiple sectors, and also requires a distinction between hours per worker and employment in order to model work-sharing.

There is a representative family, and family members work in many industries. The population grows at rate n. Preferences over consumption and leisure, and the disutility of joining the workforce, are given by:

$$\max \sum_{t=0}^{\infty} \beta^t \{\ln(c_t) + e_{at}\mu \ln(1-h_{at}) + e_{mt}\mu \ln(1-h_{mt}) - \upsilon(e_{at}+e_{mt})\}(1+n)^t. \tag{44}$$

Preferences are scaled by the population, which grows at rate n. Consumption is denoted as c, e_a denotes the number of workers in the agricultural sector, e_m denotes the number of workers in the manufacturing sector, and h_a and h_m denote the length of the workweek in agriculture and manufacturing, respectively. The function $\upsilon(e_a + e_m)$ is increasing and weakly convex, and specifies the utility cost of sending different household members to work in the market. Rank-ordering family members by their position in the distribution of this utility cost, and assuming that these costs rise linearly across family members, yields:

$$-\upsilon(e_{at}+e_{mt}) = -\int_{i=0}^{e_t} (\xi_0 + 2\xi_1 x)dx = \xi_0 e_t + \xi_1 e_t^2. \tag{45}$$

Note that there will be an optimal number of family members working, as well as an optimal number of hours per worker.

There are two production sectors, agriculture and manufacturing, and there is a continuum of industries within each sector. Industry output is given by:

$$y_i = h_i e_s(i)^{\gamma} k_s(i)^{1-\gamma}, \tag{46}$$

in which the length of the workweek is given by h, employment is given by e, and capital is given by k. Kydland and Prescott (1988), Cole and Ohanian (2002), Hayashi and

Prescott (2002), Osuna and Rios-Rull (2003), and McGrattan and Ohanian (2010) use similar production technologies to study problems that require differentiating between employment and hours per worker.

The industry-level outputs are aggregated to produce sectoral output:

$$Y_s = \left(\int_0^1 y_s(i)^\theta di \right)^{\frac{1}{\theta}} \tag{47}$$

Final output, which is divided between consumption and investment, is a CES aggregate over the two sectoral outputs:

$$Y = [\alpha Y_m^\phi + (1-\alpha) Y_a^\phi]^{\frac{1}{\phi}} \tag{48}$$

The production of final goods is competitive, and the maximization problem is given by:

$$\max \{ Y - \int p_m y_m(i) di - \int p_a y_a(i) di \} \tag{49}$$

subject to:

$$Y = \left[\alpha \left(\int_0^1 y_m(i)^\theta di \right)^{\frac{\phi}{\theta}} + (1-\alpha) \left(\int_0^1 y_a(i)^\theta di)^{\frac{\phi}{\theta}} \right) \right]^{\frac{1}{\phi}} \tag{50}$$

The solution to the final good producer's profit maximization problem is standard, and is characterized by equating the marginal product of each intermediate input to the input price.

The parameter values for the household discount factor, the depreciation rate, and the capital and labor production share parameters are standard, with $\beta = 0.95$, $\delta = 0.06$, and $\gamma = 0.67$. The values for the three parameters that govern the disutility of hours per worker (the length of the workweek), and the utility cost of employment, are jointly set to target (i) an average employment to population ratio of 0.7, (ii) the average workweek length at that time, which was about 45 hours per week, and (iii) that employment change accounts for about 80% of cyclical fluctuations in hours worked.

Ohanian (2009) discusses the fraction of the economy affected by the Hoover program, and sets the production share parameter α so that about 40% of employment was produced in industries impacted by this program. The parameter ϕ governs the substitution elasticity between agriculture and manufacturing. This elasticity is set to 1/2, which is consistent with the fact that both the manufacturing share of value added and its relative price have declined over time.

To analyze the impact of the Hoover nominal wage-fixing and work-sharing policy, the observed real manufacturing wage sequence is exogenously fed into the model. This sequence of wages is interpreted as the result of Hoover's fixed nominal wage program in conjunction with exogenous deflation. Note that the analysis is simplified considerably by abstracting from an explicit role of money in the model, such as a cash-in-advance constraint. It is unlikely that the inclusion of explicit monetary exchange in the model would change the results in any significant way, provided that a more complicated model with monetary exchange generated the same real wage path for manufacturing.

We now discuss modeling the workweek for analyzing the Hoover program. First, recall that almost all of the cyclical change in labor input prior to the Depression was due to employment, rather than changes in hours per worker. However, about 40% of the decline in labor input between 1929 and 1931 was due to a shorter workweek. This suggests that the large decline in the workweek length was due to the Hoover work-sharing policy, rather than reflecting an optimizing choice.

The Hoover workweek is also exogenously fed into the model. The evidence that indicates that the workweek was not optimally chosen suggests that the Hoover work-sharing policy was inefficient. In this model, the inefficiency of forced work-sharing results in lower productivity, since reducing the length of the workweek operates just like a negative productivity shock. To see this, note that the Cobb–Douglas composite of employment and the capital stock in the production function is scaled by the length of the workweek.

The analysis is conducted between 1929:4 and 1931:4. The wage-fixing and work-sharing policies significantly depress economic activity by raising the cost of labor, which reflects both a rising real wage and declining labor productivity. The inflexible manufacturing wage means that the manufacturing labor market does not clear, and that the amount of labor hired is solely determined by labor demand. Table 10 shows the perfect foresight model predictions and data.[o] The model generates about a 16% output decline, which accounts for over 60% of the actual decline.[p] The model also is consistent with the fact that there is a much larger decline in manufacturing than in agriculture. Manufacturing hours fall by about 30% in the model and by about 44% in the data, and agricultural hours fall by about 12% in the model and by about 4% in the data.

The agricultural sector declines much less because it is not subject to the Hoover wage and work-sharing policies. However, the agricultural sector declines because of the general equilibrium effects of the Hoover policy. This reflects the fact that manufacturing

[o] The annual NIPA data are linearly interpolated to a quarterly frequency.

[p] The deterministic path solution is the reason for the immediate increase in economic activity. This reflects the fact that producers see higher future labor costs, and thus produce before these costs rise. Future research should assess the impact of these policies in a stochastic environment.

Table 10 US Great Depression—data and model with wage fixing and work sharing policies (index, 1929:3 = 100)

	Output		Manufacturing hours		Agricultural hours	
	Data	**Model**	**Data**	**Model**	**Data**	**Model**
1929:4	97	101	91	96	99	104
1930:1	93	98	84	92	98	102
1930:2	90	96	76	89	99	99
1930:3	87	94	69	85	99	97
1930:4	84	91	67	80	99	94
1931:1	82	87	65	76	98	92
1931:2	78	86	59	71	97	90
1931:3	75	84	56	69	96	88

output is a complement to agricultural output in final goods production. Thus, depressed manufacturing output depresses the agricultural wage, which in turn depresses agricultural hours.

Note that the model is consistent with Simon's (2001) finding of excess labor supply in manufacturing, and that job seekers in manufacturing were willing to work for much less than the manufacturing wage. The model also provides a theory for why deflation was particularly depressing in the 1930s compared to the early 1920s, when a very similar deflation coincided with a much milder downturn.

While this model was tailored to study the US Great Depression, it can be used more broadly to study nominal wage maintenance policies and/or work-sharing policies.

5.4 A Neoclassical Model with Cartels and Insider–Outsider Unions

The model economy with nominal wage-fixing, deflation, and work sharing accounts for a considerable fraction of the early years of the Depression. After 1933, however, deflation ended. Moreover, productivity grew rapidly, and real interest rates declined. These factors should have promoted a strong recovery, but the economy remained far below trend for the balance of the decade. The failure of the economy to return to trend is puzzling from a neoclassical perspective, given productivity growth, and it is puzzling from a Keynesian perspective, given the end of deflation and banking crises, and given much lower real interest rates.

The empirical key to understanding the post-1933 Depression is a growing labor wedge, as the marginal product of labor was far above the marginal rate of substitution between consumption and leisure. Cole and Ohanian (2004) develop a theory of the labor wedge that is based on changes in government competition and labor market policies. One policy was the 1933 National Industrial Recovery Act, which allowed a number of nonagricultural industries to explicitly cartelize by limiting production and raising prices. The government typically approved these cartels provided that industry raised the wages of their workers. Another policy was the 1935 National Labor Relations Act (NLRA), which provided for unionization and collective bargaining. The use of the "sit-down" strike under the NLRA, in which striking workers forcibly prevented production by taking over factories, gave workers considerable bargaining power. Cole and Ohanian describe how both of these policies created an insider–outsider friction, in which insiders received higher wages than workers in sectors that were not covered by these policies.

Cole and Ohanian present industrial wage and relative price data from individual industries covered by these policies. Industry relative prices and wages jumped around the time that the industry codes were passed, and continued to rise after that. Table 9 shows that real wages rise and ultimately are about 17% above trend by the late 1930s.

Cole and Ohanian (2004) develop a multisector growth model in which the industries in the manufacturing sectors are able to cartelize provided that they reach a wage agreement with their workers. They begin with a simple neoclassical environment, and then add in cartelization policies and a dynamic, insider–outsider model of a union, in which incumbent workers (insiders) choose the size of the insider group, and bargain over the wage. The objective of the insiders is to maximize the per-worker expected, present discounted value of the union wage premium.

While this model was developed to capture specific features of US policy, it easily can be modified to analyze a variety of dynamic bargaining games in which a firm and a union repeatedly negotiate over wages, and in which the insiders choose their size by maximizing the expected, discounted payoff to union membership. The choice of the size of the union is central in any insider–outsider environment, but is typically missing from earlier insider–outsider models.

We begin with a neoclassical, multisector growth model, and then build in these policies. Preferences are given by:

$$\max \sum_{t=0}^{\infty} \beta^t \{\ln(c_t) + \mu \ln(1 - n_t)\}. \tag{51}$$

Consumption is denoted as c, and the size of the household is normalized to 1. The model is simplified by assuming that work is full-time. The term $1 - n$ is the number of household members who are engaged in nonmarket activities (leisure). The household faces a present value budget constraint:

$$\sum_{t=0}^{\infty} Q_t \left[w_{ft} n_{ft} + w_{mt} n_{mt} + \Pi_0 - c_t - \sum_s r_{st} k_{st} - x_{st} \right] \geq 0, \tag{52}$$

in which Q_t is the date-t price of output, w_f is the competitive (noncartel) wage, n_f is the number of workers in the competitive sector, w_m is the cartel wage, n_m is the number of workers in the cartel sector, Π_0 are date zero profits, r_s is the rental price of sector s capital, which in turn is denoted as k_s, and x_s is investment in sector s capital. Time allocated to market activities is given by:

$$n_t = n_{ft} + n_{mt} + n_{ut}. \tag{53}$$

This indicates that total nonmarket time, n, is the sum of household time spent working in the agricultural (noncartel) sector, n_f, the time spent working in the manufacturing (cartel) sector, n_m, and the time spent searching for a job in the manufacturing sector, n_u.

There is also a law of motion for the number of workers in the cartel sector. This transition equation is given by:

$$n_{mt} \leq \pi n_{mt-1} + \upsilon_{t-1} n_{ut-1} \tag{54}$$

The transition equation for the number of workers in the manufacturing sector indicates that the number of these manufacturing workers at date t consists of two components. One is the number who worked last period, less exogenous worker attrition, in which $(1 - \pi)$ is the probability of a manufacturing worker exogenously losing their manufacturing job. The other component is $\upsilon_{t-1} n_{ut-1}$, and this is the number of new workers hired into manufacturing jobs. This is equal to the number of family members who searched for a manufacturing job in the previous period, n_{ut-1}, multiplied by the probability of finding a manufacturing job, which is denoted as υ_{t-1}.

Note that job search is required for an outsider to be newly hired into manufacturing. This search process captures competition by the outsiders in the model for the scarce insider jobs. The insider attrition probability, $1 - \pi$, captures features that generate job loss, but that are not explicitly modeled, such as retirement, disability, and relocation. Note that if $\pi = 1$, then there is no insider attrition, and there will be no hiring (or job loss) in the cartel sector in the steady state of the model.

The law of motion for industry capital stocks is standard, and is given by:

$$k_{st+1} = (1 - \delta) k_{st} + x_{st} \tag{55}$$

Industry output in sector i is given by:

$$y(i)_t = z_t k_t^{\gamma}(i) n_t^{1-\gamma}(i) \tag{56}$$

Sector output is given by:

$$Y_s = \left[\int_{\varphi_{s-1}}^{\varphi_s} y(i)^{\theta} di \right]^{\frac{1}{\theta}}, s = \{f, m\} \tag{57}$$

Final output is given as a CES aggregate of the two sectoral outputs:

$$Y = [\alpha Y_f^{\phi} + (1-\alpha) Y_m^{\phi}]^{\frac{1}{\phi}} \tag{58}$$

Producers in the cartel sector have a profit maximization problem that features their market power, and which depends on the elasticity parameters ϕ and θ. Using the fact that industry price is given by $p = Y^{1-\phi} Y_m^{\phi-\theta}$, the industry profit function is given by:

$$\Pi = \max_{n,k} \{ Y^{1-\phi} Y_m^{\phi-\theta} ((z_t n_t)^{1-\gamma} k_t^{\gamma})^{\theta} - wn - rk \} \tag{59}$$

In the insider–outsider union model, the objective for an incumbent worker (insider) is to maximize the expected present discounted value of industry wage premia. The value of being an insider, in which there are currently n insiders, is given by:

$$V_t(n) = \max_{\bar{w}_t, \bar{n}_t} \{ \min \left[1, \frac{\bar{n}}{n} \right] ([\bar{w}_t - w_{ft}) + \pi \left(\frac{Q_{t+1}}{Q_t} \right) V_{t+1}(\pi \bar{n})] \} \tag{60}$$

The insiders propose to the firm to hire $\bar{n}$ number of workers at the wage rate $\bar{w}_t$. If the offer is accepted, the current period payoff to each insider is the wage premium, which is the cartel wage less the competitive wage: $(\bar{w}_t - w_f)$. The insider's continuation value is the expected discounted value of being an insider next period, which is $\pi \left(\frac{Q_{t+1}}{Q_t} \right) V_{t+1}(\pi \bar{n})$. Note that the number of insiders at the start of period $t+1$ is given by $\pi \bar{n}$. Note that the attrition probability, π, affects the continuation value of union membership in two different ways. First, the probability that any individual insider at date t will remain in the cartel at date $t+1$ is π, which scales the date $t+1$ value function. Second, the total number of date t insiders who will remain in the cartel at date $t+1$ is $\pi \bar{n}$.

The insiders bargain with the firm at the start of each period. If a wage agreement is reached, then the firm hires $\bar{n}$ number of workers at wage $\bar{w}$. Note that the union's offer is efficient in the sense that given the wage offer, the number of workers hired, $\bar{n}$, is consistent with the firm's labor demand schedule. The bargaining protocol is that the union makes a take-it-or-leave-it offer to the firm.

In equilibrium, the union makes an offer that the firm weakly prefers to its outside option of declining the offer. The firm's outside option is given as follows. If the offer is declined, then the firm can hire labor at the competitive wage, w_f. With probability ω the firm will be able to continue to act as a monopolist. With probability $1 - \omega$, the government will discover that the firm did not bargain in good faith with the union, and the government will force the firm to behave competitively and thus the firm earns no monopoly profits.

This feature of the model empirically captures the fact that some firms did fail to reach wage agreements, or violated wage agreements, and that the government did enforce the wage bargaining provisions of the policy. The firm's outside option therefore is the expected level of monopoly profits earned by declining the insider's offer, and the firm will only accept the insider's offer of $(\bar{n}, \bar{w})$ if it delivers at least that level of profit. It is therefore optimal for the union to make an offer that does provide the firm with its outside option.

A key parameter in this model is the share of employment in the cartelized sector. While the cartel policy was intended to cover about 80% of the nonfarm economy, there is debate regarding how much of the economy was effectively cartelized. Therefore, the model conservatively specifies that only manufacturing and mining were cartelized, which is about 1/3 of the economy. Another key parameter is ω, which governs the probability that the government will identify a firm that breaks their wage agreement. This value was chosen so that the steady state cartel wage premium is about 20% above trend. This implies that ω is around 0.10. The attrition parameter, π, is set to 0.95, which yields an average job tenure in the cartel of 20 years.

Other parameters include the substitution elasticity across industries and across sectors. For these parameters, the industry substitution elasticity is picked so that the industry markup would be 10% in the absence of wage bargaining. The sectoral substitution elasticity, which refers to the substitution possibility between manufacturing and the farm sector, is picked to be 1/2. Other parameter values, including the household discount factor, the household leisure parameter, the income shares of capital and labor, and depreciation rates, are standard, and are described in Cole and Ohanian (2004).

The quantitative analysis begins in 1934. To generate model variables, the 1933 capital stocks from the manufacturing and farm sectors from this are specified, and the sequence of TFP from 1934 to 1939 is fed into the model. The model variables then transit to their steady state values. For comparative purposes, we show the results from the cartel model to those from the perfectly competitive version of this model. Table 11, which is taken from Cole and Ohanian (2004), shows the response of the competitive version of this model. Note that the rapid return of productivity to trend fosters a rapid recovery under competition, with hours worked rising above trend to rebuild the capital stock to its steady state level. Moreover, the wage is well below trend in 1933, and then recovers quickly after that, as both productivity and the capital stocks rise.

Table 11 Equilibrium path of recovery from depression in competitive model

	Output	Consumption	Investment	Employment	Wage
1934	0.87	0.90	0.73	0.98	0.89
1935	0.92	0.91	0.97	1.01	0.91
1936	0.97	0.93	1.18	1.03	0.94
1937	0.98	0.94	1.14	1.03	0.95
1938	0.98	0.95	1.12	1.02	0.96
1939	0.99	0.96	1.09	1.02	0.97

Table 12 shows the transition of the cartel model. This transition stands in sharp contrast to the transition in the competitive economy from Table 11. The cartel economy transits to a steady state that is well below the competitive economy. Despite rising productivity, the cartel economy remains depressed through the 1930s, as cartel policies create rents that raise wage rates far above trend, despite the fact that both consumption and time allocated to market activities are below trend. These results indicate that the cartel policy accounts for about 60% of the post-1933 Depression in output, consumption, and hours worked.

5.5 Neoclassical Models of Taxes and Depressions

This section describes how tax rate changes contributed to the US Great Depression and also for more recent episodes of depressed economic activity.

Tax rates rose in the United States during the Great Depression. McGrattan (2012) studies how changes in tax rates on dividends and corporate profits affected economic activity after 1933. Specifically, a new tax rate was applied to undistributed corporate profits in 1936. The goal of this new tax was to increase corporate payments to shareholders, which in turn was expected to stimulate spending.

McGrattan analyzes a representative household economy with log preferences over consumption and leisure, and with a standard constant returns to scale Cobb–Douglas production function with capital and labor inputs. She considers two formulations for taxes. In the traditional formulation, tax rates are applied to labor income (τ_h) and to capital income net of depreciation (τ_k). Tax revenue is the sum of labor income tax revenue and capital income tax revenue:

$$\tau_h wh + \tau_k(r-\delta)k \tag{61}$$

The alternative formulation includes a finer decomposition of taxes across revenue sources, and distinguishes between business and nonbusiness capital. Tax revenue in this alternative formulation is given by:

$$\begin{aligned}\tau_h wh &+\tau_p(r-\tau_k-\delta)k_b+\tau_c c+\tau_k k_b+\tau_u(k_b'-k_b)\\ &+\tau_d\{(rk_r-x_b)-\tau_p(r-\tau_k-\delta)k_b-\tau_k k_b-\tau_u(k_b'-k_b)\}\end{aligned} \tag{62}$$

In (64), τ_p is the tax rate on profits, τ_k is now the tax rate on business property, τ_c is the consumption tax rate, τ_u is the tax rate on undistributed profits, τ_d is the dividend tax rate, and primed variables refer to period $t+1$ values.

The intertemporal first order condition that governs efficient investment shows how changes in expected taxation affect investment:

$$\frac{(1+\tau_{ut})(1-\tau_{dt})}{(1+\tau_{ct})c_t}=\beta E_t\left[\frac{(1-\tau_{dt+1})}{(1+\tau_{ct+1})c_{t+1}}\{(1-\tau_{pt+1})(r_{t+1}-\tau_{kt+1}-\delta)+(1+\tau_{ut+1})\}\right] \tag{63}$$

Table 12 Equilibrium path of recovery from depression in cartel policy model

						Employment		Wage	
	Output	**Consumption**	**Investment**	**Employment**	**Searchers**	**Cartel sector**	**Competitive sector**	**Cartel sector**	**Competitive sector**
1934	0.77	0.85	0.40	0.82	0.07	0.68	0.89	1.16	0.81
1935	0.81	0.85	0.62	0.84	0.11	0.69	0.92	1.19	0.83
1936	0.86	0.85	0.87	0.89	0.06	0.72	0.97	1.20	0.83
1937	0.87	0.86	0.90	0.90	0.04	0.73	0.98	1.20	0.83
1938	0.86	0.86	0.86	0.89	0.06	0.72	0.97	1.20	0.84
1939	0.87	0.86	0.88	0.89	0.04	0.73	0.97	1.20	0.84

Note that dividend taxes and consumption taxes in (65) do not distort investment incentives at the margin in the deterministic version of this model when these tax rates are constant over time. However, expected changes in tax rates will affect investment decisions. An expected increase in these tax rates reduces the expected returns to investment, and leads firms to increase current distributions. Tax rates rose considerably in the mid-1930s, with the dividend tax rate rising from about 14% to about 25%, the corporate profit tax rate rising from about 14% to about 19%, and the newly implemented undistributed tax rate of 5%. McGrattan shows that plausible expectations of these tax rate changes can help account for the fact that business investment remained at 50% or more below trend after 1933.

McGrattan's analysis of the US Great Depression focused on changes in capital income tax rates. Prescott (2004) and Ohanian, Raffo, and Rogerson (2008) analyze how long-run changes in labor income tax rates have affected hours worked more recently. Ohanian et al. (2008) document that hours worked per adult in the OECD vary enormously over time and across countries. Hours worked in many Northern and Western European countries declined by about 1/3 between the 1950s and 2000, including a nearly 40% decline in Germany.

Ohanian et al. use a standard neoclassical growth model with log preferences over consumption, log preferences over leisure, a flat rate labor income tax, and a flat rate consumption tax rate. The economy's technology is a constant returns to scale Cobb–Douglas production function that uses capital and labor, which is given by $Y_t = A_t K_t^{\theta} H_t^{1-\theta}$. Preferences for the representative family are given by:

$$\max \sum \beta^t \{\alpha \ln(c_t - \bar{c} + \lambda g_t) + (1-\alpha)\ln(\bar{h} - h_t)\}. \tag{64}$$

Households value private consumption, c, and public consumption, g. The term $\bar{c}$ is a subsistence consumption term to account for possible nonhomotheticities in preferences that may affect trend changes in hours worked. The parameter $\lambda, 0 < \lambda \leq 1$, governs the relative value that households place on public spending. The specification that government consumption (scaled by the parameter λ) is a perfect substitute for private consumption follows from the fact that much government spending (net of military spending) is on close substitutes for private spending, such as health care.

The first order condition governing time allocation in this economy is standard, and equates the marginal rate of substitution between consumption and leisure to the wage rate, adjusted for consumption and labor income taxes. This first order condition is presented below. Note that the marginal product of labor, $(1-\theta)\frac{Y_t}{H_t}$ is substituted into the equation for the wage rate in (67):

$$\frac{(1-\alpha)}{\bar{h} - h_t} = \frac{(1-\tau_{ht})}{(1+\tau_{ct})} \frac{\alpha}{(c_t + \lambda g_t)} (1-\theta) \frac{Y_t}{H_t}. \tag{65}$$

In the first order condition, τ_h is the labor income tax rate, and τ_c is the consumption tax rate. Ohanian et al. feed McDaniel's (2011) panel data construction of consumption and income tax rates into this first order condition, along with actual labor productivity and consumption data. They choose the value of α by country so that model hours in the first year of the dataset are equal to actual hours for each country. They set $\lambda = 1$, and labor's share of income is set to 0.67. The subsistence consumption term is set to 5% of US consumption in 1956, which represents a small departure from the standard model of homothetic preferences. Ohanian et al. describe the sensitivity of results to alternative values for these parameters.

With these parameter values and data, Ohanian et al. use this equation to construct a predicted measure of hours worked from the model economy, and compare it to actual hours worked by country and over time. Fig. 25 shows actual hours worked and

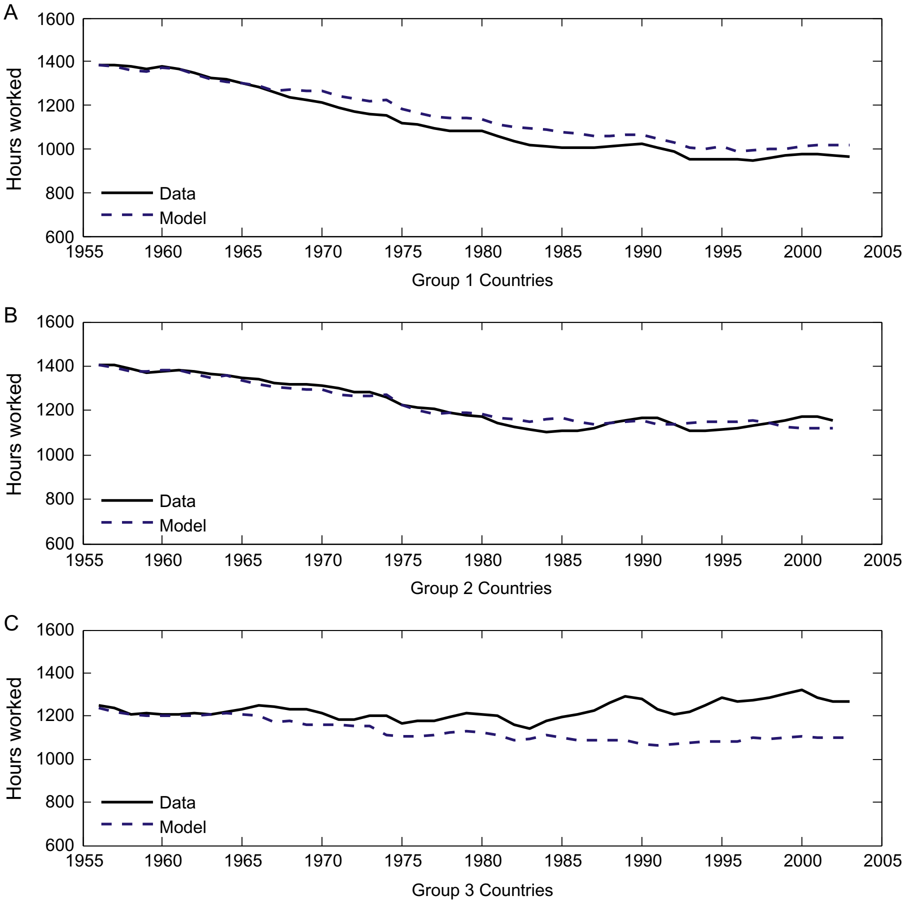


Fig. 25 Comparing OECD hours worked, model and data.

predicted hours worked from the model for 21 OECD countries.[q] Panel (A) of the graph shows results for countries which experienced at least a 25% decline in hours worked per capita. Panel (B) shows results for countries which experienced a decline in hours per capita that range between 10% and 25%. Panel (C) shows results for countries that experienced a decline in hours per capita of less than 10%, or alternatively experienced higher hours.

The figures show that the model economy accounts for much of the secular decline in hours worked, particularly for the countries which experienced the largest hours declines. Ohanian et al. also report that the contribution of tax rate changes to changes in hours worked is not sensitive to other labor market factors that may have affected hours, such as changes in employment protection policies, changes in union density, and changes in unemployment benefits.

These findings indicate that the observed increases in labor and consumption tax rates can account for the large observed declines in hours worked per adult across these countries. These neoclassical findings regarding the impact of tax rates on hours worked stand in contrast to other explanations of the decline in European hours. Other explanations include a preference shift for more leisure, or a preference shift in conjunction with policies that restrict work, and that may have been chosen in order for society to coordinate on a low-work equilibrium (see Blanchard, 2004 and Alesina et al., 2006).[r]

5.6 Summary

Depressions, which are protracted periods of substantial economic decline relative to trend, have been difficult to understand and are often presumed to extend beyond the scope of neoclassical economics. The models developed here show that government policies that depress competition can account for a considerable amount of the Great Depression, and can also account for much of the failure of economic activity to return to trend. More broadly, these models of the US Great Depression successfully confront the frequently cited view of Modigliani (1977) that neoclassical models cannot plausibly account for the behavior of labor markets during Depressions.

Modigliani interpreted the Great Depression as the failure of the market economy to right itself. This view, and associated Keynesian views of the Depression, are based on the idea that business organizations did not expand investment in the 1930s, which in turn kept employment low. The studies discussed here turn that interpretation on its head. Specifically, these new neoclassical studies indicate that the depth and persistence of the Depression was the consequence of government policies that depressed the steady

[q] Ohanian et al. (2008) describe the data sources and data construction in detail. The Group 1 countries are Austria, Belgium, Denmark, France, Finland, Germany, Italy, and Ireland. The Group 2 countries are Japan, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom. The Group 3 countries are Australia, Canada, Greece, New Zealand, and the United States.

[r] Other neoclassical studies of taxes and labor supply include Erosa, Fuster, and Kambourov (2012) Rogerson (2009), Ragan (2013), Meza (2008), Samaniego (2008), Dalton (2015), and Davis and Henrekson (2005).

state allocation of time to market work. A lower steady state level of market hours reduced the return to capital, which in turn depressed capital accumulation.

Neoclassical models can also account for more recent periods of depressed economic activity. This includes not only the secular decline in market hours worked in much of Northern and Western Europe through higher tax rates, but also the Finish Depression of the early 1990s that reflects the trade impact of the breakup of the USSR. (Gorodnichenko et al., 2012), and tax changes and productivity changes (Conesa et al., 2007). Other studies of recent Depressions include the Korean Crisis of 1998 (Otsu, 2008), and several case studies in Kehoe and Prescott (2007).

The Depression methodology presented in this section has also been used to study the flip side of Depressions, which are Growth Miracles. This includes studies of Ireland's Growth Miracle (see Ahearne et al., 2006, who analyze a standard growth model with TFP, and Klein and Ventura (2015), who study a small open economy model with taxes, labor wedges, and TFP), and Lu (2012), who analyzes the development of some East Asian countries in a neoclassical framework.

6. NEOCLASSICAL MODELING OF LARGE FISCAL SHOCKS: THE US WORLD WAR II ECONOMY

Wartime economies are interesting and important macroeconomic episodes because they feature very large, exogenous changes in government policies, particular fiscal policies, as well as large changes in macroeconomic activity. The World War II economy in the United States represents perhaps the largest fiscal policy shift of any advanced economy. This includes a nearly 400% increase in federal government spending, large increases in income tax rates, and a large increase in the number of men drafted into military service. Moreover, there was a very large resource reallocation from private use to military use that occurred in a very short period of time.

This striking period of policy changes provides information on how large aggregate and sectoral disruptions quantitatively affect a market economy, which provides a powerful test of neoclassical theory. These episodes are also informative about what a number of economists call the *government spending multiplier*, which refers to the change in output as a consequence of a change in government spending. This research area has received considerable attention since the Great Recession, when the United States and other countries increased government spending to expand economic activity (see Barro and Redlick, 2011; Mountford and Uhlig, 2009; Ramey, 2011; and Taylor, 2011).

Neoclassical analysis of fiscal policies and wars has become an active research area.[s] These studies analyze a range of issues, including the welfare costs of different wartime fiscal policies (Ohanian, 1997), the impact of the draft on economic activity

[s] Studies include Ohanian (1993, 1997), Braun and McGrattan (1993), Siu (2008), Mulligan (2005), McGrattan and Ohanian (2010), Burnside, Eichenbaum, and Fisher (2004), Baxter and King (1993), Christiano and Eichenbaum (1992), Doepke et al. (2015), and Monacelli and Perotti (2008).

(Siu, 2008), the behavior of labor productivity and investment (Braun and McGrattan, 1993), and the extent that a neoclassical model can account for aggregate time series, particularly the impact of wars on the incentives to work (Mulligan, 2005 and McGrattan and Ohanian, 2010).

This section develops a neoclassical model of the World War II US economy to study how well a neoclassical model can fit the wartime US data. The model easily can be applied to other episodes with changes in government spending, transfers, and tax rates. The model is from McGrattan and Ohanian (2010), which in turn draws on Braun and McGrattan (1993), Ohanian (1997), and Siu (2008).

There is a representative family, with two types of family members, civilians and draftees. The size of the family is denoted as N. Both types of family members have identical preferences. At date t, a_t is the number of family members in the military, and $(1 - a_t)$ is the number who are civilians. The family optimally chooses consumption of both types, which is denoted as c_{ct} for civilians, and c_{dt}, for draftees. The family also optimally chooses investment in physical capital, i_{pt}, civilian labor input, l_{ct}, and the accumulation of government bonds, b_{t+1}. The inclusion of public debt follows from the fact that there was considerable debt issue during the war. The labor input of draftees is not a choice variable for the family, but rather is set exogenously by the government, and is denoted by $\bar{l}_d$.

The maximization problem for the representative family is:

$$\max E_0 \sum_{t=0}^{\infty} \{(1-a_t)U(c_{ct}, l_{ct}) + a_t U(c_{dt}, \bar{l}_d)\} N_t \tag{66}$$

Maximization is subject to the following constraints:

$$E_t = (1-\tau_{kt})(r_{pt} - \delta)k_{pt} + (1-\tau_{lt})w_t(1-a_t)l_{ct} + R_t b_t + (1-\tau_{lt})w_t a_t \bar{l}_d + T_t \tag{67}$$

$$E_t = (1-a_t)c_{ct} + a_t c_{dt} + i_{pt} + b_{t+1} \tag{68}$$

$$k_{pt+1} = [(1-\delta)k_{pt} + i_{pt}]/(1+\gamma_n) \tag{69}$$

$$N_t = (1+\gamma_n)^t \tag{70}$$

$$c_c, c_d, i_p \geq 0 \tag{71}$$

Note that k_p is the beginning-of-period capital stock, r_p is the rental price of capital, w is the wage rate, τ_k and τ_l are flat rate tax rates on capital income and labor income, respectively, Rb is the value of matured government debt, and T is government transfers. The depreciation rate is δ. The population grows at the constant rate γ_n.

The production technology is given by:

$$Y_t = F(K_{pt}, K_{gt}, Z_t L_t). \tag{72}$$

The production inputs include private capital, labor, and public capital, K_g. Labor-augmenting productivity is denoted as Z, and is given by:

$$Z_t = z_t(1+\gamma_z)^t. \tag{73}$$

Note that z_t is a transient productivity term and γ_z is the long-run growth rate of technology.

Government purchases consist of 3 components. This is a richer specification of government spending than is typically modeled in fiscal policy studies. Government consumption, C_g is the first component, and this is the standard approach to modeling government purchases. It is common to assume that these wartime purchases of goods do not affect marginal utility or private production possibilities. The second component is government investment, I_g which enhances production possibilities by expanding the capital stock that can be used to produce output. This is typically not modeled in the fiscal policy literature, but is modeled here because of the very large government-funded investments in plant and equipment that occurred in World War II. The government made large investments in the aircraft, automotive, and aluminum industries that raised the manufacturing capital stock by 30% between 1940 and 1945. The third component of government purchases is wage payments to military personnel. Government spending is therefore given by:

$$G_t = C_{gt} + I_{gt} + N_t w_t a_t \bar{l} \tag{74}$$

The evolution of the stock of government capital, which is assumed to have the same depreciation rate as physical capital, is given by:

$$K_{gt+1} = (1-\delta)K_{gt} + I_{gt} \tag{75}$$

The period government budget constraint is given by:

$$B_{t+1} = G_t + R_t B_t - \tau_{lt} N_t w_t((1-a_t)l_{ct} + a_t \bar{l}_d) - \tau_{kt}(r_{pt}-\delta)K_{pt} - r_{gt}K_{gt} + T_t, \tag{76}$$

in which T is a residual lump-sum tax.

A competitive firm maximizes profits, which implies that the rental prices for the factors of production are equal to their marginal productivities. Government debt that is accumulated during the war is retired gradually after the war. The exogenous variables are the tax rates on factor incomes, government consumption and government investment, and the productivity shock. The equilibrium definition of this perfectly competitive economy is standard.

The functional form for preferences is given by:

$$\ln(c) + \frac{\psi}{\xi}(1-l)^{\xi} \tag{77}$$

This specification yields a compensated labor supply elasticity of $\frac{1-l}{(l(1-\xi))}$. McGrattan and Ohanian choose $\xi = 0$ (log preferences) as the benchmark specification. The parameter ψ governs the steady state allocation of time for the household, and is chosen so that model steady state hours is equal to the average time devoted to market work between 1946 and 1960. For military time allocation, they choose $\bar{l}$ such that it matches 50 h per week, which is the average hours for soldiers in basic training (see Siu, 2008). Population growth is 1.5% per year, and the growth-rate of technological progress is 2% per year.

Government capital and private capital are modeled as perfect substitutes. This reflects the fact that much of government investment at this time was in the area of manufacturing plant and equipment:

$$Y_t = F(K_{pt}, K_{gt}, Z_t L_t) = (K_{pt} + K_{gt})^{\theta} (Z_t L_t)^{1-\theta} \tag{78}$$

It is straightforward, however, to modify the aggregator between government and private capital to accommodate government capital that is not a perfect substitute for private capital.

There are six exogenous variables in the model: conscription (the draft) (a_t), the tax rate on capital income (τ_{kt}), the tax rate on labor income (τ_{lt}), government consumption (C_{gt}), government investment(I_{gt}), and productivity (z_t). The evolution of the six exogenous variables is governed by a state vector, S_t, which specifies a particular set of values for these exogenous variables. For 1939–46, these exogenous variables are equal to their data counterparts. The model is solved under different assumptions regarding household expectations about the post-1946 evolution of the exogenous variables. The discussion here focuses on the perfect foresight solution to the model that begins in 1939, and McGrattan and Ohanian discuss the other cases in detail.

While the model described here is based on the World War II US economy, it can be tailored to study other episodes, as it includes a number of features that are relevant for wartime economies, including changes in tax rates on factor incomes, changes in conscripted labor, changes in productivity, government debt issue to help pay for the war, government payments to military personnel, and government investment.

Fig. 26 shows the model's exogenous variables. Government consumption, which includes state and local spending, as well as federal spending, rises from about 14% of steady state output in 1940 to 50% of steady state output by 1944. Government investment rises from about 4% of steady state output in 1940 to about 9% by 1942. The tax rates on labor and capital income, which are average marginal tax rates taken from Joines (1981), also rise considerably, with the labor income tax rates rising from about 8% to about 20%, and with the capital income tax rates rising from about 43% to about 63%. The draft reduces potential labor supply significantly, as almost 12% of the working age population is in the military by 1944.

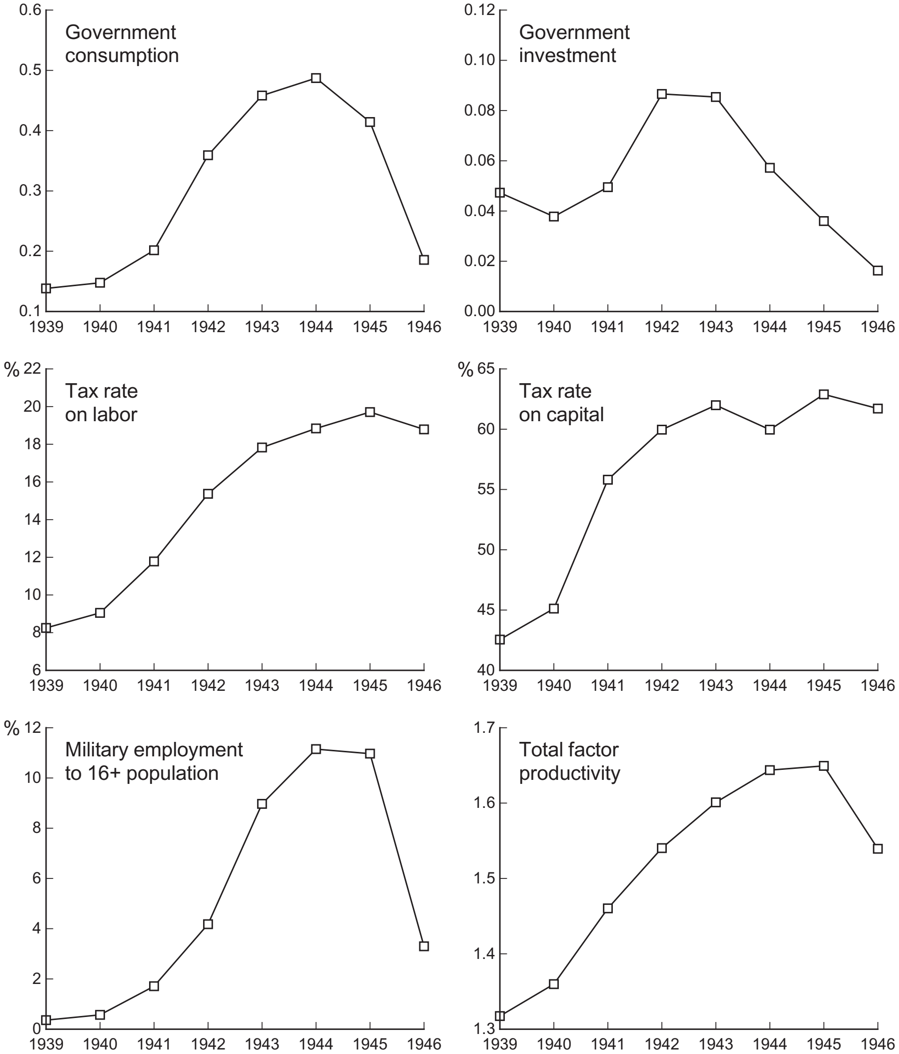


Fig. 26 US government spending, tax rates, draft, and TFP, 1939–46. *Notes*: (1) Government spending series are real and detrended by dividing by the population over 16 and by the growth trend in technology (scaled so the 1946 real detrended level of GNP less military compensation equals 1). (2) Total factor productivity is defined to be $Y/(K^{\theta}L_p^{1-\theta})$, where Y is real, detrended GNP less military compensation, K is real detrended nonmilitary capital stock, L_p is nonmilitary hours worked, and $\theta = 0.38$.

There is a considerable increase in TFP, and there are a number of good reasons why this change actually reflects higher efficiency. This includes the development of federally-funded scientific teams, the development of management science and operations research practices, and a number of technological advances during the 1940s including innovations directly or indirectly fostered by federal R & D expenditures. These include the development of modern airframes, radar, microwave technology, fertilizer, oxygen steel, synthetic rubber, nylon, sulfa drugs and chemotherapy, insecticides, and Teflon and related industrial coatings. Moreover, Herman (2012) describes how business leaders worked together in World War II to mobilize resources and to raise military output through significantly higher efficiency.

The size and diversity of these changes will affect economic activity in a variety of ways. Higher TFP will promote high labor input and output, as will public investment. In contrast, since public investment substitutes for private investment, higher public investment in plant and equipment will tend to reduce private investment. Moreover, rising tax rates and conscription of labor will tend to reduce the incentive to work.

Fig. 27 shows real GNP, real consumption, and real investment, all measured as a percent of trend output. The model output series is very close to actual output, as both increase by more than 50% over the course of the war, and then decline after the war, back to near trend. Model consumption is very flat during the war, and is close to actual consumption. Model investment has a very similar pattern as actual investment. The model investment is somewhat higher than actual investment through 1942, which reflects the perfect foresight solution. Specifically, investment rises considerably in order to build the capital stock by the time that government consumption is high. By 1944, the high level of government investment in plant and equipment, coupled with the enormous resource drain of the war, leads to investment declining significantly. Fig. 28 shows the behavior of total hours worked, and nonmilitary hours, which is the choice variable

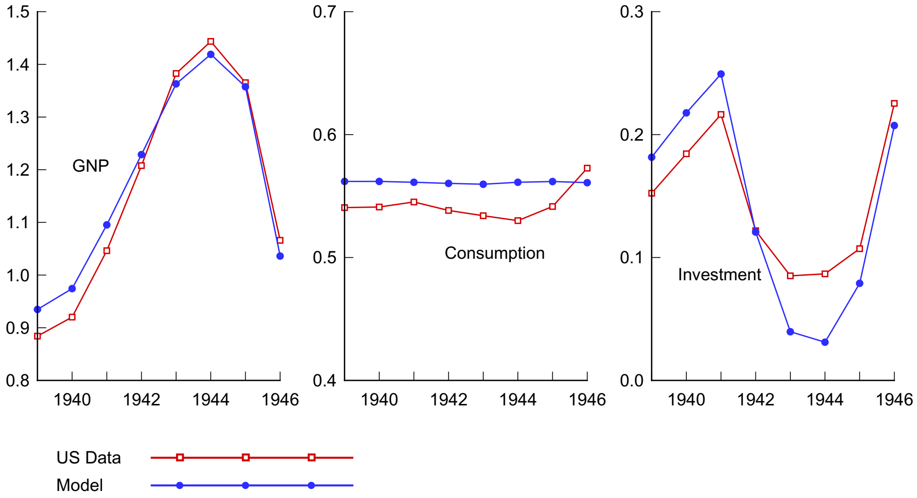


Fig. 27 Real detrended GNP, private consumption, and private investment. *Note*: Data series are divided by the 1946 real detrended level of GNP less military compensation.

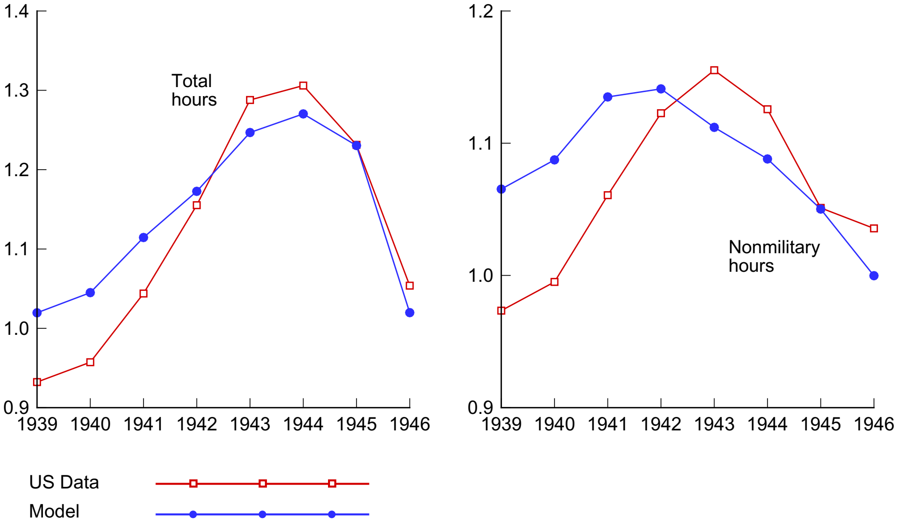


Fig. 28 Per capita total and nonmilitary hours of work, 1939–46. *Note*: Hours series are divided by the 1946–60 US averages.

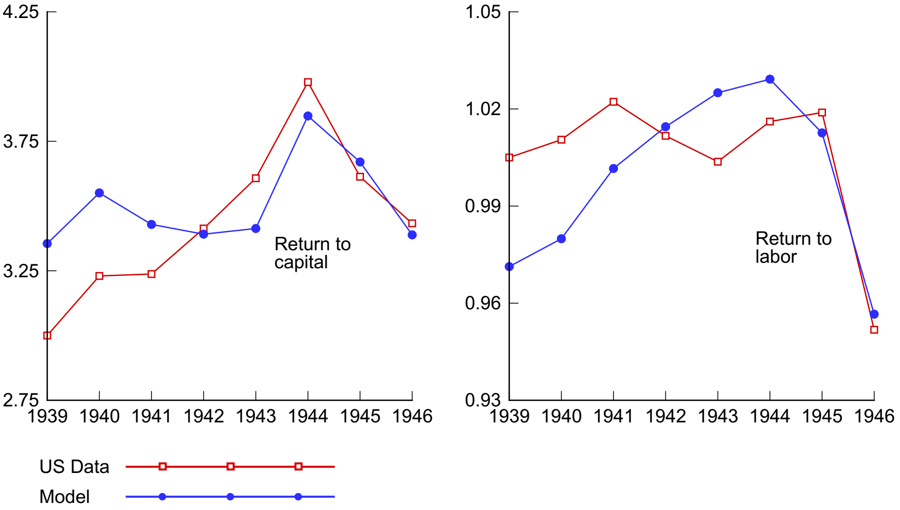


Fig. 29 After-tax returns to capital and nonmilitary labor, 1939–46. *Note*: Return to capital is equal to $100(1-\tau_k)(\theta Y/K-\delta)$. Return to labor is after-tax nonmilitary labor productivity normalized by the 1946–60 US averages.

for the family. Both hours series rise significantly in the data and in the model. The nonmilitary hours in the model rises earlier than in the data, and this again partially reflects the perfect foresight assumption. Fig. 29 shows the after-tax returns to private capital and labor. These are also quite similar to the data.

The dominant factor driving these results is the enormous expansion of government consumption that occurred during the war. This resource drain of wartime government consumption creates a sizeable wealth affect within the model that leads to higher labor input and output, and this effect is much larger than that of any of the other shocks. McGrattan and Ohanian (2010) analyzed the impact of each of the six shocks in the model on hours worked. The impact of just government consumption in the absence of any other shocks raises nonmilitary labor input by about 27% on average between 1943–45. Adding productivity shocks raises this to about a 29% increase. Adding in the draft to these two preceding shocks results in about a 25% increase. Adding in the labor and capital income tax increases has a sizeable depressing effect, and results in an increase in nonmilitary hours of about 10%. Overall, the negative wealth effect arising from government consumption is the dominant factor, followed by the impact of tax increases.

These results shed light on a number of issues that are analyzed in the literature on the macroeconomics of fiscal policy. One issue is regarding the government spending multiplier. A difficulty facing many studies of government spending multipliers is that they are primarily based on peacetime episodes, and episodes even with relatively large peacetime shifts in fiscal policy still involve small changes in fiscal policy compared to policy changes during wartime episodes. Moreover, many of these studies require exogenous changes in fiscal policy, and this can be problematic during peacetime. Consequently, it is challenging to draw sharp conclusions about the size of the multiplier based on peacetime policy changes.

The results from this World War II analysis indicate a multiplier that is considerably less than one. This is informative, not only because the wartime fiscal policy shock is so large, but also because the model explicitly distinguishes between different types of government spending. The analysis conducted here makes it possible to isolate the impacts of different types of spending and taxes on economic activity.

To see that the multiplier from this episode is fairly small, consider the following case in which we account for the impact of all government expenditures, but omit the negative impact of the tax increases and the draft. By omitting these latter two items, we construct the maximum possible effect of fiscal policy, even though tax increases, which depress labor supply, are certainly part of fiscal policy. In this experiment, the World War II episode shows that the multiplier would be about 0.6, reflecting a hypothetical 30% increase in output resulting from government purchases of goods. This multiplier is very similar to Barro and Redlick's (2011) estimates and Mountford and Uhlig's (2009) short-run estimates and is in the lower end of the range of estimates discussed in Ramey (2011).

The results have broader implications regarding neoclassical analyses of large shocks. They indicate that the US economy responded to the enormous wartime economic dislocations, as well as the peacetime reversal of these dislocations, very much along the lines of a simple neoclassical growth model augmented with several large policy changes.

These policy shifts include the massive reallocation of economic activity from peacetime to wartime production, the enormous drain of resources resulting from government purchases, the reduction of the labor endowment through the draft, higher taxes, and government-funded investment. This also includes the rapid unwinding of these unique factors after the war. While this represents just a single episode, this analysis provides a strong test of the neoclassical model in response to large fiscal policy changes.

7. NEOCLASSICAL MODELS OF PRODUCTIVITY SHOCKS

Productivity change is an important feature of the models and the data that we have used to analyze the US historical macroeconomic record in this chapter. This includes a large TFP decline in the Great Depression, a large TFP increase in World War II, and large TFP and equipment-specific productivity fluctuations in the post-Korean War US economy.

There are long-standing questions about the nature and sources of these productivity changes. Much of the profession has viewed TFP declines during downturns, and particularly during depressions, with skepticism, and naturally so. But economists are now analyzing TFP deviations during short-run and longer-run episodes from alternative perspectives than the narrow interpretation that TFP declines reflect a loss of technological know-how and knowledge.

7.1 Resource Misallocation and TFP

Restuccia and Rogerson (2008) analyze the impact of *resource misallocation* on TFP in a competitive economy. The idea is to assess how the misallocation of production inputs across locations affects measured TFP. Their model is related to Hopenhayn and Rogerson (1993), in which there is a representative family and there are different producers, or alternatively, different production locations, each with a decreasing returns to scale technology with potentially different TFP levels, and which are indexed by i. The simplest case of production heterogeneity is the case of a single final good produced at multiple locations, y_i, that is produced with a single production input, labor (h_i). The production relationship at location i is given by:

$$y_i = z_i f(h_i) \tag{79}$$

In this economy, the technology f is twice continuously differentiable, with $f' > 0, f'' < 0$. The term z_i denotes exogenous productivity. Assume that z_i is drawn from the set $\{z_1, z_2, \ldots z_I\}$, and let $\mu(i)$ be the distribution of productivity across these locations.

The efficient allocation of labor requires equating the marginal product of labor across production locations. For the isoelastic technology, $z_i h_i^{\theta}, 0 < \theta < 1$, the efficient allocation of labor between any two locations depends on the differences in productivities at those locations, and the amount of curvature in the production technology:

$$\frac{h_i}{h_j} = \left(\frac{z_i}{z_j}\right)^{\frac{1}{1-\theta}}. \tag{80}$$

We construct an economy-wide measure of TFP by aggregating TFP across all locations. Aggregate TFP in this economy is given by:

$$z = \sum_i z_i^{\frac{1}{1-\theta}} \mu(i)^{1-\theta}. \tag{81}$$

The efficient allocation of labor at any specific location depends on the location's productivity relative to aggregate productivity, as well as the amount of curvature in the technology, and is given by:

$$h_i = \left(\frac{z_i}{z}\right)^{\frac{1}{1-\theta}}. \tag{82}$$

Note that as $\theta \rightarrow 1$, even small differences in productivity generate very large differences in the efficient allocation of production inputs across locations.

Atkeson et al. (1996) use data on differences in worker firing costs and job reallocation rates between the United States and Europe to argue that θ is around 0.85. Restuccia and Rogerson use this value for specifying the level of decreasing returns in their economy, and they study how misallocation of production inputs across locations affects aggregate productivity, z. Resource misallocation means that the marginal product of labor is not equated across production locations, which implies that (82) and (84) are not satisfied.

Restuccia and Rogerson (2008) analyze various government policies that tax the output of some producers, and that subsidize the output of other producers, and they calculate the aggregate productivity and welfare losses from these policies. There is a large literature that has built on Restuccia and Rogerson along many dimensions. This includes the application of misallocation to specific Depressions and Crises (see Oberfield, 2013 and Chen and Irarrazabal, 2013 on the Chilean Depression of the early 1980s, and Sandleris and Wright, 2014 on the Argentinian Depression of 2001), the connection between financial market imperfections and misallocation (see Moll, 2014; Buera and Moll, 2015; and Midrigan and Xu, 2014) and the connection between trade barriers and productivity during the US Great Depression (see Bond et al., 2013). Other studies of misallocation focus on longer-run issues, including studies of the role of misallocation in the development experiences of China and India (Hsieh and Klenow, 2009), entry regulation and productivity (Poschke, 2010), size-dependent policies and productivity (Guner et al., 2008), imperfect information and productivity (David et al., forthcoming), the misallocation of managerial talent and productivity (Alder, 2016), and the magnification of misallocation on productivity in economies with production chains (Jones, 2013).

7.2 Intangible Investments and TFP

Neoclassical models with intangible capital are being developed to construct new measures of TFP. These studies focus on intangible investments that traditionally have not been counted as part of national product. Prior to 2013, the Bureau of Economic Analysis (BEA) counted only software as investment among the intangible categories. In 2013, the BEA implemented a comprehensive revision of the National Income and Product Accounts to include other business purchases that previously were counted as business expenses as investment, including research and development, artistic products, mineral exploration, and intellectual property. The shift of these purchases from an expensed item to business investment increases output. This BEA revision improves the measurement of real output, but the BEA does not currently count other intangible investments in the national accounts, such as marketing, advertising, and organization capital investments. These investment omissions indicate that output is mismeasured, which implies that productivity is also mismeasured.

McGrattan and Prescott (2012, 2014), and McGrattan (2016), go beyond the new NIPA measures of GDP by constructing real output measures that include other expensed items, including advertising, marketing, computer design, management consulting, public relations, and engineering expenses as intangible investment. McGrattan (2016) develops a model of the US economy that includes both tangible and intangible production, with a focus on intersectoral linkages.

McGrattan develops a model with tangible output and intangible output. Intangibles are a nonrival good. There are s sectors that use both tangibles and intangibles. There is a Cobb–Douglas aggregate over consumption goods from the S sectors. The technologies differ in terms of a sector-specific technology shock, and technology share parameters. The outputs for tangibles and intangibles is given by:

$$Y_{st}=(K_{Tst}^{1})^{\theta_S}(K_{Ist})^{\phi_S}(\Pi_l(M_{lst}^{1})^{\gamma_{lS}})(Z_tZ_{st}^{1}H_{st}^{1})^{1-\theta_S-\phi_S-\gamma_S} \tag{83}$$

$$I_{st}=(K_{Tst}^{2})^{\theta_S}(K_{Ist})^{\phi_S}(\Pi_l(M_{lst}^{2})^{\gamma_{lS}})(Z_tZ_{st}^{1}H_{st}^{1})^{1-\theta_S-\phi_S-\gamma_S} \tag{84}$$

Y_s denotes the output of the tangible sector, K_{Ts}^{1} is tangible capital that is used to produce tangible output in sector S, K_{Ts}^{2} is tangible capital used to produce intangible output in sector S, K_{Ist} is intangible capital, which is assumed to be nonrival, M_{ls}^{1} and M_{ls}^{2} are intermediate inputs used to produce tangibles in sector S, and intangibles in sector S, respectively. Z is the aggregate productivity shock and Z_s is a sector-specific productivity shock. H_s^1 and H_s^2 are labor input for tangibles in sector S, and intangibles in sector S, respectively.

McGrattan (2016) uses maximum likelihood to estimate the parameters of the stochastic processes for Z_t and for Z_{st}, and compares two economies, one with intangibles, and another without intangibles. The mismeasurement of productivity in the economy without intangibles generates a large labor wedge, and McGrattan argues that this may account for the empirical labor wedge measured from NIPA data. McGrattan also shows that the economy with intangibles closely accounts for the 2008–14 US economy, despite the fact that the standard measure of TFP based on NIPA data is not highly correlated with hours worked during this period.

Another literature that relates intangible investments to productivity is in the area of organization capital. As noted above, these investments are not counted in the NIPA. Atkeson and Kehoe (2005) study a neoclassical model in which an organization stochastically accumulates intangible knowledge over time. They find that the payments from these intangibles are about one-third as large as the payment from tangible capital, which suggests that organization capital is very large.

7.3 Neoclassical Models of Network Linkages and TFP

The impact of industry and/or sectoral shocks on the aggregate economy motivates a significant component of the real business cycle literature, including the seminal contribution of Long Jr and Plosser (1983), and subsequent research by Dupor (1999) and Horvath (2000). One theme of this research is to provide a theory for aggregate productivity shocks that hit the economy.

This idea is now being developed further in network models, which focus on the idea that production is organized through networks of supply chains, and that small disruptions in networks can have significant aggregate consequences, particularly if there are only a small number of suppliers of a particular input, and if there are no particularly close substitutes for that input. Carvalho (2014) describes much of the recent literature on networks and macroeconomics.

Carvalho describes a simple model of production networks in which individual sectors produce a specialized output. This output is produced using homogeneous labor and intermediate inputs from other sectors. The output of sector i is given by:

$$y_i = (z_i h_i)^{1-\theta} \left(\prod_{i=1}^{n} y_{ij}^{\omega_{ij}} \right)^{\theta}. \tag{85}$$

In this technology, y_i denotes sectoral output, z_i is a sectoral productivity shock, h_i is labor employed in sector i, and the exponents ω_{ij} denote the share of intermediate input j used in producing good i. Note that labor is supplied inelastically by a representative household, so aggregate labor is in fixed supply. For simplicity, preferences are symmetric over the i goods in the household utility function.

The empirical importance of network linkages can be identified from a standard input–output matrix. Since aggregate labor is in fixed supply, aggregate output is a weighted average of the sectoral productivity shocks:

$$\ln(y) = \sum_{i=1}^{n} \nu_i \ln(z_i). \tag{86}$$

In this expression, y is aggregate output and the ν_i are weights that are constructed from the input–output table. Note that measured aggregate productivity in this economy, which is $\frac{y}{h}$, will fluctuate even though there is no aggregate productivity shock. This simple model shows how a single shock to an important sector can have significant aggregate affects that will be observationally equivalent to a one-sector model with an aggregate productivity shock.

8. NEOCLASSICAL MODELS OF INEQUALITY

Neoclassical modeling is also making considerable progress in characterizing and quantifying how technological change has affected income distribution and wage inequality. Neoclassical studies of inequality analyze how biased technological change differentially affects the demand for different types of workers.

Early empirical studies by Katz and Murphy (1992), among others, concluded that skill-biased technological change was responsible for the widening wage gap between highly-educated workers and workers with less education. This conclusion reflects the fact that the relative supply of highly-skilled workers rose considerably, and the relative wage of these workers also rose.

Krusell et al. (2000) develop a neoclassical model to analyze how technological change has affected the relative wage of skilled to less-skilled workers. This relative wage is often called the *skill premium*. Krusell et al. provide an explicit theory of skill-biased technological change, show how to measure this change, and develop a neoclassical model to quantify its effect on inequality through observable variables.

The model features two different types of labor: high-skilled labor, who are workers with 16 or more years of education, and unskilled labor, who have fewer than 16 years of education.[t] Skill-biased technological change in this model is the combination of capital equipment-specific technological change, coupled with different substitution elasticities between the two types of labor. Krusell et al. construct a four factor production function

[t] Note that the term *unskilled* is used here not as a literal description of worker skill, but rather to clearly differentiate the two types of labor from each other.

that allows for different types of labor, and for different types of capital goods. The technology is given by:

$$y_t = A_t k_{st}^{\alpha}[\mu u_t^{\sigma} + (1-\mu)(\lambda k_{et}^{\rho} + (1-\lambda)s_t^{\rho})^{\frac{\sigma}{\rho}}]^{\frac{1-\alpha}{\sigma}} \tag{87}$$

The term A_t is a neutral technology parameter. The inputs are capital structures (k_{st}), unskilled labor input (u_t), which is the product of unskilled hours and unskilled labor efficiency ($\psi_{ut}h_{ut}$), capital equipment (k_{et}), and skilled labor input (s_t), which is the product of skilled labor hours and skilled labor efficiency ($\psi_{st}h_{st}$). These inputs are specified within a nested CES technology in which the curvature parameters σ and ρ govern the substitution elasticities among the inputs. In this technology, rapid growth of capital equipment raises the wage of skilled workers relative to the wage of unskilled workers only if capital equipment is more complementary with skilled labor than with unskilled labor. This requires that $\sigma > \rho$, which Krusell et al. call *capital-skill complementarity*.

It is straightforward to see this requirement of $\sigma > \rho$ by assuming that ψ_{st} and ψ_{ut} are constant, log-linearizing the ratio of the marginal productivities of the two types of labor, and expressing variables in terms of growth rates between periods t and $t + 1$:

$$g_{\pi t} \simeq (1-\sigma)(g_{h_{ut}} - g_{h_{st}}) + (\sigma-\rho)\lambda\left(\frac{k_{et}}{s_t}\right)^{\rho}(g_{k_{et}} - g_{h_{st}}) \tag{88}$$

In (90), g_π is the growth rate of the skill premium, g_{h_u} and g_{h_s} are the growth rates of unskilled and skilled hours, and g_{k_e} is the growth rate of capital equipment. Since the parameter σ is less than one, the first term on the right hand side of (90) shows that the skill premium declines if the growth rate of skilled hours exceeds the growth rate of unskilled hours. Krusell et al. call this first term the *relative quantity effect*. The second term is called the *capital-skill complementarity effect*. This second term shows that the skill premium rises if the growth rate of capital equipment exceeds the growth rate of skilled hours, and if there is relatively more complementarity between skilled labor and equipment ($\sigma > \rho$).

Krusell et al. construct a dataset of skilled and unskilled labor input using data from the Current Population Survey. They use Gordon's (1990) data on equipment prices to construct a measure of the stock of capital equipment, and they use the NIPA measure of capital structures.

They estimate the parameters of the nonlinear production function with data from 1963 to 1992 using two-step simulated pseudo-maximum likelihood. They fit the model using the equations that measure the deviation between model and data for total labor's share of income, and the ratio of skilled labor income to unskilled labor income. The third equation in the criterion function measures the deviation between the rate of return to investment in structures to equipment. They estimate substitution elasticities of about 1.67 between unskilled labor and equipment, and of about 0.67 between skilled labor and

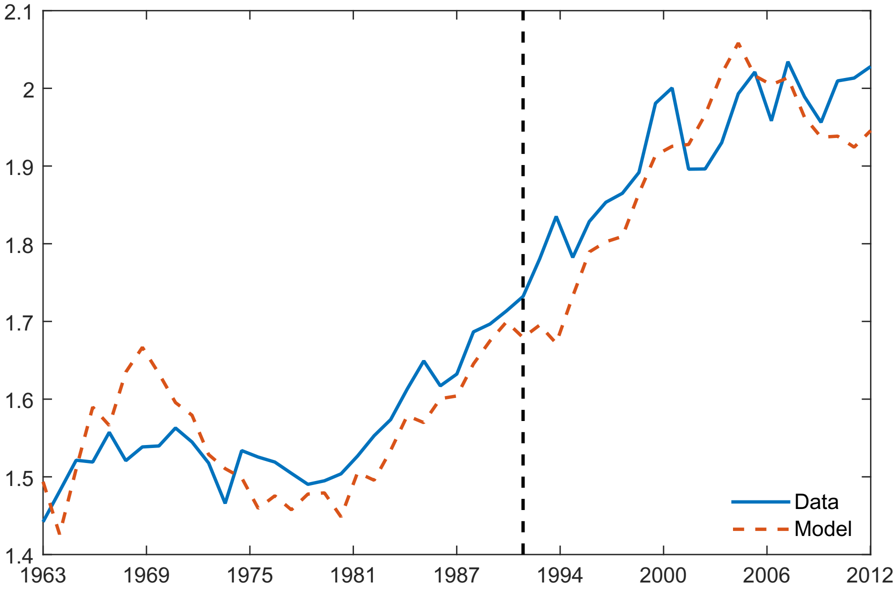


Fig. 30 Comparing college skill premium, model and data.

equipment, which provides strong support for capital-skill complementarity. They find that the model accounts for much of the movements in the skill premium over the 1963–92 period.

Given that the Krusell et al. data end in 1992, Ohanian and Orak (2016) analyze this same model, but extend the dataset through 2013 to assess the contribution of capital-skill complementarity to wage inequality for the last 20 years. Fig. 30 shows the skill premium in the model and in the data from 1963 to 2013. To compare the analysis to Krusell et al., Ohanian and Orak also estimate the model from 1963 to 1992. The dashed line in Fig. 30 corresponds to the end of the estimation period for the parameters (1992). Although Ohanian and Orak use the same sample period to estimate the parameters, they use revised data in the estimation. They find very similar elasticities to those in Krusell et al. Ohanian and Orak estimate an elasticity of about 1.78 between unskilled labor and equipment, and about 0.69 between skilled labor and equipment. The figure shows that the model accounts for the major changes in the skill premium, including the very large rise that has occurred in the last 30 years.[u]

The Krusell et al. model also fits aggregate labor share very well up until the mid-2000s. After that, the model overpredicts labor's share. This finding led Orak (2016) to analyze the same type of production function with different substitution possibilities

[u] Krusell et al. normalize the skill premium to 1 in 1963, and report fluctuations relative to the normalized value. To show the actual level of the skill premium, Ohanian and Orak estimate the model with normalized data as in Krusell et al. and then reconstruct the levels data. See Ohanian and Orak for details.

between capital equipment and different types of skills, but with three types of labor, as opposed to two types of labor. The labor types in Orak are classified based on occupational tasks, as in Autor et al. (2003), rather than on education levels, as in Krusell et al.

Orak specifies the three types of labor based on whether an occupation primarily performs cognitive tasks, manual tasks, or routine tasks. He estimates a relatively high elasticity of substitution between capital equipment and workers who perform routine tasks, and he estimates lower substitution elasticities between equipment and cognitive workers, and between equipment and manual workers. He finds that this augmented neoclassical model can account for much of the recent and significant decline in labor's share of income.

9. NEOCLASSICAL MACROECONOMICS: CRITICAL ASSESSMENTS AND FUTURE DIRECTIONS

This section discusses the open questions in the area of neoclassical macroeconomics, and presents our views on interesting future avenues for research that will address these questions. Perhaps the major open question for neoclassical models—and which is also a major question for other classes of macroeconomic models—is accounting for fluctuations in hours worked. The multisector models developed in this chapter account for considerably more of the fluctuations in hours worked than the standard one-sector neoclassical model, but there are also changes in hours that these models do not capture. Below, we describe the research areas that we view as important and promising in addressing this issue and others.

9.1 Biased Technological Change and the Labor Market

Analysis of biased technological change, and its impact on both aggregate variables and on labor market outcomes of workers with different skill levels, is an interesting avenue for future research. The home production results from the model motivated by Greenwood et al. (2005) indicate interesting trend changes in hours worked from the early 1980s through the 1990s, which coincide with the increase in women's hours worked. Important future research will further connect this demographic increase in hours worked with general equilibrium models of home production.

More broadly, it will be important to further develop models in the area of directed technological change and the shape of the production function, as in Acemoglu (2002) and Jones (2005), the relationship between technologies and secular sectoral shifts, as in Lee and Wolpin (2006), human capital accumulation and technological change, as in Heckman et al. (1998), and demographic shifts, technological change, and wage shifts as in Jeong et al. (2015). A related area is studying movements in factor income shares, as in Karabarbounis and Neiman (2014) and Orak (2016), and the impact of factor endowments on how societies choose among biased technologies, as in Caselli and Coleman (2006).

All of these research areas are in relatively early stages of development, and merit additional analysis. Research in this area can also be combined with broader empirical studies of time allocation, including the analysis and documentation of home and market time allocation, as in Aguiar and Hurst (1997) and Aguiar et al. (2013), and studies of the allocation of time across rich and poor countries, as in Bick et al. (2016).

9.2 Neoclassical Analyses of the Great Recession and Its Aftermath

Several open questions remain about the Great Recession and its aftermath. This includes accounting for macroeconomic aggregates from 2008 and onwards, particularly for hours worked. The results presented in this chapter indicate that neoclassical models with standard measures of equipment-specific productivity shocks, and TFP shocks, and without any policy components, miss some features of the Great Recession. McGrattan (2016) argues that output mismeasurement resulting from the omission of intangible investments in GDP has important implications for measured TFP and labor wedge measures during the Great Recession. Further research in this important area is needed.

There are also interesting aspects of economic policies during this period that merit additional analysis. Mulligan (2012, 2013) argues that changes in social insurance programs and the Affordable Care Act depressed labor by implicitly raising tax rates on labor. Kydland and Zarazaga (2016) study how expectations of different types of tax policies may have contributed to the weak recovery from the Great Recession. Baker et al. (2015) measure the evolution of economic policy uncertainty during the Great Recession. These uncertainty measures can be used in models in which uncertainty can depress an economy, as in Bloom (2009) and Fernández-Villaverde et al. (2015). These factors may have implications for understanding changes in hours worked in recent years.

9.3 The Effect of Policy Changes and Institutions on Macroeconomic Performance

An important area for future research is quantifying the impact of observed departures from competitive markets on economies. Cole and Ohanian (2004) developed and applied a particular methodology in their study of cartelization and unionization in the US Great Depression. This approach was also applied by Lahiri and Yi (2009) in evaluating the affect of noncompetitive policies in West Bengal Indian development. A similar approach has been used by Cheremukhin et al. (2013, 2015) to study the impact of Lenin's policies and institutions on economic development in the USSR at that time, and to study the impact of Mao's policies and institutions on Chinese development in the 1940s and 1950s. Alder (2016) uses a related approach to analyze the contribution of labor union hold-up and imperfect competition on the decline of America's Rust Belt region in the postwar United States. Similar methods also can be used to study the recent evolution of the post-Soviet Union economies, to study recent Indian and Chinese development patterns (see Dekle and Vandenbroucke (2012) for a neoclassical study of recent trends in China's economy), and to study long-run Latin American development (see Cole et al., 2005 for a long-run analysis of Latin America). As better data becomes available, these methods can also be used to study how policies and institutions have affected the stagnation and development of very poor countries. Future research along these lines will allow us to understanding the relative importance of various noncompetitive policies across countries, and will be an important input in developing growth-enhancing policies in poor countries.

9.4 Analyses of TFP

Since productivity is central in neoclassical growth models, advancing our understanding of changes in TFP is another important area for future research. In the last 10 years, progress in evaluating TFP has been made along three different research lines: resource misallocation, intangible investments, and network economies. Advancements in misallocation analysis of TFP will be facilitated by the assessment of how actual economic policies have affected resource allocation and productivity loss. Continued advances in computing power will facilitate the analysis of network economies and intersectoral linkages in the study of TFP. The continued expansion of intangible investments into NIPA data will advance our understanding of intangibles investment and TFP.

An area that to our knowledge has not been studied in detail is to link changes in what Decker et al. (2014) call "business dynamism" to aggregate measures of TFP. Specifically, Decker et al. document lower rates of resource reallocation in the United States, and also a lower rate of successful start-ups that have occurred over time. This decline has coincided with a secular decline in productivity growth. Analyzing theoretical and empirical connections between these observations has the potential to advance our understanding of secular movements in productivity.

9.5 Taxes and Macroeconomic Activity

The impact of tax and fiscal policies on economic activity in neoclassical models is another interesting area for future work, and may advance our understanding of changes in hours worked. Research in this area has been constrained by the availability of data on tax rates and hours worked. Constructing tax rates along the lines of McDaniel's (2011) tax measurements for the OECD can in principle be extended to other countries. In terms of hours worked, Ohanian and Raffo (2011) construct panel data on hours in the OECD, and similar data constructions can be made for other countries.

10. CONCLUSIONS

This chapter presented aggregate data and a series of neoclassical models to show how the historical evolution of the US economy reflects much longer-run changes in economic activity than previously recognized, and that much of this evolution is plausibly interpreted as the consequences of long-run shifts in technologies and government policies.

This chapter shows that neoclassical models can shed light on relatively stable periods of aggregate economic activity, such as the post-Korean War US economy, but also on very turbulent periods that are typically considered to be far beyond the purview of neoclassical economics, including the Great Depression and World War II. Moreover, neoclassical analysis not only provides insights into purely aggregate issues, but also sheds light on how technological change has affected individual labor market outcomes.

Future macroeconomic analyses of fluctuations should shift from the standard practice of narrowly studying business cycle frequencies, and to include the quantitatively important lower frequency component of fluctuations that dominates much of the US historical economic record. We anticipate that neoclassical research along these lines will continue to advance the profession's knowledge in a number of areas reflecting both longer-run events and business cycle fluctuations. This includes Depressions, Growth Miracles, the macroeconomic effects of various types of government regulatory and fiscal policies, the sources and nature of productivity shocks, the effects of biased technological change on the macroeconomy and on individual labor market outcomes, and understanding cyclical and longer-run fluctuations in hours worked.

ACKNOWLEDGMENTS

We thank John Cochrane, Jesus Fernández-Villaverde, Kyle Herkenhoff, Per Krusell, Ed Prescott, Valerie Ramey, John Taylor, Harald Uhlig, seminar participants at the Handbook of Macroeconomics Conference, and at the 2015 Federal Reserve Bank of Saint Louis Policy Conference for comments. Adrien D'Avernas Des Enffans, Eric Bai, Andreas Gulyas, Jinwook Hur, and Musa Orak provided excellent research assistance.

REFERENCES

Acemoglu, D., 2002. Directed technical change. Rev. Econ. Stud. 69 (4), 781–809.

Aguiar, M., Hurst, E., 1997. Life-cycle prices and production. Am. Econ. Rev. 5 (3), 1533–1599.

Aguiar, M., Hurst, E., Karabarbounis, L., 2013. Time use during the Great Recession. Am. Econ. Rev. 103 (5), 1664–1696.

Ahearne, A., Kydland, F., Wynne, M.A., 2006. Ireland's Great Depression. Econ. Soc. Rev. 37 (2), 215–243.

Alder, S., 2016. In the wrong hands: complementarities, resource allocation, and TFP. Am. Econ. J.: Macroecon. 8 (1), 199–241.

Alesina, A.F., Glaeser, E.L., Sacerdote, B., 2006. Work and leisure in the US and Europe: why so different?. In: Gertler, M., Rogoff, K. (Eds.), NBER Macroeconomics Annual 2005, vol. 20. MIT Press, Cambridge, MA, pp. 1–100.

Amaral, P.S., MacGee, J.C., 2015. Re-examining the role of sticky wages in the U.S. Great Depression: a multi-sector approach, University of Western Ontario.

Arias, A., Hansen, G., Ohanian, L.E., 2007. Why have business cycle fluctuations become less volatile? Econ. Theory 32 (1), 43–58.

Atkeson, A., Kehoe, P.J., 2005. Modeling and measuring organization capital. J. Polit. Econ. 113 (5), 1026–1053.

Atkeson, A., Khan, A., Ohanian, L.E., 1996. Are data on industry evolution and gross job turnover relevant for macroeconomics? Carn. Roch. Conf. Ser. Public Policy 44 (2), 215–250.

Autor, D.H., Levy, F., Murnane, R.J., 2003. The skill content of recent technological change: an empirical exploration. Q. J. Econ. 118 (4), 1279–1333.

Baker, S.R., Bloom, N., Davis, S.J., 2015. Measuring economic policy uncertainty. National Bureau of Economic Research. Working Paper No. 21633.

Barro, R.J., 1981. Intertemporal substitution and the business cycle. Carn. Roch. Conf. Ser. Public Policy 14 (1), 237–268.

Barro, R.J., King, R.G., 1984. Time-separable preferences and intertemporal-substitution models of business cycles. Q. J. Econ. 99 (4), 817–839.

Barro, R.J., Redlick, C.J., 2011. Macroeconomic effects from government purchases and taxes. Q. J. Econ. 126 (1), 51–102.

Braun, R.A., McGrattan, E.R., 1993. The macroeconomics of war and peace. In: Blanchard, O., Fischer, S. (Eds.), NBER Macroeconomics Annual 1993. In: MIT Press 8, Cambridge, MA, pp. 197–258.

Baxter, M., King, R.G., 1993. Fiscal policy in general equilibrium. Am. Econ. Rev. 83 (3), 315–334.

Baxter, M., King, R.G., 1999. Measuring business cycles: approximate band-pass filters for economic time series. Rev. Econ. Stat. 81 (4), 575–593.

Benhabib, J., Rogerson, R., Wright, R., 1991. Homework in macroeconomics: household production and aggregate fluctuations. J. Polit. Econ. 99 (6), 1166–1187.

Bernanke, B.S., 1983. Nonmonetary effects of the financial crisis in the propagation of the Great Depression. Am. Econ. Rev. 73 (3), 257–276.

Beveridge, S., Nelson, C.R., 1981. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the ‘business cycle’. J. Monet. Econ. 7 (2), 151–174.

Bick, A., Lagakos, D., Fuchs-Schundeln, N., 2016. How do average hours worked vary with development: cross-country evidence and implications. Unpublished paper.

Blanchard, O., 2004. The economic future of Europe. J. Econ. Perspect. 18 (4), 3–26.

Bloom, N., 2009. The impact of uncertainty shocks. Econometrica 77 (3), 623–685.

Bond, E.W., Crucini, M.J., Potter, T., Rodrigue, J., 2013. Misallocation and productivity effects of the Smoot-Hawley tariff. Rev. Econ. Dyn. 16 (1), 120–134.

Bordo, M.D., Erceg, C.J., Evans, C.L., 2000. Money, sticky wages, and the Great Depression. Am. Econ. Rev. 90 (5), 1447–1463.

Brinca, P., Chari, V.V., Kehoe, P.J., McGrattan, E.R., 2016. Accounting for business cycles. In: Taylor, J., Uhlig, H. (Eds.), Handbook of Macroeconomics, vol. 2A. Elsevier, Amsterdam, Netherlands, pp. 1013–1063.

Buera, F.J., Moll, B., 2015. Aggregate implications of a credit crunch: the importance of heterogeneity. Am. Econ. J. Macroecon. 7 (3), 1–42.

Burns, A.F., Mitchell, W.C., 1946. Measuring Business Cycles. National Bureau of Economic Research, New York.

Burnside, C., Eichenbaum, M., Fisher, J.D.M., 2004. Fiscal shocks and their consequences. J. Econ. Theory 115 (1), 89–117.

Carvalho, V.M., 2014. From micro to macro via production networks. J. Econ. Perspect. 28 (4), 23–47.

Caselli, F., Coleman II, J.W., 2006. The world technology frontier. Am. Econ. Rev. 96 (3), 499–522.

Chang, Y., Schorfheide, F., 2003. Labor-supply shifts and economic fluctuations. J. Monet. Econ. 50 (8), 1751–1768.

Chari, V.V., Kehoe, P.J., McGrattan, E.R., 2007. Business cycle accounting. Econometrica 75 (3), 781–836.

Chen, K., Irarrazabal, A., 2013. Misallocation and the recovery of manufacturing TFP after a financial crisis. Norges Bank. Working Paper 2013-01.

Cheremukhin, A., Golosov, M., Guriev, S., Tsyvinski, A., 2013. Was Stalin necessary for Russia's economic development? National Bureau of Economic Research. Working Paper No. 19425.

Cheremukhin, A., Golosov, M., Guriev, S., Tsyvinski, A., 2015. The economy of People's Republic of China from 1953. National Bureau of Economic Research. Working Paper No. 21397.

Christiano, L.J., Eichenbaum, M., 1992. Current real-business-cycle theories and aggregate labor-market fluctuations. Am. Econ. Rev. 82 (3), 430–450.

Christiano, L.J., Fitzgerald, T.J., 2003. The band pass filter. Int. Econ. Rev. 44 (2), 435–465.

Cociuba, S., Prescott, E., Ueberfeldt, A., 2012. U.S. hours and productivity behavior using CPS hours worked data: 1947-III to 2011-IV. Discussion paper.

Cole, H.L., Ohanian, L.E., 1999. The Great Depression in the United States from a neoclassical perspective. Fed. Reserve Bank Minneapolis Q. Rev. 23 (1), 2–24.

Cole, H.L., Ohanian, L.E., 2002. The Great UK Depression: a puzzle and possible resolution. Rev. Econ. Dyn. 5 (1), 19–44.

Cole, H.L., Ohanian, L.E., 2004. New Deal policies and the persistence of the Great Depression. J. Polit. Econ. 112 (4), 779–816.

Cole, H.L., Ohanian, L.E., 2016. The impact of cartelization, money, and productivity shocks on the international Great Depression. Unpublished.

Cole, H.L., Ohanian, L.E., Riascos, A., Schmitz, J.A., 2005. Latin America in the rearview mirror. J. Monet. Econ. 52 (1), 69–107.

Comin, D., Gertler, M., 2006. Medium-term business cycles. Am. Econ. Rev. 96 (3), 523–551.

Conesa, J.C., Kehoe, T.J., Ruhl, K.J., 2007. Modeling great depressions: the depression in Finland in the 1990s. In: Kehoe, T.J., Prescott, E.C. (Eds.), Great Depressions of the 20th Century. Federal Reserve Bank of Minneapolis, Minneapolis, MN.

Cooley, T.F. (Ed.), 1995. Frontiers of Business Cycle Research. Princeton University Press, Princeton, NJ.

Cummins, J.G., Violante, G.L., 2002. Investment-specific technical change in the United States (1947-2000): measurement and macroeconomic consequences. Rev. Econ. Dyn. 5 (2), 243–284.

Dalton, J.T., 2015. The evolution of taxes and hours worked in Austria, 1970-2005. Macroecon. Dyn. 19 (8), 1800–1815.

David, J.M., Hopenhayn, H.A., Venkateswaran, V., forthcoming. Information, misallocation and aggregate productivity. Q. J. Econ.

Davis, S.J., Haltiwanger, J., 1992. Gross job creation, gross job destruction, and employment reallocation. Q. J. Econ. 107 (3), 819–863.

Davis, S.J., Henrekson, M., et al., 2005. Tax effects on work activity, industry mix, and shadow economy size: evidence from rich country comparisons. In: Gómez-Salvador, R. (Ed.), Labour Supply and Incentives to Work in Europe. Edward Elgar, Cheltenham, UK.

Decker, R., Haltiwanger, J., Jarmin, R., Miranda, J., 2014. The role of entrepreneurship in US job creation and economic dynamism. J. Econ. Perspect. 28 (3), 3–24.

Dekle, R., Vandenbroucke, G., 2012. A quantitative analysis of China's structural transformation. J. Econ. Dyn. Control 36 (1), 119–135.

DiCecio, R., 2009. Sticky wages and sectoral labor comovement. J. Econ. Dyn. Control. 33 (3), 538–553.

Doepke, M., Hazan, M., Maoz, Y.D., 2015. The baby boom and World War II: a macroeconomic analysis. Rev. Econ. Stud. 82 (3), 1031–1073.

Dupor, B., 1999. Aggregation and irrelevance in multi-sector models. J. Monet. Econ. 43 (2), 391–409.

Ebell, M., Ritschl, A., 2008. Real origins of the Great Depression: monopoly power, unions and the American business cycle in the 1920s. CEP Discussion Paper No. 876.

Eggertsson, G.B., 2012. Was the New Deal contractionary? Am. Econ. Rev. 102 (1), 524–555.

Erosa, A., Fuster, L., Kambourov, G., 2012. Labor supply and government programs: a cross-country analysis. J. Monet. Econ. 59 (1), 84–107.

Feenstra, R.C., Inklaar, R., Timmer, M.P., 2015. The next generation of the Penn World Table. Am. Econ. Rev. 105 (10), 3150–3182.

Fernald, J., 2014. A quarterly, utilization-adjusted series on total factor productivity. Federal Reserve Bank of San Francisco. Working Paper 2012-19.

Fernández-Villaverde, J., Guerrón-Quintana, P., Kuester, K., Rubio-Ramírez, J., 2015. Fiscal volatility shocks and economic activity. Am. Econ. Rev. 105 (11), 3352–3384.

Field, A.J., 2003. The most technologically progressive decade of the century. Am. Econ. Rev. 93 (4), 1399–1413.

Fisher, J.D.M., 2006. The dynamic effects of neutral and investment-specific technology shocks. J. Polit. Econ. 114 (3), 413–451.

Friedman, M., Schwartz, A.J., 1963. Monetary History of the United States, 1867-1960. Princeton University Press, Princeton, NJ.

Gali, J., van Rens, T., 2014. The vanishing procyclicality of labor productivity. Unpublished paper.

Gordon, R.J., 1990. The Measurement of Durable Goods Prices. University of Chicago Press, Chicago, IL.

Gorodnichenko, Y., Mendoza, E.G., Tesar, L.L., 2012. The Finnish Great Depression: from Russia with love. Am. Econ. Rev. 102 (4), 1619–1643.

Greenwood, J., Hercowitz, Z., Krusell, P., 1997. Long-run implications of investment-specific technological change. Am. Econ. Rev. 87 (3), 342–362.

Greenwood, J., Seshadri, A., Yorukoglu, M., 2005. Engines of liberation. Rev. Econ. Stud. 72 (1), 109–133.

Greenwood, J., Yorukoglu, M., 1997. Carn. Roch. Conf. Ser. Public Policy 46 (1), 49–95.

Guner, N., Ventura, G., Yi, X., 2008. Macroeconomic implications of size-dependent policies. Rev. Econ. Dyn. 11 (4), 721–744.

Hansen, G.D., 1985. Indivisible labor and the business cycle. J. Monet. Econ. 16 (3), 309–327.

Hansen, G.D., 1997. Technical progress and aggregate fluctuations. J. Econ. Dyn. Control 21 (6), 1005–1023.

Hansen, G.D., Prescott, E.C., 1993. Did technology shocks cause the 1990-1991 recession? Am. Econ. Rev. 83 (2), 280–286.

Hayashi, F., Prescott, E.C., 2002. The 1990s in Japan: a lost decade. Rev. Econ. Dyn. 5 (1), 206–235.

Heckman, J.J., Lochner, L., Taber, C., 1998. Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. Rev. Econ. Dyn. 1 (1), 1–58.

Herman, A., 2012. Freedom's Forge: How American Business Produced Victory in World War II. Random House, New York City, NY.

Hodrick, R., Prescott, E.C., 1997. Postwar U.S. business cycles: an empirical investigation. J. Money Credit Bank. 29 (1), 1–16.

Hopenhayn, H., Rogerson, R., 1993. Job turnover and policy evaluation: a general equilibrium analysis. J. Polit. Econ. 101 (5), 915–938.

Horvath, M., 2000. Sectoral shocks and aggregate fluctuations. J. Monet. Econ. 45 (1), 69–106.

Hsieh, C.T., Klenow, P.J., 2009. Misallocation and manufacturing TFP in China and India. Q. J. Econ. 124 (4), 1403–1448.

Jeong, H., Kim, Y., Manovskii, I., 2015. The price of experience. Am. Econ. Rev. 105 (2), 784–815.

Joines, D.H., 1981. Estimates of effective marginal tax rates on factor incomes. J. Bus. 54 (2), 191–226.

Jones, C.I., 2005. The shape of production functions and the direction of technical change. Q. J. Econ. 120 (2), 517–549.

Jones, C.I., 2013. Misallocation, economic growth, and input-output economics. In: Acemoglu, D., Arellano, M., Dekel, E. (Eds.), Advances in Economics and Econometrics, Tenth World Congress. vol. II. Cambridge University Press, Cambridge.

Justiniano, A., Primiceri, G.E., Tambalotti, A., 2010. Investment shocks and business cycles. J. Monet. Econ. 57 (2), 132–145.

Karabarbounis, L., Neiman, B., 2014. The global decline of the labor share. Q. J. Econ. 129 (1), 61–103.

Katz, L.F., Murphy, K.M., 1992. Changes in relative wages, 1963-1987: supply and demand factors. Q. J. Econ. 107 (1), 35–78.

Kehoe, T.J., Prescott, E.C., 2007. Great Depressions of the Twentieth Century. Federal Reserve Bank of Minneapolis, Minneapolis, MN.

Kendrick, J., 1961. Productivity Trends in the United States. Princeton University Press, Princeton, NJ.

Klein, P., Ventura, G., 2015. Making a miracle: Ireland 1980-2005. Unpublished paper.

Kovacic, W.E., Shapiro, C., 2000. Antitrust policy: a century of economic and legal thinking. J. Econ. Perspect. 14 (1), 43–60.

Krusell, P., Ohanian, L.E., Ríos-Rull, J.V., Violante, G.L., 2000. Capital-skill complementarity and inequality: a macroeconomic analysis. Econometrica 68 (5), 1029–1053.

Kydland, F.E., Prescott, E.C., 1982. Time to build and aggregate fluctuations. Econometrica 50 (6), 1345–1370.

Kydland, F.E., Prescott, E.C., 1988. The workweek of capital and its cyclical implications. J. Monet. Econ. 21 (2), 343–360.

Kydland, F.E., Zarazaga, C.E.J.M., 2016. Fiscal sentiment and the weak recovery from the Great Recession: a quantitative exploration. J. Monet. Econ. 79, 109–125.

Lahiri, A., Yi, K.-M., 2009. A tale of two states: Maharashtra and West Bengal. Rev. Econ. Dyn. 12 (3), 523–542.

Lee, D., Wolpin, K.I., 2006. Intersectoral labor mobility and the growth of the service sector. Econometrica 74 (1), 1–46.

Lilien, D.M., 1982. Sectoral shifts and cyclical unemployment. J. Polit. Econ. 90 (4), 777–793.

Ljungqvist, L., Sargent, T.J., 1998. The European unemployment dilemma. J. Polit. Econ. 106 (3), 514–550.

Long Jr., J.B., Plosser, C.I., 1983. Real business cycles. J. Polit. Econ. 91 (1), 39–69.

Lu, S.S., 2012. East Asian growth experience revisited from the perspective of a neoclassical model. Rev. Econ. Dyn. 15 (3), 359–376.

Lucas, R.E., 1973. Expectations and the neutrality of money. J. Econ. Theory 4 (2), 103–124.

Lucas, R.E., Rapping, L.A., 1969. Real wages, employment, and inflation. J. Polit. Econ. 77 (5), 721–754.

Lucas, R.E., Rapping, L.A., 1972. Unemployment in the Great Depression: is there a full explanation? J. Polit. Econ. 80 (1), 186–191.

Manuelli, R.E., Seshadri, A., 2014. Frictionless Technology Diffusion: The Case of Tractors. Am. Econ. Rev. 104 (4), 1368–1391.

McDaniel, C., 2011. Forces shaping hours worked in the OECD, 1960-2004. Am. Econ. J.: Macroecon. 3 (4), 27–52.

McGrattan, E.R., 2012. Capital taxation during the US Great Depression. Q. J. Econ. 127 (3), 1515–1550.

McGrattan, E.R., 2016. Intangible Capital and Measured Productivity. Working Paper, University of Minnesota, Minneapolis, MN.

McGrattan, E.R., Prescott, E.C., 2014. A reassessment of real business cycle theory. Am. Econ. Rev. Papers and Proceedings 104 (5), 177–187.

McGrattan, E.R., Ohanian, L.E., 2010. Does neoclassical theory account for the effects of big fiscal shocks? Evidence from World War II. Int. Econ. Rev. 51 (2), 509–532.

McGrattan, E.R., Prescott, E.C., 2012. The Great Recession and delayed economic recovery: a labor productivity puzzle? In: Ohanian, L.E., Taylor, J.B., Wright, I. (Eds.), Government Policies and the Delayed Economic Recovery. Hoover Press, Stanford, CA.

Meza, F., 2008. Financial crisis, fiscal policy, and the 1995 GDP contraction in Mexico. J. Money Credit Bank. 40 (6), 1239–1261.

Midrigan, V., Xu, D.Y., 2014. Finance and misallocation: evidence from plant-level data. Am. Econ. Rev. 104 (2), 422–458.

Modigliani, F., 1977. The monetarist controversy or, should we forsake stabilization policies? Am. Econ. Rev. 67 (2), 1–19.

Moll, B., 2014. Productivity losses from financial frictions: can self-financing undo capital misallocation? Am. Econ. Rev. 104 (10), 3186–3221.

Monacelli, T., Perotti, R., 2008. Fiscal policy, wealth effects, and markups. National Bureau of Economic Research. Working Paper No. 14584.

Morley, J., Nelson, C.R., Zivot, E., 2003. Why are unobserved component and Beveridge-Nelson trend-cycle decompositions of GDP so different? Rev. Econ. Stat. 85 (2), 235–243.

Mountford, A., Uhlig, H., 2009. What are the effects of fiscal policy shocks? J. Appl. Econ. 24 (6), 960–992.

Mulligan, C., 2012. The Redistribution Recession: How Labor Market Distortions Contracted the Economy. Oxford University Press, Oxford.

Mulligan, C., 2013. Average marginal labor income tax rates under the Affordable Care Act. National Bureau of Economic Research. Working Paper No. 19365.

Mulligan, C.B., 2005. Public policies as specification errors. Rev. Econ. Dyn. 8 (4), 902–926.

Oberfield, E., 2013. Productivity and misallocation during a crisis: evidence from the Chilean crisis of 1982. Rev. Econ. Dyn. 16 (1), 100–119.

Ohanian, L.E., 1997. The macroeconomic effects of war finance in the United States: World War II and the Korean War. Am. Econ. Rev. 87 (1), 23–40.

Ohanian, L.E., 2001. Why did productivity fall so much during the Great Depression? Am. Econ. Rev. 91 (2), 34–38.

Ohanian, L.E., 2009. What - or - who started the Great Depression? J. Econ. Theory 144 (6), 2310–2335.

Ohanian, L.E., 2010. The economic crisis from a neoclassical perspective. J. Econ. Perspect. 24 (4), 45–66.

Ohanian, L.E., 2011. Comment on 'what fiscal policy is effective at zero interest rates'? In: Acemoglu, D., Woodford, M. (Eds.), NBER Macroeconomics Annual 2010, vol. 25. University of Chicago Press, Chicago, IL, pp. 125–137.

Ohanian, L.E., Orak, M., 2016. Capital-skill complementarity, inequality, and labor's share of income, 1963-2013. Discussion paper.

Ohanian, L.E., Raffo, A., 2011. Aggregate hours worked in OECD countries: new measurement and implications for business cycles. National Bureau of Economic Research. Working Paper 17420.

Ohanian, L., Raffo, A., Rogerson, R., 2008. Long-term changes in labor supply and taxes: evidence from OECD countries, 1956–2004. J. Monet. Econ. 55 (8), 1353–1362.

Orak, M., 2016. Capital-Task Complementarity and the Decline of the US Labor Share of Income. Working Paper, UCLA, Los Angeles, CA.

Osuna, V., Rios-Rull, J.-V., 2003. Implementing the 35 hour workweek by means of overtime taxation. Rev. Econ. Dyn. 6 (1), 179–206.

Otsu, K., 2008. A neoclassical analysis of the Korean crisis. Rev. Econ. Dyn. 11 (2), 449–471.

Poschke, M., 2010. The regulation of entry and aggregate productivity. Econ. J. 120 (549), 1175–1200.

Prescott, E.C., 2004. Why do Americans work so much more than Europeans? Federal Reserve Bank of Minneapolis. Q. Rev.Vol. 28, 2–13. No. 1, July 2004.

Ragan, K.S., 2013. Taxes and time use: fiscal policy in a household production model. Am. Econ. J.: Macroecon. 5 (1), 168–192.

Ramey, V.A., 2011. Can government purchases stimulate the economy? J. Econ. Lit. 49 (3), 673–685.

Rees, A., 1970. On equilibrium in labor markets. J. Polit. Econ. 78 (2), 306–310.

Restuccia, D., Rogerson, R., 2008. Policy distortions and aggregate productivity with heterogeneous establishments. Rev. Econ. Dyn. 11 (4), 707–720.

Rogerson, R., 2008. Structural transformation and the deterioration of European labor market outcomes. J. Polit. Econ. 116 (2), 235–259.

Rogerson, R., 2009. Market work, home work, and taxes: a cross-country analysis. Rev. Int. Econ. 17 (3), 588–601.

Samaniego, R.M., 2008. Can technical change exacerbate the effects of labor market sclerosis? J. Econ. Dyn. Control. 32 (2), 497–528.

Sandleris, G., Wright, M.L.J., 2014. The costs of financial crises: resource misallocation, productivity, and welfare in the 2001 Argentine crisis. Scand. J. Econ. 116 (1), 87–127.

Sargent, T.J., 1973. Rational expectations, the real rate of interest, and the natural rate of unemployment. Brook. Pap. Econ. Act. 2, 429–480.

Sargent, T.J., Wallace, N., 1975. 'Rational' expectations, the optimal monetary instrument, and the optimal money supply rule. J. Polit. Econ. 83 (2), 241–254.

Schumpeter, J., 1927. The explanation of the business cycle. Economica 21, 286–311.

Simon, C.J., 2001. The supply price of labor during the Great Depression. J. Econ. Hist. 61 (4), 877–903.

Siu, H.E., 2008. The fiscal role of conscription in the US World War II effort. J. Monet. Econ. 55 (6), 1094–1112.

Smets, F., Wouters, R., 2007. Shocks and frictions in US business cycles: a Bayesian DSGE approach. Am. Econ. Rev. 97 (3), 586–606.

Stock, J.H., Watson, M.W., 1988. Variable trends in economic time series. J. Econ. Perspect. 2 (3), 147–174.

Stock, J.H., Watson, M.W., 2003. Has the business cycle changed and why? NBER Macroeconomics Annual 2002, vol. 17. MIT Press, Cambridge, MA, pp. 159–230.

Taylor, J.B., 2010. Getting back on track: macroeconomic policy lessons from the financial crisis. Fed. Reserve Bank St. Louis Rev 92 (3), 165–176.

Taylor, J.B., 2011. An empirical analysis of the revival of fiscal activism in the 2000s. J. Econ. Lit. 49 (3), 686–702.

Uhlig, H., 1999. A toolkit for analysing nonlinear dynamic stochastic models easily. Ramon marimon and andrew scott: computational methods for the study of dynamic economies. Oxford University Press, Oxford and New York, pp. 30–61.

Valentinyi, A., Herrendorf, B., 2008. Measuring factor income shares at the sector level. Rev. Econ. Dyn. 11 (4), 820–835.

Watson, M.W., 1986. Univariate detrending methods with stochastic trends. J. Monet. Econ. 18 (1), 49–75.

Ziebarth, N.L., 2014. Misallocation and productivity in the Great Depression. Unpublished.

CHAPTER 27

Macroeconomics of Persistent Slumps

R.E. Hall
Hoover Institution, Stanford University, CA; National Bureau of Economic Research, Cambridge, MA, United States

Contents

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.010

Abstract

In modern economies, sharp increases in unemployment from major adverse shocks result in long periods of abnormal unemployment and low output. This chapter investigates the processes that account for these persistent slumps. The data are from the economy of the United States, and the discussion emphasizes the financial crisis of 2008 and the ensuing slump. The framework starts by discerning driving forces set in motion by the initial shock. These are higher discounts applied by decision makers (possibly related to a loss of confidence), withdrawal of potential workers from the labor market, diminished productivity growth, higher markups in product markets, and spending declines resulting from tighter lending standards at financial institutions. The next step is to study how driving forces influence general equilibrium, both at the time of the initial shock and later as its effects, persist. Some of the effects propagate the effects of the shock—they contribute to poor performance even after the driving force itself has subsided. Depletion of the capital stock is the most important of these propagation mechanisms. I use a medium-frequency dynamic equilibrium model to gain some notions of the magnitudes of responses and propagation.

Keywords

Financial crisis, Great recession, Slump, Unemployment, Labor-force participation, Stagnation, Sources of economic fluctuations, Economic driving forces, Economic shocks, Confidence, Propagation

JEL Classification Codes

E24, E32, G12

Beginning in 2008, output and employment in the United States dropped well below its previous growth path. Eight years later, unemployment is back to normal, but output remains below the growth path. Japan has been in a persistent slump for two decades. And many of the advanced economies of Europe are in slumps, several quite deep. This chapter reviews the macroeconomics of slumps taking the American experience as a leading example.

The adverse shock that launches a slump generally triggers a rapid contraction of output and employment, with a substantial jump in unemployment. This phase—the recession—is usually brief. It ended in mid-2009 in the recent case. The recovery from the trough often lasts many years. The slump is the entire period of substandard output and employment and excess unemployment. In the recent U.S. case, the slump lasted

Table 1 Unemployment in the four serious slumps since 1948

		Ratio of later unemployment rate to peak rate, by number of years later			
Peak year	**Peak rate**	**1**	**2**	**3**	**4**
1975	8.5	0.91	0.84	0.72	0.69
1982	9.7	0.99	0.77	0.74	0.72
1992	7.5	0.92	0.81	0.75	0.72
2010	9.6	0.93	0.84	0.77	0.65

from late 2008 until around the end of 2014. Dating the end of a slump is challenging because some of the state variables accounting for depressed output, notably the capital stock, take many years to return to normal. Output in 2014 was well below its earlier trend path.

Persistent slumps did not begin with the one that originated from the financial crisis of 2008. The Great Depression remains much the deepest and longest slump in the American record since the beginning of national income accounting. Table 1 shows that the persistence of unemployment was about equally high in the four major slumps that occurred after the introduction of the household unemployment survey in 1948. Normal unemployment in the United States, measured as its average over the period starting in 1948, is 6.0%. In all four slumps, unemployment remained above normal 3 years following the peak of unemployment, and in only one slump, the milder one associated with the recession of 1990–91, did unemployment drop below normal 4 years after the peak of unemployment.

Other accounts of persistent shortfalls in output and employment, focusing on the financial crisis and its aftermath, include Kocherlakota (2013), Christiano et al. (2016), Christiano et al. (2010), Benigno and Fornaro (2015), Petrosky-Nadeau and Wasmer (2015), Gertler et al. (2008), Mian and Sufi (2010), Reifschneider et al. (2013), Hall (2013, 2014).

1. THE SLUMP FOLLOWING THE 2008 FINANCIAL CRISIS

This section provides the factual foundation for the chapter by describing events in the U.S. economy around the time of the 2008 crisis, through to 2014. I provide plots of key macroeconomic variables with brief discussions. The rest of the chapter considers the ideas and models that seem most relevant to understanding those events.

Fig. 1 shows that real GDP fell dramatically right after the crisis and remained below its prior growth path even 6 years after the crisis. Plainly the crisis had a persistent effect on the total output of goods and services. Fig. 2 shows that real consumption expenditures behaved similar to real GDP, with no sign of regaining its earlier growth path over the

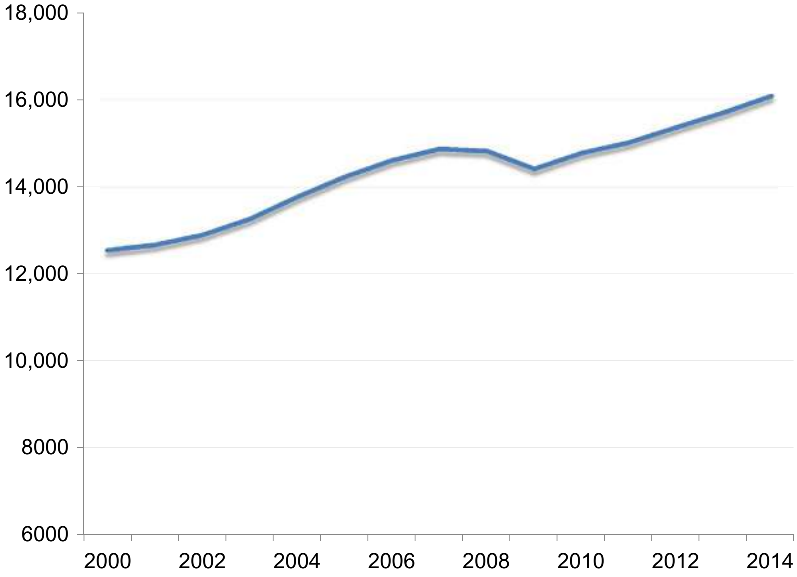


Fig. 1 Real GDP, 2000–14, billions of 2009 dollars.

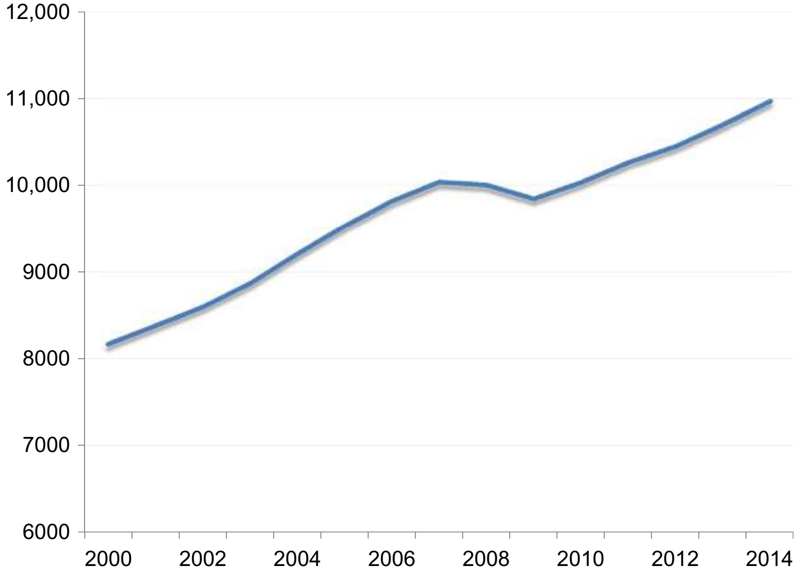


Fig. 2 Real consumption expenditure, 2000–14, billions of 2009 dollars.

period following the 2008 crisis. Fig. 3 shows persistent shortfalls from the growth path of employment. Fig. 4 shows that unemployment rose to a high level and returned to its long-run average of 5.8% at the end of 2014, 6 years after the crisis. The unemployment rate is the only major macroeconomic indicator that returned to normal within the 6-year

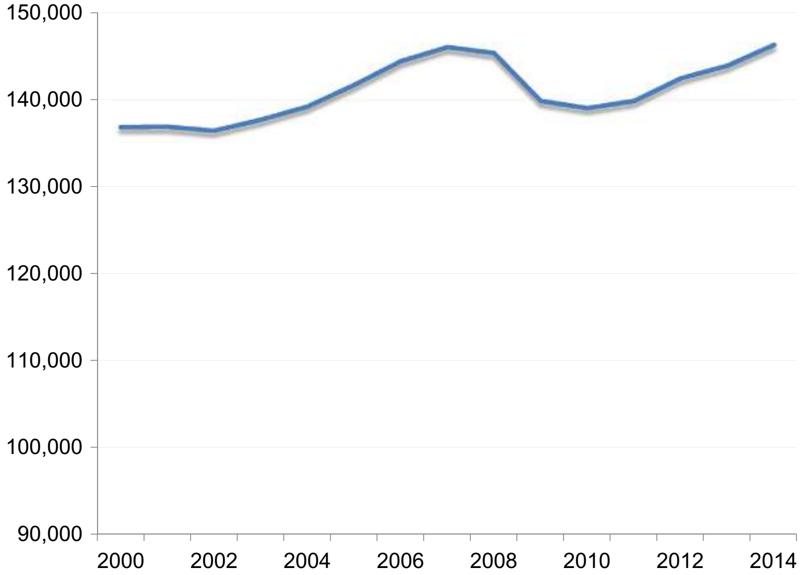


Fig. 3 Employment, 2000–14, thousands of workers.

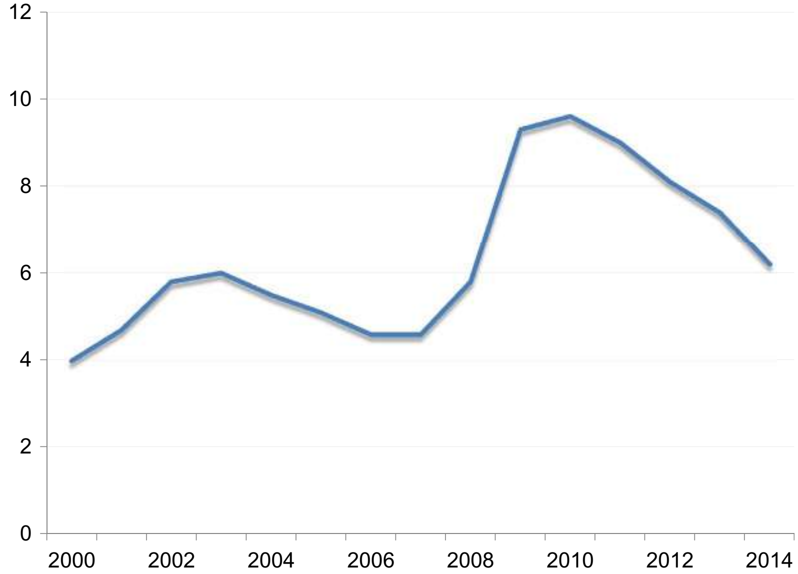


Fig. 4 Unemployment, 2000–14, percent of labor force.

period considered here. Fig. 5 shows that the labor force shrank after the crisis, relative to the working-age population, and that no recovery of the labor force occurred during the recovery. Fig. 6 shows that average real compensation per household, which had grown briskly through 2000, flattened before the crisis, fell sharply just after the crisis, and only

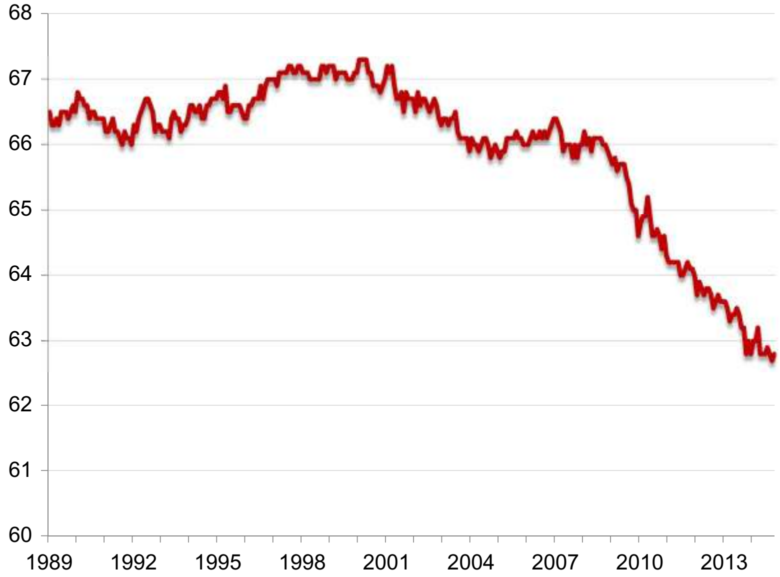


Fig. 5 Percent of working-age population in the labor force, 2000–14.

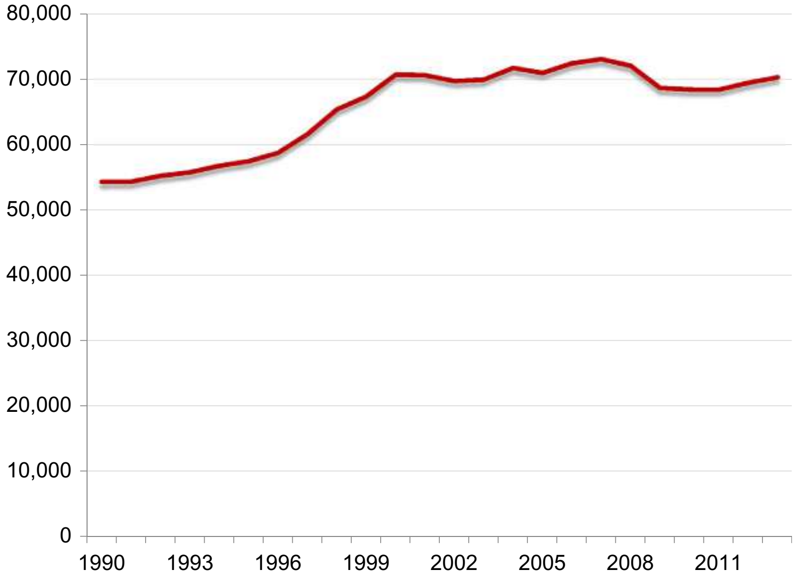


Fig. 6 Average real earnings per household, 2009 dollars, 1990–2014.

regained its previous level in 2014. Fig. 7 shows that the business capital stock—in the sense of an index of capital services available to private businesses—grew much less rapidly than normal immediately after the crisis. Its growth rate returned closer to normal, but left a considerable shortfall in capital relative to trend, as of 2014. Fig. 8 shows that

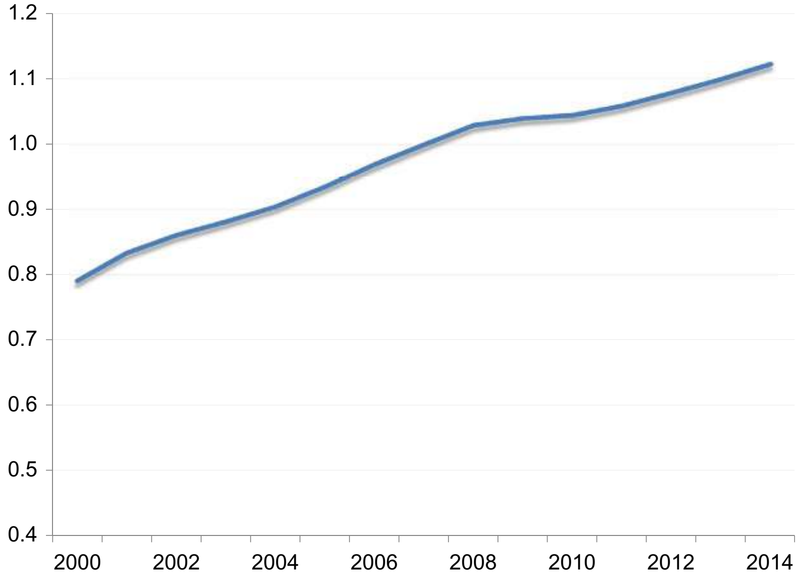


Fig. 7 Index of capital services, 2007 = 1, 2000–14.

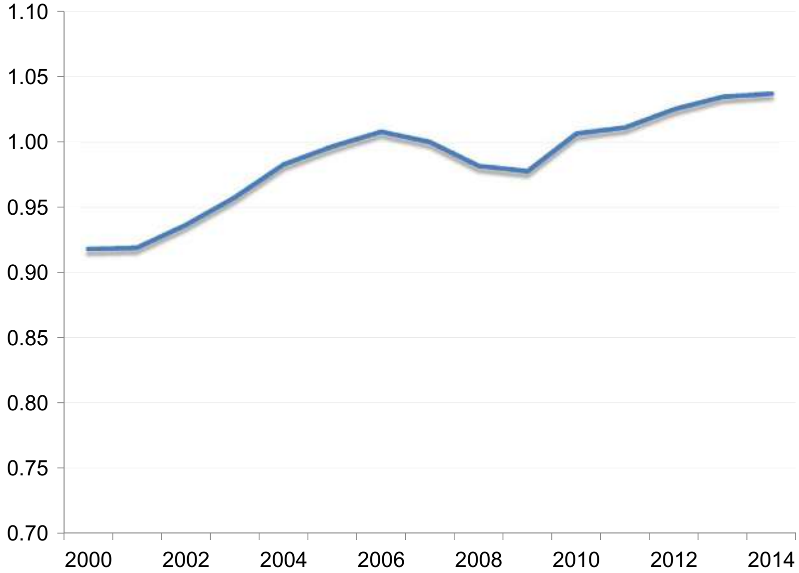


Fig. 8 Index of total factor productivity, 2007 = 1, 2000–14.

private business total factor productivity grew rapidly from 1989 through 2006. A dip in productivity began in 2007. Though productivity grew at normal rates during the recovery, it did not make up for the cumulative decline just after the crisis. Fig. 9 shows the index of the share of the total income generated in the U.S. economy that accrues to

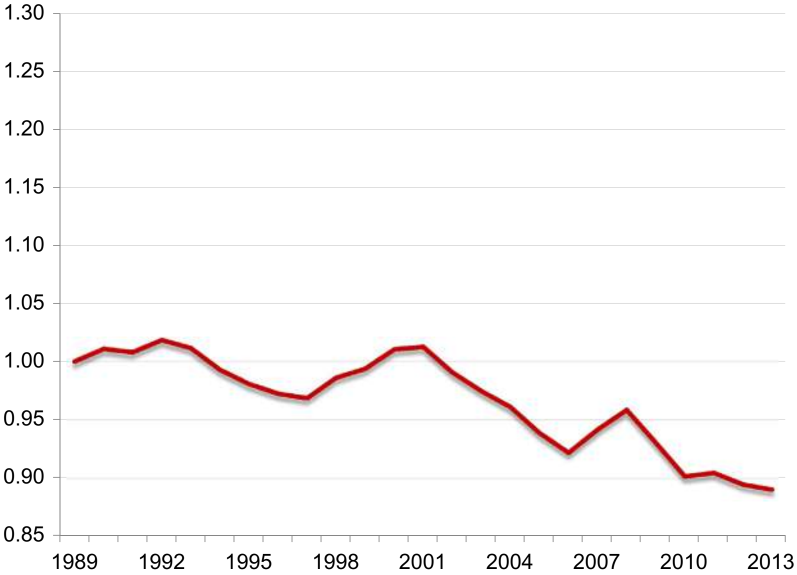


Fig. 9 Labor share.

workers, including fringe benefits. It tends to have a high level in recession years, to fall during the first half of the ensuing expansion, then to rise back to a high level at the next recession. But superimposed on that pattern is a general decline that cumulates to about 10% over the period. Like the general declining trend in earnings, the decline in the share seems to have started around 2000.

2. DRIVING FORCES

I use the term *driving force* to mean either an exogenous variable or an endogenous variable that is taken as an input to a macro model. An example of the latter case is a rise in the discount rate for investment and job creation, triggered by a financial crisis. There is no claim that the discount increase is exogenous. Rather, the hypothesis is that a process outside the model—say a collapse of house prices—influences the model through a higher discount rate. The same process outside the model may enter the model through more than one driving force. For example, the collapse of housing prices may also affect consumption demand by lowering borrowing opportunities of constrained households.

Here I provide an informal review of the driving forces that macroeconomics has identified to account for persistent slumps.

2.1 Labor-Force Participation

A discovery in recent U.S. experience has been the importance of a major decline in labor-force participation. In past slumps, participation remained close to unchanged—the economy has not had a consistent tendency for the labor force to shrink when job finding became more difficult. As of 2015, the U.S. labor market had returned to normal tightness, as measured by job-finding and job-filling rates, yet a large decline in participation starting around 2000 has not reversed. The decline in participation is an important contributor to the divergent behavior of output and employment, on the one hand, and labor-market tightness, on the other hand. Judged by the latter, the slump triggered by the financial crisis of 2008 is over, yet output and employment are far below the paths expected just prior to the crisis.

Movements in participation not directly tied to labor-market tightness need to be added to the list of phenomena associated with episodic slumps. Even if a major shock did not cause a subsequent decline in participation, if a decline happens to occur during a slump, the shortfall in employment and output will be negatively affected.

Elsby et al. (2013) is a recent investigation of the decline in participation. Autor (2011) describes the disability benefits that may be a contributor to that decline.

2.2 The Capital Wedge

A key fact in understanding the slump following the financial crisis is the stability of business earnings. Fig. 10 shows the earnings of private business (the operating surplus from the NIPAs, revenue less noncapital costs) as a ratio to the value of capital (plant, equipment, software, and other intangibles, from the Fixed Assets account of the NIPAs). Earnings fell in 2007 from their normal level of just over 20%, but recovered most of the way by 2010, when output and employment remained at seriously depressed levels.

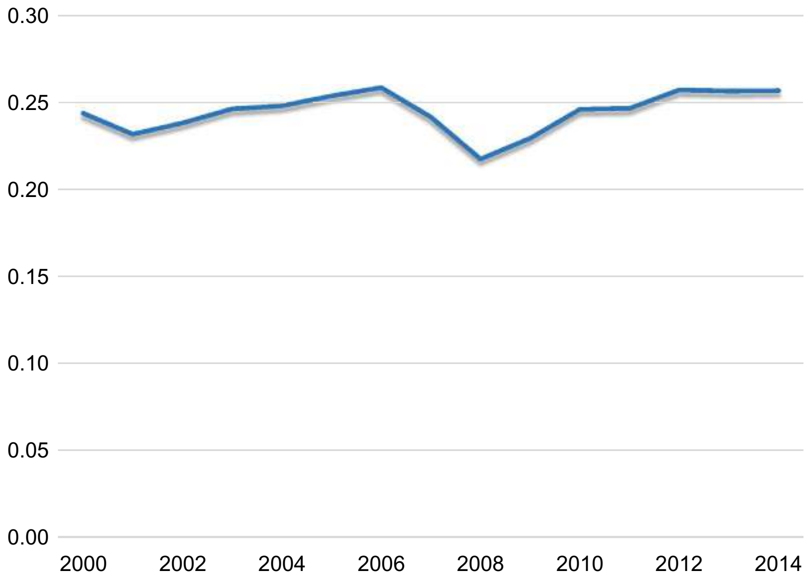


Fig. 10 Business earnings as a ratio to the value of capital.

A basic question is why investment fell so much despite the continuing profitability of business activities. Macroeconomics has gravitated toward an analysis of wedges as ways of describing what seem to be failures of incentives. The capital wedge is the difference between the measured return to investment and the financial cost of investment. I take the latter to be the risk-free real interest rate. The risk premium is one component of the wedge between the return to business capital and the risk-free interest rate. Other components are taxes, financial frictions, and liquidity premiums. To measure the total wedge, I calculate the annual return to capital and subtract the 1-year safe interest rate from it. Later, I decompose the total wedge into one component, interpreted as an extra discount on risky capital earnings not explained by finance theory, and a second, interpreted as an extra premium on safe returns not explained by finance theory.

The calculation of the return to capital uses the following thought experiment: A firm purchases one extra unit of investment. It incurs a marginal adjustment cost to install the investment as capital. During the year, the firm earns incremental gross profit from the extra unit. At the end of the year, the firm owns the depreciated remainder of the one extra unit of installed capital. Installed capital has a shadow value measured by Tobin's q.

Installation incurs a marginal cost at the beginning of the period of $\kappa(k_t/k_{t-1} - 1)$. Thus the shadow value of a unit of installed capital at the beginning of the year is

$$q_t = \kappa\left(\frac{k_t}{k_{t-1}} - 1\right) + 1 \tag{1}$$

units of capital. From its investment of a unit of capital at the beginning of year t together with the marginal installation cost—with a total cost of $q_t p_{k,t}$—the firm's nominal return ratio is the gross profit per unit of capital π_t/k_t plus the depreciated value of the capital in year $t + 1$, all divided by its original investment:

$$1 + r_{k,t} = \frac{1}{q_t p_{k,t}}\left[\frac{\pi_t}{k_t} + (1 - \delta_t) q_{t+1} p_{k,t+1}\right]. \tag{2}$$

Gross profit includes pretax accounting profit, interest payments, and accounting depreciation. In principle, some of proprietors' income is also a return to capital—noncorporate business owns significant amounts of capital—but attempts to impute capital income to the sector result in an obvious shortfall in labor compensation measured as a residual. The reported revenue of the noncorporate business sector is insufficient to justify its observed use of human and other capital. Note that business capital as measured in the NIPAs now includes a wide variety of intangible components in addition to plant and equipment.

The implied wedge between the return to capital and the risk-free real interest rate $r_{f,t}$ is the difference between the nominal rate of return to capital and the 1-year safe nominal interest rate:

$$r_{k,t} - r_{f,t}. \tag{3}$$

This calculation is on the same conceptual footing as the investment wedge in Chari et al. (2007), stated as an interest spread. Note that the wedge is in real units—the rate of inflation drops out in the subtraction.

Fig. 11 shows the values of the business capital wedge for two values of the adjustment cost parameter κ, calculated from Eq. (3), combining plant, equipment, and intellectual property. On the left, κ is taken as 0 and on the right, as 2. The former value accords with the evidence in Hall (2004) and the latter with the consensus of other research on capital adjustment costs. The value $\kappa = 2$ corresponds to a quarterly parameter of 8.

The two versions agree about the qualitative movements of the wedge since 1990, but differ substantially in volatility. The wedge was roughly steady or falling somewhat during the slow recovery from the recession of 1990, rose to a high level in the recession of 2001, declined in the recovery, and then rose to its highest level after the crisis. The two calculations agree that the wedge remained at a high level of about 18% per year through 2013.

Hall (2011a) discusses the surprising power of the financial wedge over general economic activity. The adverse effect of the wedge on capital formation cuts market activity in much the same way as taxes on consumption or work effort.

One branch of the recent literature on the propagation of financial collapse into a corresponding collapse of output and employment emphasizes agency frictions in businesses and financial intermediaries. The simplest model in the case of an intermediary—completely dominant in this literature though not obviously descriptive of the actual U.S. economy—grants the intermediary the opportunity to abscond with the investors' assets. Absconding takes place if the intermediary's continuation value falls short of the value of absconding, taken to be some fraction of the amount stolen from the investors. If the intermediary's equity falls on account of a crisis—for example, if mortgage-backed securities suffer a large capital loss—the investors need to restore the intermediary's incentive to perform by granting a larger spread between the lending rate the intermediary earns and the funding rate it pays to the investors. Hence spreads rise after a financial crisis. This view is consistent with the actual behavior of the spread between the return to capital and the risk-free rate.

The same type of agency friction can occur between a nonfinancial business and its outside investors. Depletion of the equity in the business will threaten the investors' capital. They need to raise the rents earned by the business to increase the continuation values of the insiders, and again spreads will rise.

Gertler and Kiyotaki (2011) cover this topic thoroughly in a recent volume of the *Handbook of Monetary Economics*. Brunnermeier et al. (2012) is another recent survey. Key contributions to the literature include Bernanke et al. (1999), Kiyotaki and Moore (2012), Gertler and Karadi (2011), Brunnermeier and Sannikov (2014), and Gertler and Kiyotaki (2011). See also Krishnamurthy and Vissing-Jorgensen (2013), He and Krishnamurthy (2015), Adrian et al. (2012), and Korinek and Simsek (2014).

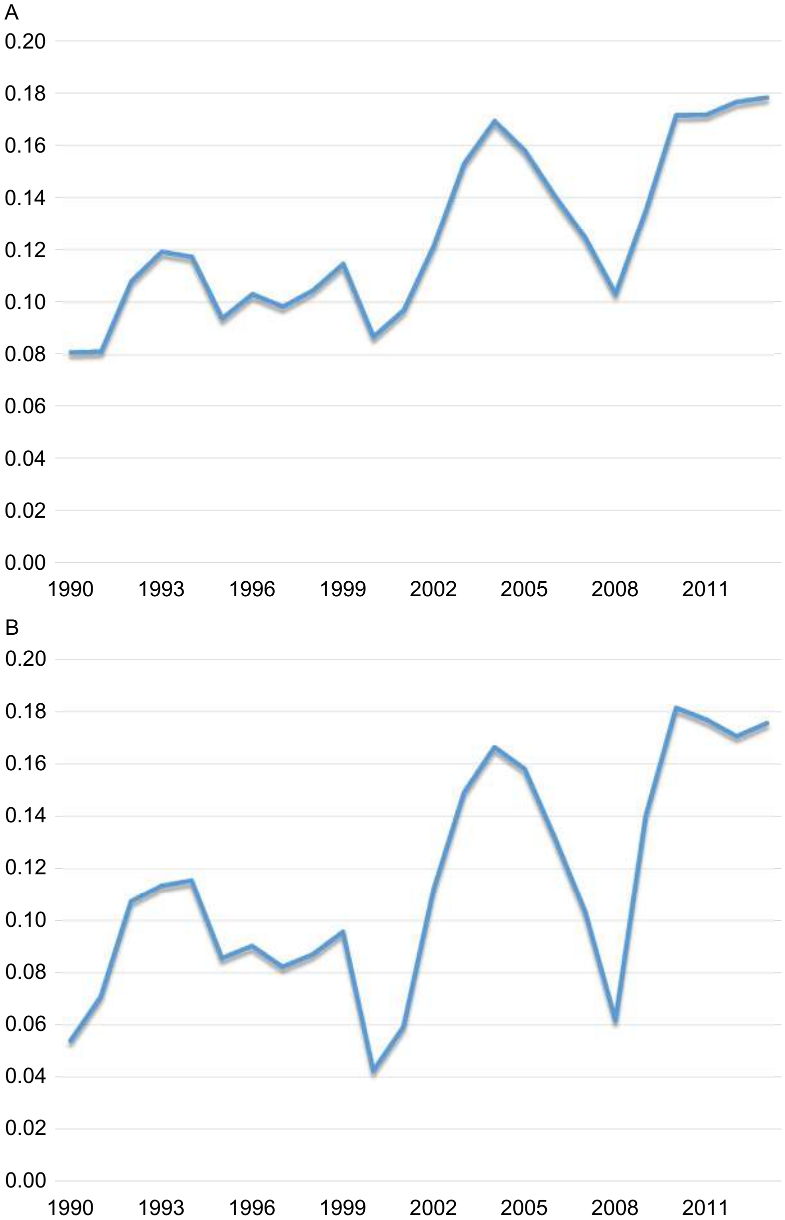


Fig. 11 The capital wedge for two values of the adjustment cost κ. (A) $\kappa = 0$ and (B) $\kappa = 2$.

2.3 Discounts and Confidence

A second branch of the literature linking financial collapse to rising spreads considers widening risk premiums in crises and ensuing slumps. Cochrane (2011) discusses the high volatility of the risk premium in the stock market, measured as the discount rate less

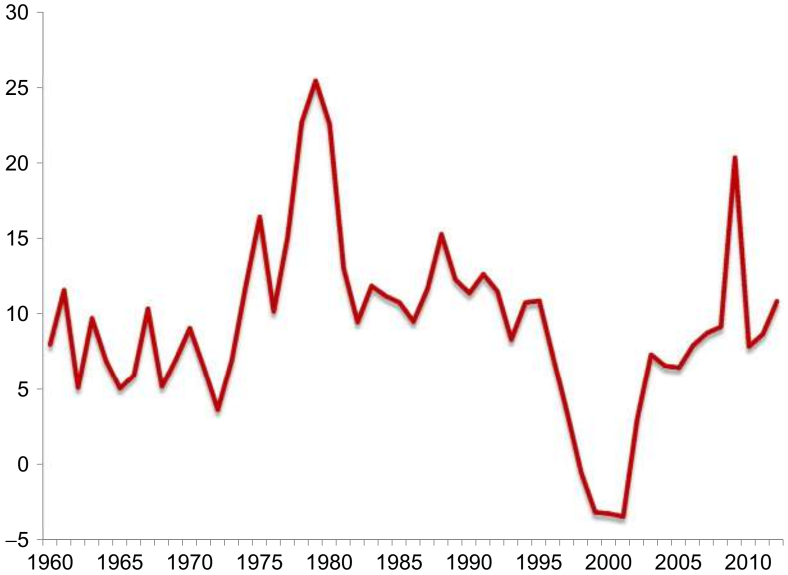


Fig. 12 The S&P Risk Premium, 1960 through 2012.

the risk-free rate. Lustig and Verdelhan (2012) document the tendency for discounts to rise in slumps.

A basic property of the stock market is that, when the level of the stock market is low, relative to a benchmark such as dividends, discounts are higher—see Campbell and Shiller (1988). Normalized consumption is another reliable predictor of returns. Fig. 12 shows the equity premium for the S&P stock-price index from a regression of annual returns on those two variables (see Hall, 2015 for further discussion and details of its construction). The risk premium spiked in 2009. Notice that it is not nearly as persistent as the slump itself—the premium was back to normal well before unemployment fell back to normal and long before investment recovered.

Macroeconomics and finance are currently debating the explanation for the high volatility of discounts. In principle, high discounts arise when the marginal utility of future consumption is high. Generating this outcome in a model is a challenge. Marginal utility would need to be highly sensitive to consumption to generate observed large movements in discounts from the modest expected declines in consumption that occur even in severe contractions. Contractions in consumption appear to be almost completely surprises. If a model implied that occasional drops in consumption occurred as surprises, and consumption then grew faster than normal to regain its previous growth path, the discount rate would *fall* after a crisis because marginal utility would be lower in the future.

Fig. 13 shows the history of the growth of real consumption of nondurable goods per person from 2001 through 2014. The largest decline was in 2009, at 2.5%, about 3.5% below its normal growth. With a coefficient of marginal utility with respect to

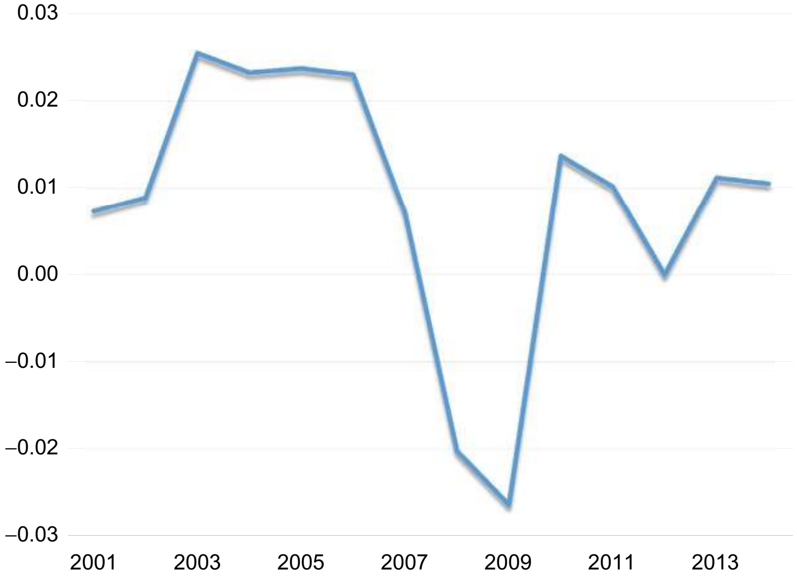


Fig. 13 Growth rate of real consumption of nondurable goods per person.

consumption of 2 (elasticity of intertemporal substitution of 0.5), the effect on marginal utility would be a substantial 7%. But this applies to a fully foreseen decline. The process for consumption change is close to white noise, so the hypothesis of a large negative expected change seems untenable.

Bianchi et al. (2012) propose a mechanism to overcome the problem that expected increases in marginal utility are inconsistent with the observed behavior of consumption. They disconnect discounts from rational expectations of changes in marginal utility by invoking ambiguity aversion. Investors form discounts based on their perceptions of a bad-case realization of marginal utility. During periods when investors have unusually pessimistic views, discounts are high.

Angeletos et al. (2014) overcome the problem in a related way. Investors form expectations about the future state of the economy based on biased beliefs about beliefs of other decision makers. When these second-order beliefs are unusually pessimistic, investors believe that their own future consumption will be lower and their future marginal utility higher, and thus apply higher discounts. The authors use the term *confidence* to refer to optimism in second-order beliefs.

In general, if a financial crisis or other salient event causes investors to shift their beliefs toward higher future marginal utility, discounts will rise. To the extent that the mean of future marginal utility rises, the safe real rate will increase along with the discounts applied to risky returns. To harness the mechanism to explain the decline in the safe rate in the Great Slump along with the rise in the risky discount, the change in the distribution of future marginal utility needs to lower the mean but raise the expected product of marginal utility and the payoffs that govern the levels of employment and output.

The spreads between yields on risky and safe bonds of the same maturity are informative about variations in discounts. Philippon (2009) argues that the bond spread may be more informative. Because the difference in the values of a risky bond and a safe bond is sensitive only to shocks that alter payoffs conditional on default, and default is relatively rare for bonds, the bond spread encodes information about the rare, serious events that could account for high discounts on business income and low discounts on safe payoffs. Fig. 14 shows the option-adjusted spread between BBB-rated bonds and Treasurys of the same maturity.

The spread widened dramatically in 2009, supporting the hypothesis that the perceived probability of a collapse of business cash flow had increased substantially. But the widening was transitory. The spread returned to historically normal levels in 2010 and remained there subsequently. It would take a powerful propagation mechanism for the change in perceptions to account for the persistent slump after 2010.

Gilchrist et al. (2014b), figs. 2 and 3, show IRFs for a spread shock, derived from a vector autoregression. These show relatively little persistence in the shock, but substantial persistence in investment and GDP responses. See also Cúrdia and Woodford (2015).

Other contributions relating to discounts and confidence include Kozlowski et al. (2015), Farmer (2012), He and Krishnamurthy (2013), Gourio (2012), Bianchi et al.

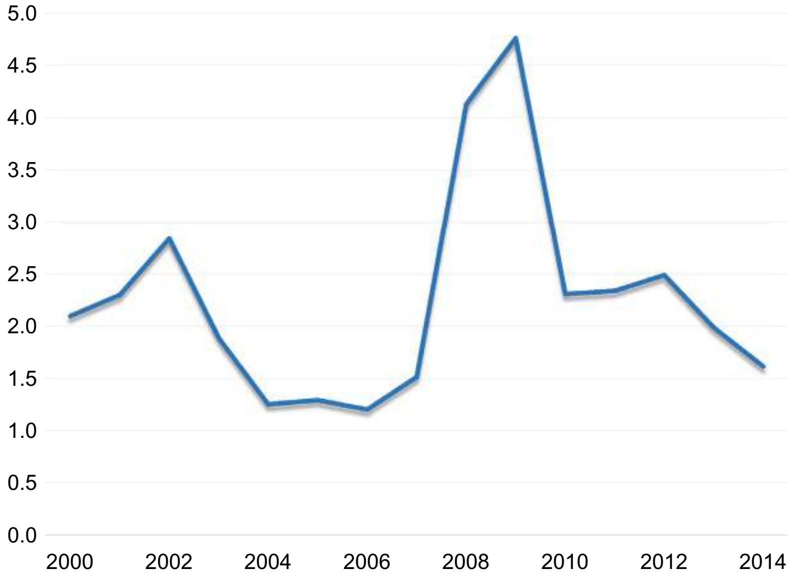


Fig. 14 BBB-treasury bond spread.

(2012), Lustig et al. (2013), and Eckstein et al. (2015). A related topic is the role of fluctuations in uncertainty as a driving force—see Ludvigson et al. (2015) for cites and discussion.

2.4 Productivity

A decline in TFP growth was an important factor in the shortfall of output during the post-crisis U.S. slump. Fernald (2014) makes the case that the productivity slowdown was unrelated to the crisis. Rather, he argues, it was a slowdown only relative to rapid TFP growth in the late 1990s and the early 2000s, associated with adoption of modern information technology. The episode illustrates the importance of TFP growth as a driving force of medium-term fluctuations, even though TFP is not a consistent driver of sharp contractions.

2.5 Product-Market Wedge

Market power in product markets creates a wedge that has been discussed extensively as a driving force of fluctuations, mainly in the context of the new Keynesian model. Rotemberg and Woodford (1999) discuss how sticky product prices result in cyclical fluctuations in markups—in a slump, prices fall less than costs, so market power rises. In almost any modern macro model, the market-power wedge has a negative effect on employment and output. Nekarda and Ramey (2013) question the evidence supporting this view, with respect to shocks apart from productivity. Bils et al. (2014) defend the view, using new evidence.

Gilchrist et al. (2014a) show that firms facing higher financial stress after the crisis raised prices (and thus the wedge) relative to other firms, a finding that supports the idea that the product-market wedge rose in general when overall financial stress worsened. The likely mechanism is different from the one in Rotemberg and Woodford (1999)—it is an idea launched in Phelps and Winter (1970). Financially constrained firms borrow, in effect, by raising prices relative to cost and shedding some of their customer bases.

Chari et al. (2007) provide a comprehensive discussion of wedges in general. See also Gourio and Rudanko (2014).

2.6 Household Deleveraging

Survey data also show a belief that lending standards to households tightened, for mortgages, loans against home equity, and unsecured borrowing (mostly credit cards). Mian and Sufi (2010) use detailed geographic data to argue that household credit restrictions caused declines in consumption. Mian and Sufi (2012), Mian et al. (2013), and Dynan (2012) document the relation between economic activity and household debt. Bhutta (2012) uses household data to show that families did not repay debt more quickly than usual during the slump. Rather, they took on less debt as it became more difficult to qualify for loans, thanks to rising lending standards and declining equity for existing homeowners who prior to 2008 were using cash-out refinancing and home-equity loans. See also Blundell et al. (2008), Petev et al. (2012), and De Nardi et al. (2011).

3. PROPAGATION MECHANISMS

3.1 Capital

The capital stock is an important source of propagation in slumps, a point that has escaped analysis in the cycle-around-trend view of fluctuations. Investment falls sharply in slumps, leaving a depleted capital stock in a slump that lasts several years. Capital depletion also helps account for the divergent behavior of output and labor-market tightness. See Gilchrist et al. (2014b) and Gomme et al. (2011).

3.2 Unemployment Dynamics

In the standard search-and-matching model, calibrated as in Shimer (2005), the unemployment rate is a fast-moving state variable. With job-finding rates around 50% per month even during slumps, unemployment converges to the stationary level dictated by tightness and the job-finding rate within a few months. Unemployment dynamics have essentially nothing to do with the persistence of slumps.

Some facts about the U.S. labor market call this view into question. Hall (1995) observed that research on the experiences of workers who lost jobs after gaining substantial tenure gave a quite different view of unemployment. Davis and von Wachter (2011) summarize more recent results with the same conclusion and emphasize the discord between the quick recovery from job loss implicit in the basic search-and-matching model and the actual experience of workers with three or more years of tenure following job loss. That experience involves an extended period of low employment—much greater loss than a 50% per month reemployment rate—and years of loss of hourly earnings. Jarosch (2014) confirms this view. The aggregate implications are that a wave of layoffs from a major shock, such as the financial crisis, results in an extended period of unemployment and a much longer period of lower productivity of the higher-tenure workers who lose jobs from the shock. Ravn and Sterk (2012) develop a model with two kinds of unemployment to capture this type of heterogeneity among the unemployed.

Some progress has been made in reconciling high monthly job-finding rates with the low recovery from high unemployment following a shock. Hyatt and Spletzer (2013) show that short jobs are remarkably frequent—the distribution of job durations is utterly unlike the exponential distribution with a constant separation hazard usually assumed in search-and-matching models. This finding explains the high job-finding rates found in the CPS—there is a huge amount of churn in the U.S. labor market. Hall and Schulhofer-Wohl (2015) show that job-finding rates over year-long periods are well below what would be expected from monthly job-finding rates. The obvious explanation of this finding is that job-seekers often take interim jobs during much longer spells of mixed unemployment and brief employment.

Shimer (2008) discusses the labor-market wedge as a convenient summary of the effects of labor-market frictions.

Other contributions relating to propagation through unemployment dynamics include Valletta and Kuang (2010b), Cole and Rogerson (1999), Chodorow-Reich and Karabarbounis (2015), Davis and von Wachter (2011), Davis et al. (2012), Petrosky-Nadeau and Wasmer (2013), Fujita and Moscarini (2013), Jarosch (2014), Rothstein (2011), Petrosky-Nadeau and Zhang (2013), Mortensen (2011), Valletta and Kuang (2010a), Sahin et al. (2012), Daly et al. (2011a,b, 2012), Kuehn et al. (2013), Mulligan (2012a), Barnichon and Figura (2012), Estevão and Tsounta (2011), Krueger et al. (2014), Herz and van Rens (2011), Sahin et al. (2012), Farber and Valletta (2013), Kaplan and Menzio (2016), Elsby et al. (2011), Krueger and Mueller (2011), Davis and Haltiwanger (2014), Hall (2012), Fujita (2011), Hagedorn et al. (2013), Mulligan (2012b), Restrepo (2015), Farber (2015), and Ravn and Sterk (2012).

3.3 The Zero Lower Bound

The policy of every modern central bank is to issue two types of debt: reserves and currency. The bank pays interest or collects negative interest on reserves. No direct force constrains the rate on reserves. It is impractical to pay or collect interest on currency. Central banks keep currency and reserves at par with each other by standing ready to exchange currency for reserves or reserves for currency in unlimited amounts. If the bank sets a reserve rate below the negative of the storage cost of currency, owners of reserves will convert them to higher-yielding currency. A number of European central banks have experimented recently with increasingly negative reserve rates.

The lower bound on the real interest rate is the bound on the nominal rate less the expected rate of inflation. Fig. 15 shows three time series relevant for measuring expected inflation. The top line is the median expected rate of inflation over the coming year for the Michigan Survey of Consumers. The line starting in 2007 is the median forecast of the average annual rate of change of the PCE price index over the coming 5 years, in the Survey of Professional Forecasters of the Philadelphia Federal Reserve Bank. The bottom line is the breakeven inflation rate in the 5-year TIPSs and nominal 5-year note—the rate of inflation that equates the nominal yields of the two instruments. See also Fleckenstein et al. (2013) on extracting expected inflation from inflation swaps.

The three measures agree that essentially nothing happened to expected inflation over the period of the post-crisis slump. All recorded a drop around the time of the crisis, but

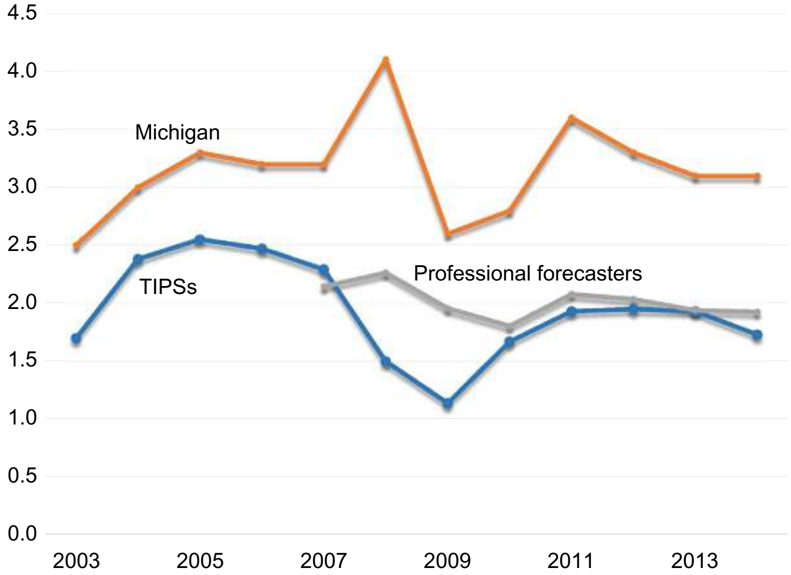


Fig. 15 Inflation expectations and forecasts.

then returned to close to precrisis levels despite high unemployment. This finding pretty much eliminates an idea that permeated macroeconomics over the past 50 years, that slack more or less automatically results in lower inflation. Some combination of factors in 2008 prevented the collapse of the price level that occurred, for example, in the much deeper slump following the contraction of 1929–33.

Had expected inflation declined by the amounts that occurred in the earlier slumps of the past 50 years, the influence of the zero lower bound on the real interest rate would have been more severe. And if deflation at the rate experienced in 1929–33 had occurred, a catastrophe similar to the Great Depression would probably have occurred. Good fortune kept expected inflation at normal levels and avoided high real interest rates and their likely adverse effects on output and employment.

In view of the importance of the inflation rate in determining the real interest rate corresponding to a zero nominal rate, the complete absence of a model of inflation is a considerable shortcoming of current macroeconomic thinking. About the best that macro modeling can do is to take expected inflation as an exogenous constant, currently around 2%. It is common for macroeconomists to say that "inflation is firmly anchored at the Fed's target of two percent" as if that amounted to a model. But it is not—at best it is an observation that expected inflation has remained at about that level despite large changes in output, employment, and other macro variables.

With exogenous, constant inflation, the bound on the nominal interest rate places a bound on the safe real rate at the nominal bound minus the rate of inflation—minus

2% in the recent slump if the nominal bound is zero; minus 3% if the nominal bound is minus 1%.

Stock and Watson (2010) study the joint behavior of inflation and unemployment with conclusions similar to those stated here. Ball and Mazumder (2011) argue in favor of the conventional view that inflation has a stable relation to slack.

3.3.1 Incorporating the Zero Lower Bound in Macro Models

Hall (2011b) discusses the issues in modeling an economy with a safe real rate fixed above the value that would clear the output market under normal conditions. In brief, the high real rate creates the illusion of an opportunity to defer consumption spending when deferral is actually infeasible. Because of the mispricing of the benefit of saving, consumers create congestion as they try to save and defer spending. Congestion arises from the same force that slows traffic on a highway that is underpriced, so more drivers try to use it than its capacity. As a practical matter, the congestion appears to take the form of low job-finding rates and abnormally high unemployment.

Modeling of the congestion resulting from the mispricing of saving is still at a formative stage. To frame the issue, consider a simple frictionless general-equilibrium macro model with a unique equilibrium. The model will describe an equilibrium value of the short-term safe real interest rate. Now implant a central bank in the model with a policy of setting that rate at a value above the equilibrium value. In particular, suppose that the bank's interest rate is elevated by the zero lower bound. What happens in the model? It cannot have an equilibrium—its only equilibrium is ruled out by assumption. One solution in macro theory is to disable one equation. Then the model has one less endogenous variable, the interest rate (made exogenous by the zero lower bound), and one less equation. One example is to drop a clearing condition for the labor market and to interpret the gap between labor supply and labor demand as unemployment. When the central bank sets a rate above equilibrium, labor demand will fall short of labor supply and unemployment will be above its normal level. This approach has some practical appeal and often gives reasonable answers.

A closely related approach is to place the demand gap in the product market. Krugman (1998) and Korinek and Simsek (2014) are examples of that approach. Farhi and Werning (2013) present a general analysis of demand gaps, where any set of prices and wages can be jointly restricted and gaps can occur in any market.

The new Keynesian tradition takes a different and more subtle approach to this issue by adding the price level as another endogenous variable without any corresponding equation. The model has demand gaps in the product market associated with temporarily sticky prices that adjust over time to close the gaps. Eggertsson and Krugman (2012) and Christiano et al. (2011) apply the NK model to the zero lower bound issue. One branch of the NK literature—notably Walsh (2003), Gertler et al. (2008), and, most recently, Christiano et al. (2016)—uses the Diamond–Mortensen–Pissarides framework to describe the labor market, so the only role of demand gaps is in the product market.

Hall (2016) tackles the congestion issue directly, in the DMP setup. Both the output and labor markets suffer from congestion when the central bank elevates the real rate above the market-clearing level. The central bank's acceptance of deposits at the elevated real rate creates an outside option in the product-price bargain that creates slack according to standard DMP principles.

In general, a model that combines the DMP view of unemployment with a real interest rate held above its market-clearing level will incorporate an additional variable, analogous to congestion in the highway case, that changes the DMP unemployment rate and the demand-gap rate until they are equal. To be concrete about that variable, suppose it is matching efficiency. A decline in efficiency increases hiring cost, raises the cost of labor, lowers the demand for labor, and raises demand-gap unemployment. The decline in efficiency lowers the job-finding rate and raises the DMP unemployment rate. The second effect is robust in the DMP model and presumably exceeds the effect on demand-gap unemployment. In equilibrium, unemployment is less than demand-gap unemployment would be at normal matching efficiency but higher than DMP unemployment would be at normal efficiency. The model would need to tie matching efficiency to the spread between the bank's interest rate and the rate that cleared the output and labor markets. Though this mechanism is attractive because matching efficiency did fall after the 2008 crisis, I do not have a model embodying variations in matching efficiency. The model in Hall (2016) is rather more complicated and invokes DMP principles in both product and labor markets.

If the effect of congestion in the labor market on labor demand is small enough to be neglected, the gap between labor supply and labor demand controls unemployment. In this case, the traditional view that ignores DMP-type considerations applies. In that case, the general-equilibrium model simply omits the DMP-based equations. In the background, labor-market congestion fluctuates to bring unemployment into line with the level dictated by product demand. In the model later in this chapter, I take this approach as an interim solution pending development of fully articulated models of congestion induced by above-equilibrium real interest rates.

Michaillat and Saez (2014) build a model of labor- market congestion that differs from the DMP model in one crucial respect—it lacks a resource decision to control the tightness of the market. In the DMP model, recruiting effort determines the tightness of the labor market. Employers expand recruiting effort until the payoff to creating an incremental vacancy equals the expected recruiting cost. In a simple real business-style macro model with a DMP labor market, equilibrium is determinate. By contrast, in the model of Michaillat and Saez, the corresponding basic model is indeterminate. It has a continuum of equilibria indexed by the real interest rate, with tightness depending on that rate. A monetary intervention that sets the real interest rate picks out one of those equilibria.

Adding that monetary intervention to the DMP-based model would make it overdetermined.

This discussion presupposes that the central bank can set any path it chooses for the real interest rate. Friedman (1968) reached the opposite conclusion. In his view, a bank that tried to keep the real rate below the market-clearing level would cause exploding inflation (the case that concerned him in 1967), and a policy aiming to keep the real rate above the market-clearing level presumably would cause exploding deflation. Recent experience does not bear his prediction out—the lower bound froze the safe real rate at around minus 2% because expected inflation remained unchanged at around 2% per year. Our lack of understanding of inflation stands in the way of fully satisfactory modeling of central bank policies that control the real interest rate.

See also Attanasio and Weber (1995), Correia et al. (2010), Eggertsson and Krugman (2012), Cochrane (2014), Hall (2016), and Eggertsson and Mehrotra (2014).

3.3.2 The Zero Lower Bound and Product Demand

The zero lower bound, together with low expected inflation, has prevented central banks from lowering interest rates as much as would seem appropriate. Lower rates should stimulate output and employment. The Federal Reserve and the Bank of Japan have kept rates slightly positive since the crisis, while the European Central Bank did the same until recently, when it pushed the rate just slightly negative. All three economies had combinations of high unemployment and substandard inflation that unambiguously called for lower rates, according to standard principles of modern monetary economics. Under normal conditions, fluctuations in product demand are not a source of important fluctuations in output and employment, because interest rates change as needed to clear those markets. Under almost any view of purposeful monetary policy, the central bank adjusts its policy rate in response to those demand fluctuations. But the zero lower bound is an exception to that principle. Economies with low inflation rates and low equilibrium real interest rates run the danger of episodic slumps when the lower bound is binding.

In the slump that began in 2008, three driving forces for product demand appeared to be important: rising discounts, tightening lending standards to businesses, and tightening lending standards to households. All three of these declines may also reflect the rising importance of another driving force, financial frictions. Other sources could be declining government purchases and transfers and declining export demand. In the recent slump, government purchases fell slightly relative to trend, transfers rose dramatically, and exports fell.

3.3.3 Discounts

As documented elsewhere in this section, discounts applied to future risky cash flows appeared to rise dramatically during and immediately after the financial crisis. Basic principles of investment theory hold that purchases of new capital goods decline when discounts rise. In fact, all three major categories of investment fell sharply: (1) business purchases of new plant, equipment, and intellectual property, (2) residential construction, and (3) autos and other consumer durables. Eggertsson and Krugman (2012) describe how a rise in discounts pushes the economy into a regime where the zero lower bound binds.

3.3.4 Lending Standards to Businesses

Survey data show unambiguously that bank officials believe that they tightened lending standards after the crisis. It remains controversial whether the tightening is an independent driving force or just a symptom of other adverse forces. Chodorow-Reich (2014), using data on individual bank–borrower relationships, argues for a separate role for tightening standards. Tighter standards may also be a driving force for the sharp decline in residential construction, given the dependence of major house-builders on bank lending.

3.3.5 Lending Standards to Households

I noted earlier that rising lending standards and declining equity resulted in cutbacks in consumption because families who had previously financed high consumption levels in part by taking on more and more debt could no longer qualify for those loans.

4. FISCAL DRIVING FORCE AND MULTIPLIER

The multiplier is the derivative of total GDP or a component, such as consumption, with respect to an exogenous shift in product demand. The obvious source of such a shift is government purchases, but the same multiplier describes the propagation of other shifts in product demand, notably those induced by changes in household access to credit.

Ramey (2011a) is a recent survey of the literature on the multiplier, and her chapter in this volume also treats the subject in detail. See also Coenen et al. (2012), Shapiro and Slemrod (2009), Spilimbergo et al. (2009), Hall (2009), Barro and Redlick (2011), Parker et al. (2011), Kaplan and Violante (2014), and Ramey (2011b).

5. OTHER ISSUES

5.1 Decline in the Labor Share

Economists have pursued multiple explanations of the decline, but no consensus has formed. Rognlie (2015) provides a comprehensive discussion of this topic. See also Karabarbounis and Neiman (2014).

5.2 Time Use

Some indication about the changing balance between work and other uses of time comes from the American Time Use Survey, which began in 2003. Table 2 shows the change in

Table 2 Changes in weekly hours of time use, between 2003 and 2013, people 15 and older

	Personal care	Household work	Market work	Education	Leisure	Other
Men	1.3	0.1	−2.5	0.2	1.3	−0.4
Women	1.6	−0.7	−0.8	−0.1	0.8	−0.8

weekly hours between 2003 and 2013 in a variety of activities. For men, the biggest change by far is the decline of 2.5 h per week at work, a big drop relative to a normal 40-h work week. A small part of the decline is attributable to higher unemployment—the unemployment rate was 6.0% in 2003 and 7.4% in 2013. The decline for women is much smaller, at 0.8 h per week. For both sexes, the big increases were in personal care (including sleep) and leisure (mainly video-related activities). Essentially no change occurred in time spent in education. Women cut time spent on housework. See also Aguiar et al. (2013).

6. A MODEL

Many macro-fluctuations models omit slower-moving driving forces and are correspondingly estimated or calibrated to data filtered to remove slower movements. Growth models generally omit cyclical and medium-frequency driving forces. A small literature—notably including Comin and Gertler (2006)—deals explicitly with medium-frequency driving forces and corresponding movements of key macro variables. That paper focuses on technology and productivity. The model developed here considers other medium-frequency driving forces, such as labor-force participation and discounts. Hall (2005) discusses evidence of the importance of medium-frequency movements and argues against the suitability of superimposing a high-frequency business-cycle model on an underlying growth model. Instead, a unified model appears to be a better approach.

The model is inherently nonstationary—its labor force grows randomly and so does productivity. Solution methods widely used for stochastic macro models, either near-exact solutions using projection methods or approximate solutions based on log-linearization, require that models be restated in stationary form. I take a different approach. The model has random driving forces that are functions of a Markov discrete state. Over a finite horizon the model has an event space with a large but finite set of nodes. Models with this structure are widely used in finance and banking. I find essentially exact solutions for the contingent values of continuous state variables and other key macro variables at each node. Finance models, such as the binary option-pricing model, have backward-recursive solutions, but macro models require solving the entire model as a system of simultaneous equations. Recursive models are highly sparse, and solution methods that fully exploit the sparsity are fast.

6.1 Specification

The equations of the model are the familiar first-order conditions for optimization by the decision makers in the model and laws of motion of the state variables, together with initial and terminal conditions. The framework does not require that the model be recursive, though the model here is actually recursive—it can be expressed in equations that consider only three dates: *Now* (for example, k), *Soon* (for example, k'), and *Later* (for example, k''). Each value *Now* branches stochastically into N_t values in the *Soon* period and N_t^2 values in the *Later* period. Here N_t is the number of states in the discrete Markov process in period t. The economy operates for T periods.

The driving forces of the model are:

a: increment to total factor productivity

l: increment to the labor force

d_k: discount or confidence with respect to capital

d_n: discount or confidence with respect to job creation

d_f: discount or flight to safety factor with respect to safe 1-year returns (found to be negative, implying a safety premium)

z: product-market wedge arising from market power

g: government purchases of goods and services, serving as a proxy for shifts in product demand arising from forces not considered explicitly in the model

The continuous state variables are:

k: physical capital stock (endogenous)

A: total factor productivity (exogenous)

L: labor force (exogenous)

Endogenous variables that are functions of the state variables are:

y: output

n: employment

c: consumption

q: Tobin's q, the value of installed capital

r': the realized return to holding installed capital from now to later

r_f: the safe real interest rate from now to later, known now

m: the stochastic discounter, not including d_k, d_n, and d_f

x: the marginal revenue product of labor

6.2 States

An integer-valued state s governs the outcomes of random influences on the economy. It follows a Markov process:

$$\text{Prob}[s'|s] = \pi_{s,s'}. \tag{4}$$

6.3 Technology

Output at the beginning of a period combines labor and capital services according to a Cobb–Douglas technology:

$$y' = A' n^{1-\alpha} k^{\alpha}. \tag{5}$$

Installation of capital incurs quadratic adjustment costs. The marginal cost of adjustment, q, is

$$q' = \kappa\left(\frac{k'}{k} - 1\right) + 1. \tag{6}$$

Total factor productivity evolves as

$$A' = \exp(a_{s'})A. \tag{7}$$

Here $a_{s'}$ is a state-dependent log-increment to TFP. The law of motion of the capital stock is

$$k' = (1-\delta)k + y' - c' - g'. \tag{8}$$

Here δ is the rate of depreciation of capital.

6.4 Financial Markets

The realized rate of return to holding capital is

$$r_k' = \frac{\alpha \frac{y'}{z'k} + (1-\delta)q'}{q} - 1. \tag{9}$$

Here z is a product-marked wedge. The economy's normal stochastic discount factor is

$$m' = \beta\left(\frac{c'}{c}\right)^{-1/\sigma}. \tag{10}$$

The pricing condition for the return to capital is

$$\mathbb{E}[(1 + r_k')m'] - d_k = 1. \tag{11}$$

Here d_k is a distortion of the discounter for the return to capital, interpreted as loss of confidence or increased pessimism, that lowers the perceived present value of the future payoff to capital.

The pricing condition for the risk-free rate is

$$\mathbb{E}\ [(1 + r_f)m'] - d_f = 1. \tag{12}$$

Here d_f is a distortion of the discounter for the safe real return, whose negative value is interpreted as a liquidity premium or flight to safety premium.

6.5 The Zero Lower Bound

The model does not embody a bound on the short safe interest rate. Rather, it identifies conditions when the rate is low—generally negative. Times of negative rates are times when the lower bound would be binding, and the model's equilibrium would not actually hold. As noted earlier, macroeconomics has yet to provide a coherent account of equilibrium with a binding lower bound. All the literature simply assumes that a demand gap implies output and employment gaps, without further explanation of why economic behavior results in gaps. The predictions of the demand-gap model may well be correct—the point is that models do not meet normal standards of explanation imposed on modeling other economic phenomena. See Hall (2016) for further discussion of this point.

6.6 Initial and Terminal Values of the Capital Stock

The capital stock grows stochastically along with growth in TFP, A, and the labor force, L. I calculate the initial capital stock and the stock at each terminal node as

$$k^* = (1-u^*)L\left(\frac{\alpha A}{r^*+\delta}\right)^{1/(1-\alpha)}. \qquad (13)$$

Here u^* is the normal unemployment rate. The quantity r^* is the constant discount rate equivalent to actual stochastic discounting, including the extra discount d_k. I pick the value of r^* that generates roughly constant growth of capital. If r^* is below that level, the capital stock grows more rapidly at first until it reaches the stochastic turnpike path, then shrinks back to the terminal condition toward the end. The stock sags below its initial level and grows extra-rapidly at the end of the period if r^* is too high.

6.7 The Labor Market

The model incorporates the idea that hiring is a form of investment, as in the Diamond–Mortensen–Pissarides model of the labor market. As with other forms of investment, the discount rate influences hiring, as discussed with citations in Hall (2015). The equation also takes the marginal revenue product of labor as the measure of the benefit of a hire—subject to variation through changes in market power as in Rotemberg and Woodford (1999), stated in DMP terms in Walsh (2003).

DMP employment depends on the present value of the ratio, $x'/\bar{x}'$ of the actual future marginal revenue product of labor to the normal level based on future technology A' and current capital k. The numerator is

$$x' = (1-\alpha)A'\left(\frac{k}{z'n}\right)^{\alpha} \qquad (14)$$

and the denominator is

$$\bar{x}' = (1-\alpha)A'\left(\frac{k}{\bar{n}}\right)^{\alpha}. \tag{15}$$

There is a downward distortion, d_n, in the discounted value of the ratio. Employment is

$$\begin{aligned} n &= \bar{n}\left[\frac{\mathbb{E}(m'x')}{\mathbb{E}(m'\bar{x}')}\exp(-d_n)\right]^{\omega} \\ &= \bar{n}\left[\left(\frac{\bar{n}}{zn}\right)^{\alpha}\exp(-d_n)\right]^{\omega}. \end{aligned} \tag{16}$$

The value of d_n implied by the data is

$$d_n = -\left(\alpha + \frac{1}{\omega}\right)\log\frac{n}{\bar{n}} - \alpha\log z. \tag{17}$$

Given d_n, the resulting solution is

$$\log\frac{n}{\bar{n}} = -\left(\frac{\omega}{1+\alpha\omega}\right)(\alpha\log z + d_n). \tag{18}$$

The labor force evolves as

$$L' = \exp(l_{s'})L. \tag{19}$$

Unemployment is

$$n = (1-u)L. \tag{20}$$

6.8 Timing

Timing is easiest to understand in the nonstochastic case, where $N_t = 1$ for all periods t. In period 1, capital is at its specified initial value k_I. No consumption occurs in period 1. In period T, capital is at its specified terminal value, k_T. No employment occurs. Consumption c_T is an unknown to be determined. Thus there are $T-2$ values of capital to be determined, k_2 through k_{T-1}, and $T-1$ values of consumption, c_2 through c_T. Given candidate values for these, and the exogenous variables A_t and L_t, one can calculate corresponding candidate values of the other variables, y_t, q_t, $r_{k,t}$, $m_{t,t+1}$, x_t, n_t, and u_t. The $T-1$ residuals of the material balance condition

$$\epsilon_{M,t} = k' - [y' + (1-\delta)k - c' - g'], t = 1 : T-1 \tag{21}$$

and the $T-2$ residuals of the Euler equation

$$\epsilon_{E,t} = \mathbb{E}_t(1 + r_{k,t+1})(m_{t,t+1} - d_t) - 1, t = 2 : T-1 \tag{22}$$

define a system of equations

$$\epsilon(x) = 0. \tag{23}$$

Here $\epsilon(x)$ is the combined vector of the $2T-3$ residuals and x is a vector of the $2T-3$ unknown values of k_t and c_t. A standard nonlinear equation solver finds a solution, which is the dynamic stochastic contingent equilibrium of the model.

6.9 Summary

Equations with a zero on the right-hand side enter the solution with discrepancies ϵ which are driven to zero by Newton's method:

$$\text{Prob}[s'|s] = \pi_{s,s'}, \tag{24}$$

$$k' - (1-\delta)k - y' + c' + g' = 0, \tag{25}$$

$$A' = \exp(a)A, \tag{26}$$

$$L' = \exp(l)L, \tag{27}$$

$$y' = A'n^{1-\alpha}k^{\alpha}, \tag{28}$$

$$q' = \kappa\left(\frac{k'}{k} - 1\right) + 1, \tag{29}$$

$$r_k' = \frac{\alpha\dfrac{y'}{zk} + (1-\delta)q'}{q} - f' - 1, \tag{30}$$

$$m' = \beta\left(\frac{c'}{c}\right)^{-1/\sigma}, \tag{31}$$

$$x' = (1-\alpha)\frac{y'}{zn}, \tag{32}$$

$$\mathbb{E}[(1+r_k')(m'-d)] - 1 = 0, \tag{33}$$

$$r_f = \frac{1}{\mathbb{E}m'} - 1, \tag{34}$$

$$n = (1-u)L, \tag{35}$$

$$\log\frac{n}{\bar{n}} = \left(\frac{\omega}{1+\alpha\omega}\right)\left(\alpha\log\frac{k}{z\bar{k}} - d_n\right). \tag{36}$$

7. APPLICATION TO THE U.S. ECONOMY

7.1 States of the Economy

The model operates at an annual frequency. I constructed its states by the *k*-cluster method with six clusters, based on the following variables measured over the period from 1953 through 2014:

- TFP growth, from Fernald (2012), without utilization adjustment
- The discount implicit in the S&P stock-market index, measured as the expected real return based on the Livingston survey
- The annual growth rate of the civilian labor force
- The unemployment rate

Table 3 shows the discrete states of the model, in terms of the values of the four variables. It also shows the classification of years by state. Each of the four variables defining the states has six state-dependent values. In a row in the table, *Low* refers to the two lowest values of a variable across the states, *Med* to the middle two values, and *High* to the upper two values. Table 4 shows the state-contingent values of the variables that define the states. The states are:

1. Strong economy with low discount, low unemployment, high growth of labor force, and high productivity growth
2. Strong economy with medium TFP growth
3. Mediocre economy with low TFP growth
4. Mediocre economy with high discount and low TFP growth
5. Slump with average TFP growth
6. Slump with high TFP growth

Table 3 The states of the model

State	TFP growth	Discount	Labor-force growth	Unemployment	Years in state
1	Low	Low	High	High	1955, 1957, 1959, 1960, 1964, 1966, 1968, 1969, 1972, 1995, 1996, 1997, 1999, 2000, 2006
2	Low	Low	High	Med	1953, 1956, 1962, 1965, 1973, 1978, 1988, 1989, 1998
3	Med	Med	Med	Low	1954, 1958, 1963, 1967, 1971, 1977, 1979, 1980, 1985, 1986, 1987, 1990, 1991, 1993, 1994, 2007, 2013, 2014
4	High	Med	Med	Low	1961, 1970, 1974, 1975, 1981, 1982, 1983, 2001, 2002, 2004, 2005, 2008
5	High	High	Low	Med	2003, 2009
6	Med	High	Low	High	1976, 1984, 1992, 2010, 2011, 2012

Table 4 State-contingent values of the variables defining the states

	State-contingent value (%)			
State	**Discount**	**Unemployment**	**Labor-force growth**	**TFP growth**
1	2.79	4.67	1.68	2.00
2	−1.84	4.81	1.79	1.80
3	5.40	6.22	1.48	0.43
4	10.73	6.63	1.40	0.27
5	20.74	7.63	0.52	0.92
6	3.94	8.22	1.03	2.42

Table 5 Transition matrix and ergodic distribution

		To state						
		1	**2**	**3**	**4**	**5**	**6**	**Ergodic probability**
From state	1	0.33	0.27	0.20	0.20	0.00	0.00	0.25
	2	0.33	0.11	0.44	0.11	0.00	0.00	0.13
	3	0.35	0.12	0.35	0.12	0.00	0.06	0.30
	4	0.08	0.08	0.08	0.42	0.17	0.17	0.20
	5	0.00	0.00	0.00	0.50	0.00	0.50	0.03
	6	0.00	0.00	0.67	0.00	0.00	0.33	0.10

Table 5 shows the annual transition matrix among the four states together with the ergodic probabilities of the states. Fig. 16 illustrates the persistence of the six states. It shows the expected value of the unemployment rate starting in each of the six states and evolving toward the ergodic distribution over a 10-year period. For example, the curve labeled 6 shows unemployment starting at the state-contingent level for state 6, which is over 8%. The rate falls quickly, dropping slightly below its ergodic value before converging back to that value. In the first few years, the dynamics of these impulse response functions differ, corresponding to the differences in the rows of the transition matrix. In later years, the paths are similar, because they are all controlled by the largest eigenvalue of the transition matrix.

The model starts in period 1 with initial values of TFP and the labor force both equal to one. In the base case, the distribution of the state in period 2 is the ergodic distribution. For 4 years, the transition matrix governs the succeeding states. In year 5, the economy has $6^4 = 1296$ possible configurations. For the next 10 years, the economy continues to evolve, but no further random events occur. The exogenous variables—TFP and the labor force—grow at constant rates equal to the average of the state-contingent rates, weighted by the ergodic distribution. The model has $1+6+36+216+1296+10\times 1296=14{,}515$ nodes, each with distinct values of all the variables of the model.

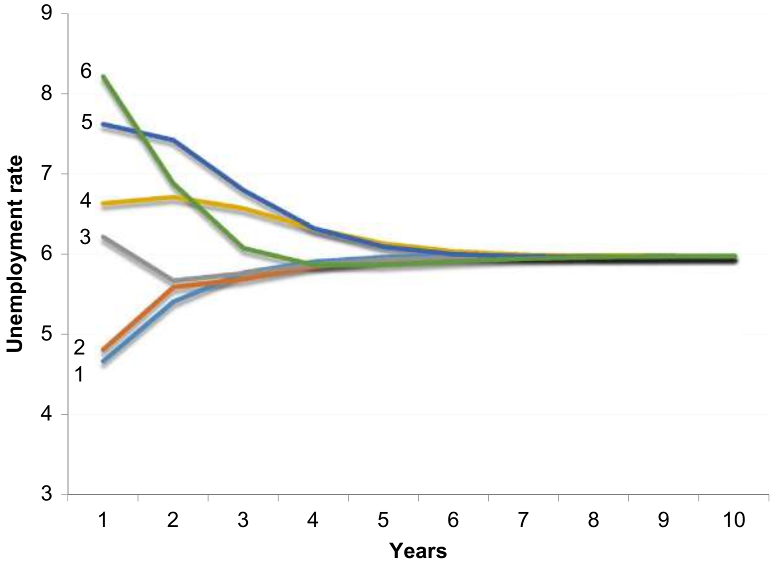


Fig. 16 Persistence of the states.

7.2 State-Based Driving Forces

Two of the variables used to define the states are also treated as driving forces in the model. These are the increments to TFP and the labor force. Another two driving forces are calculated from the data. These are the discount shock for capital, calculated as the residual in the pricing condition for capital,

$$d_k = \mathbb{E}[(1 + r_k')m'] - 1, \tag{37}$$

averaged over states to measure the expectation, and the discount shock for job creation, calculated as

$$d_n = -\left(\alpha + \frac{1}{\omega}\right)\log\frac{n}{\bar{n}}. \tag{38}$$

I also calculate the values of the discount shock for the safe 1-year interest rate, as

$$d_f = \mathbb{E}[(1 + r_f)m'] - 1, \tag{39}$$

but this value does not feed back into the rest of the model, so it is not a driving force, provided no bound on the rate is binding. Table 6 shows the state-contingent values of the driving forces.

TFP growth varies substantially across the economy's states. It is generally higher in the better, lower-numbered, states, but is highest in state 6. The reason is that most of the

Table 6 Values of the driving forces

	State-contingent value (%)				
State	TFP growth	Labor-force growth	Capital discount	Liquidity discount	Labor discount
1	2.00	1.68	15.01	−1.09	−0.81
2	1.80	1.79	14.93	−1.23	−0.72
3	0.43	1.48	13.39	−2.25	0.17
4	0.27	1.40	12.14	−0.90	0.44
5	0.92	0.52	12.98	−4.87	1.08
6	2.42	1.03	14.25	−3.05	1.46

years classified into state 6 are in the later years of slumps, when the economy is recovering. Historically, recoveries enjoyed high measured TFP growth because of improving utilization (recall that the model uses Fernald's TFP measure without his utilization adjustment). Labor-force growth, a driving force omitted from most models of fluctuations, also shows substantial variability across states, in a pattern similar to TFP. The capital discount is high, definitely in excess of almost all measures of the equity premium. The reason is that it includes factors that cause the return to capital to exceed the payout to owners that are not normally included in the equity premium. These include corporate taxes and agency frictions. The capital discount is higher in the favorable, lower-numbered states, again with the exception of state 6, so it is not much of a contributor to the business cycle. The table shows the calculated values of the liquidity discount, though it is not actually a driving force. The negative of the discount is a safety premium, associated with liquidity services and, in the bad states, a flight to safety. The most negative value of the discount is in the rare state 5 when the economy is in an unusually bad condition. That fact is important for the model's message about the conditions when the zero lower bound on the safe rate will matter. Finally, the labor discount, calculated from the unemployment rate, naturally tracks unemployment perfectly, because the other determinant of unemployment in the model, the product-market wedge, is taken to be the same in all states, for want of a reliable basis for computing it.

Two additional driving forces are present in the model, but do not have empirical counterparts. These are the product-market wedge, z, and the variable g, interpreted as a shift in product demand. The product-market wedge plays a central role in the new Keynesian model, but the measurement has proven controversial. Shifts in product demand resulting from tightening financial constraints on consumption have played a big role in understanding the financial crisis of 2008 and its aftermath, but again measurement of the shifts has proven controversial. The model tracks the effects of z and g, but its base case does not include their actual movements as driving forces in the economy. They both play important roles in the application of the model to the crisis of 2008 and the ensuing slump.

7.3 Parameters

Table 7 shows the parameter values used in the model. All are standard except for r_k^*, which is special to this framework, to ensure that the model's initial and terminal capital are close to its turnpike level of capital in relation to TFP and the size of the labor force.

7.4 Equilibrium

An equilibrium of the model is a complete set of values of the variables at every node. Fig. 17 provides some basic information about the equilibrium—it shows the means and standard deviations of the two exogenous variables, TFP and the size of the labor force, and two key endogenous variables, consumption and the unemployment rate, in each year. The distributions are conditional on the state of the economy in the first year. The standard deviations are calculated across the nodes for each year. Each should be interpreted as the standard deviation of the corresponding variable, conditional on the state of the economy in year 1, defined by the initial values of TFP, the labor force, and the capital stock. Because the capital stock is chosen to start the economy on its (stochastic) turnpike path, the subsequent values of the variables are distributed symmetrically around the path as time passes. Some of the variables grow and some have stable distributions around constant means. The upper left graph shows the distribution of TFP, A, which is close to a random walk. Its mean grows smoothly and its standard deviation fans out, rising approximately as the square root of the year number. The size of the labor force, L, shown in the upper right, behaves similarly, but its growth rate is somewhat higher and its conditional standard deviation is smaller. The variables in the lower part of Fig. 17 are not defined in period 1, but, again, the figures show the distributions conditional on the state of the economy in period 1. The conditional standard deviation of consumption, shown in the lower left, evolves by the same square-root principle as the ones for TFP and the labor force. The unemployment rate, shown in the lower right, has a stationary distribution along the turnpike path.

Table 7 Parameter values

Parameter	Interpretation	Value
α	Elasticity of output with respect to capital	0.35
δ	Depreciation rate of capital	0.1
β	Household discount ratio	0.95
σ	Intertemporal elasticity of substitution	0.5
κ	Capital adjustment parameter	2
u^*	Normal unemployment rate	0.0596
ω	Elasticity of employment function with respect to present value of a worker's contribution	4
rk^*	Effective discount rate for initial and terminal capital	0.3

Table 8 compares the volatility of some of the model's variables to the volatility of the corresponding data. In the case of variables that share the random-walk character of TFP and the labor force, the table describes rates of growth. The left column shows the standard deviations of the variables in the original annual data. The middle column shows the standard deviation, calculated using the model's ergodic distribution, of the state-contingent averages calculated from the original annual data. The right column shows

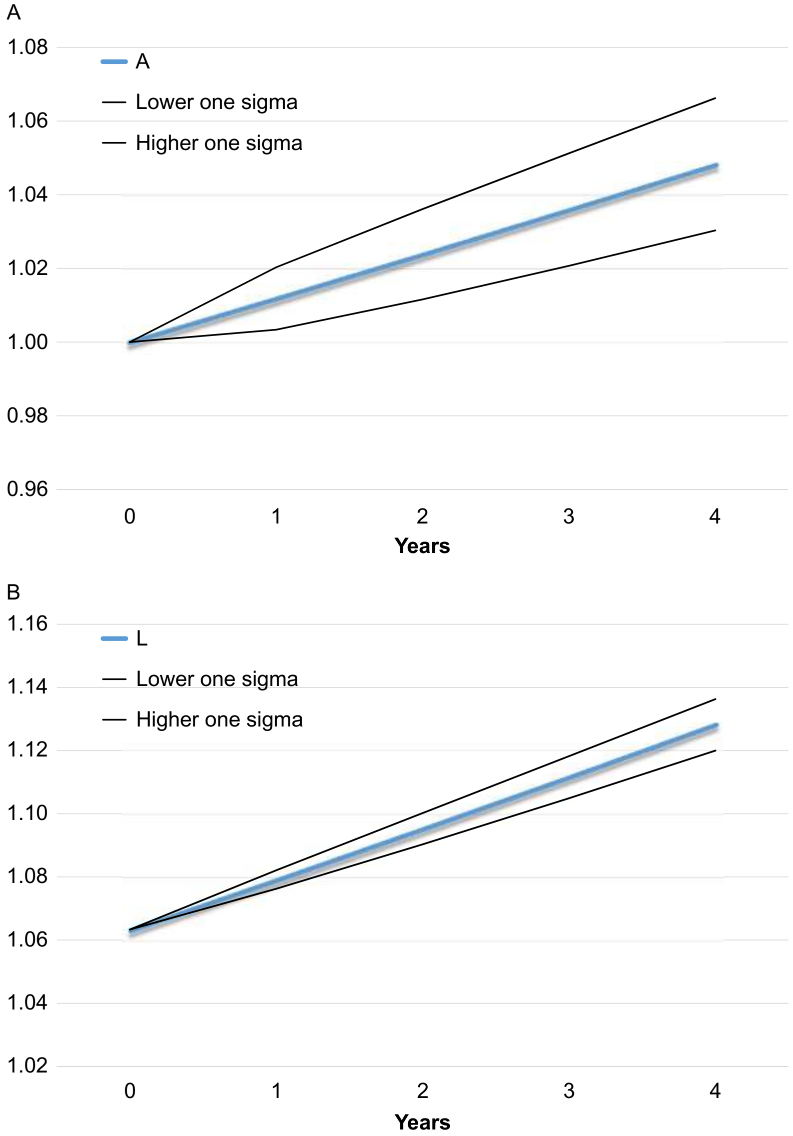


Fig. 17 Distributions of four variables in equilibrium. (A) TFP, (B) labor force,

(Continued)

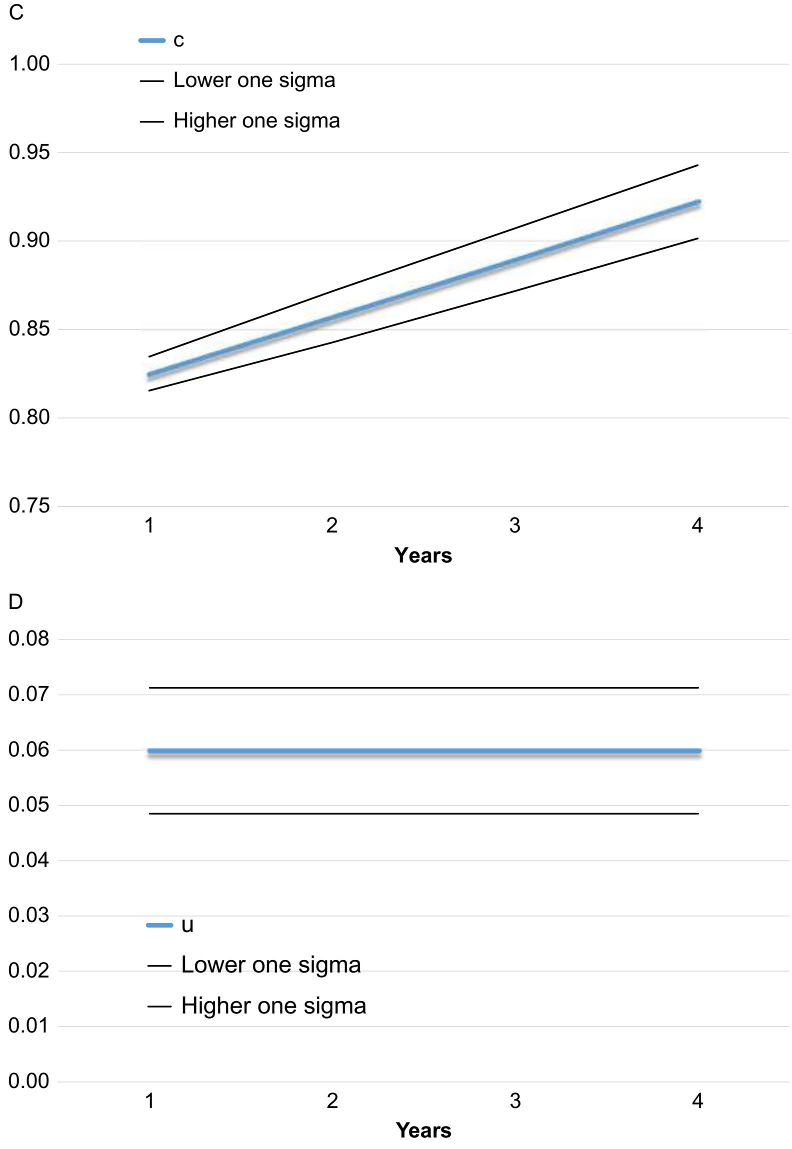


Fig. 17—Cont'd (C) consumption, and (D) unemployment rate.

the standard deviations in year 5 of the equilibrium. Comparison of the middle to the left column shows the success of the state setup in capturing the volatility of the corresponding variable. By necessity, the state setup falls short of full success. In most cases, the standard deviation across the states, weighted by the ergodic distribution, is around half of the actual standard deviation. Employment, unemployment, output, consumption,

Table 8 Standard deviations of selected variables in the data and in the model's equilibrium

	Standard deviation		
Variable	**Data**	**State-based data**	**Model**
TFP growth	1.65	0.83	0.84
Labor-force growth	0.81	0.27	0.27
Capital wedge	NA	1.42	1.42
Employment wedge	1.02	0.73	0.73
Output growth	2.19	1.34	1.18
Consumption growth	1.81	1.04	1.17
Investment growth	8.88	5.32	4.63
Return to capital	3.81	1.05	1.63
Unemployment	1.59	1.13	1.14

and investment do better than half, while labor-force growth and the return to capital fall short. Comparison of the right column to the middle column of Table 8 shows the success of the model in matching its target, the state-contingent values in the middle. For TFP growth, labor-force growth, the capital wedge, and the employment wedge, the match is perfect by construction. The match is reasonably good for the other variables.

7.5 Effects of the Driving Forces

The popular vector autoregression framework emphasizes *shocks* as the starting point for dynamic macro models. Shocks are uncorrelated with each other contemporaneously and uncorrelated with all lagged variables. See Ramey's chapter "Macroeconomic shocks and their propagation" in this handbook for a discussion of these assumptions. The framework of this chapter is different. Each year, a new value of the underlying state, s, occurs. Its probability distribution is known from the transition probabilities of the Markov process, but the realization from that distribution is a shock. The realization determines the new values of the driving forces. These movements are mutually correlated. Because the model incorporates the hypothesis of rational expectations, adjusted by the known state-dependent distortions, the model incorporates the notion that rational actors respond to the surprise elements of current realizations.

In this framework, it is interesting but challenging to answer questions about the separate effects of the driving forces. Because those forces are correlated, the variance decomposition often presented along with a VAR model is not available—potentially large components of the variance of a given endogenous variable arise from the covariance of a pair of driving forces, so their distinct contributions are not defined. The position of the VAR modeler, as Ramey explains, is that shocks must be uncorrelated, because otherwise they would not have distinct contributions. The approach in this chapter is that driving forces are fundamental and that their correlation is a matter of measurement, not assumption.

One way to understand the roles of the driving forces is to consider a set of counterfactual economies, each with only one driving force. Table 9 shows the results of that exercise. The top row shows the standard deviations of annual output growth for the base case, with all four driving forces in action, and for the four counterfactuals, with single driving forces. Table 10 shows the correlation matrix of the driving forces, based on the state-contingent values, using the ergodic probabilities (the one-period-ahead correlation matrix is state dependent). Two correlations stand in the way of even an approximate allocation of explanatory role: The capital wedge is correlated 0.83 with TFP growth and the labor wedge is correlated −0.89 with the labor-force growth.

Table 9 suggests that all four driving forces have important roles in economic fluctuations. An economy with only TFP fluctuations has substantial fluctuations in all of the variables except unemployment. An economy with only labor-force fluctuations has moderate volatility of investment growth—but recall that this driving force is not well captured by the states of the model, so this finding probably understates the importance of labor-force fluctuations. An economy with only a capital wedge has some volatility of consumption, quite a bit of volatility of the return to capital, and a lot of volatility of investment. An economy with only a labor wedge has substantial volatility of all the variables.

In addition to the ambiguities associated with the correlation among the four observed driving forces, the results in Table 9 need to be interpreted in the light of

Table 9 Standard deviations of selected variables in counterfactual economies with single driving forces

	All driving forces	TFP growth only	Labor-force growth only	Capital wedge only	Labor wedge only
Output growth	1.18	0.84	0.15	0.09	0.73
Consumption growth	1.17	0.80	0.12	0.44	0.41
Investment growth	4.63	1.18	0.71	2.64	3.26
Return to capital	1.63	0.62	0.26	0.84	1.03
Unemployment	1.14	0.00	0.00	0.00	1.14

Table 10 Correlations of driving forces

	TFP growth	Labor-force growth	Capital wedge	Labor wedge
TFP growth	1.00			
Labor-force growth	0.15	1.00		
Capital wedge	0.83	−0.03	1.00	
Labor wedge	−0.28	−0.89	−0.18	1.00

the inability to measure other driving forces, notably fluctuations in product demand. The large role of the labor wedge in the table may actually reflect effects operating through shifts of consumption and investment from forces not included in the model. A later section of this chapter on the forces unleashed by the 2008 crisis shows the potential importance of the product demand and product-market-wedge driving forces.

The model takes a simplified view of the role of confidence, ambiguity aversion, and other factors that may discourage economic activity in ways not included in traditional macro models. Both the capital wedge and the labor wedge are modeled as extra discounts that have adverse effects, but the labor wedge appears to be much the more important of the two. In the model, a decline in confidence and the corresponding increase in the labor discount d_n have a direct effect on job creation through the mechanisms associated with the DMP model. Lower job creation results in lower job-finding rates and higher unemployment. The result enters the rest of the economy as an adverse shift in net labor supply resulting in declines in output, shared between consumption and investment. As Table 9 shows, in the base model, there is no effect on unemployment from other driving forces—the rise in unemployment in bad times is entirely assigned to a decline in confidence among businesses that cuts back their job-creation flows. Obviously this property is an oversimplification, but the macro-labor research community has made more progress recently in demonstrating the near-irrelevance of driving forces of unemployment such as productivity than in finding driving forces to account for fluctuations in unemployment as responses to other forces. The later section on the crisis shows how the product-market wedge influences unemployment.

8. CRISIS AND SLUMP

This section explores the model's properties when the driving forces are tuned to data from the years 2009 through 2012, the years of the maximum effects of the crisis of late 2008 and its aftermath. This exercise assigns those 4 years to an altered state 5 with more negative effects, including values of the two driving forces not measured for the base model covering all the years starting in 1953. Table 11 shows the values for the six driving

Table 11 Values of driving forces hypothesized for crisis slump

Driving force	Value in state 5
TFP growth	0.92
Labor-force growth	0.10
Capital discount	16.70
Liquidity discount	−6.00
Labor discount	1.96
Product-market wedge	3.00
Product demand	−5.00

forces. TFP growth retains its value from the base case, which was close to actual growth over 4 years. Labor-force growth is much lower than normal, just above zero. The capital discount is well above its actual value in any state in the base case, reflecting the belief that agency frictions and a loss of confidence occurred during the immediate post-crisis years. The liquidity discount for the safe 1-year interest rate is lower than in any state in the base case, reflecting an unusual flight to safety after the crisis. The labor discount is 0.4 percentage points higher than in state 5 in the base case, corresponding to an unemployment rate (with no product-market wedge) around 9%, that actually occurred after the crisis. The product-market wedge is taken at the hypothetical value of 3% and the product demand shift at minus 5%.

Table 12 shows the average effects of the driving forces over 4 years of adverse shocks, in comparison to an economy that stayed all 4 years in a different version of state 5 in which the driving forces all had the average of their values from the base case. Thus the figures in the table are the effects of the crisis in the sense of the differences in the outcomes between an economy with the special crisis driving forces and one with driving forces typical of the U.S. economy historically in normal times. The left column shows the average with all the crisis-specific driving forces in action. The rise of 4.54 percentage points of unemployment resembles the actual behavior of the economy. The decline in output is substantial but falls short of the actual decline of about 10%. But the *positive* numbers for consumption and investment are dramatically the opposite of the actual sharp decline in consumption and collapse of investment. This result is not a failure of the model, but rather a consequence of the model's implication of a huge decline in the safe interest rate. This decline could not have occurred, because of the zero lower bound. The story of the table is that the decline in the interest rate unhindered by the lower bound would have brought about an increase in interest-sensitive consumption and investment that would more than offset the direct decline in the spending shift g and the adverse effects of other driving forces.

Table 12 Effects of crisis shocks on key variables, averaged over 4 years

	Driving force						
Variable	**All**	**Capital discount**	**Labor discount**	**Safe rate discount**	**Labor force**	**Product-market wedge**	**Spending shift**
Unemployment, percentage points	4.54	0.00	2.98	0.00	0.00	1.61	0.00
Consumption, %	0.23	0.32	−1.33	0.00	−1.72	−0.71	2.51
Output, %	−3.30	−0.26	−1.84	0.00	−1.30	−1.00	0.53
Investment, %	1.47	−0.57	−0.51	0.00	0.43	−0.28	1.77
Safe interest rate, percentage points	−11.37	−1.48	−0.33	−4.48	−0.88	−0.17	−3.20

The right panel of Table 12 breaks down the effects by the driving forces. Because the model is nonlinear, the sum of the effects on the right side is slightly different from the combined effect on the left. The increase in the capital discount had no effect on unemployment, moved a small amount of spending from investment to consumption, lowered output modestly, and depressed the safe interest rate. The rise in the labor discount raised unemployment substantially and cut output by 1.84%, 1.33% of output from consumption and 0.51% from investment. The rise in the liquidity-safety premium for the short rate had an effect only on that rate, as there is no direct feedback from changes in that rate induced by changes in the premium in the model. The adverse effect of the crisis on the labor force cut output by 1.30%. Consumption fell by 1.72% of normal output, while investment rose by 0.43%. The rise in the product-market wedge accounted for 1.61 percentage points of the rise in unemployment, by raising market power and lowering the marginal revenue product of labor and thus cutting the incentive to create jobs. The spending shift, modeled as a decline in government purchases, resulted in increases of 2.51% of output in consumption and 1.77% in investment, thanks to the income effect of lower implied taxes and the induced decline in the safe short rate of 3.20 percentage points.

8.1 The Zero Lower Bound

Obviously the main lesson of Table 12 is the central importance of the zero lower bound for the severity of the post-crisis slump in the U.S. economy. Although the model does not implement a lower bound on the safe real rate, the results are informative about the incidence of the bound and, to some extent, about the magnitude of adverse effects that would have resulted from the bound. During the slump following the 2008 crisis, the short safe nominal rate was essentially zero, at its bound as perceived by the Federal Reserve, and the expected inflation rate was around two percent—see Fig. 15—so the corresponding bound on the real rate is around minus 2%.

In the model, the normal value of the safe real rate during the years after 2008 is about 3%. According to the lower left figure in Table 12, with all driving forces active at the levels in Table 11, the rate would have been about 11% lower, or minus 8%. Most macroeconomists would probably agree that the effects of a monetary force that raised the safe real rate 8 percentage points above its equilibrium would be severely contractionary. More than half of that is the direct result of the depression of the safe rate on account of the flight to safety hypothesized in the crisis-slump scenario. The other big negative force is the downward shift in product demand, shown in the lower right corner of Table 12. The model supports the idea that the collapse of house prices and tightening of bank lending battered the economy by discouraging consumption and investment. The third-biggest contributor to the decline in the safe rate was the capital discount, good for about 1.5 percentage points of decline. The labor discount, on the other hand, had only a small effect—it is a supply effect. Whereas demand effects are more than fully offset by the decline in the safe rate, reductions in supply cannot be offset that way.

9. PERSISTENCE

Effects lasting longer than the driving forces themselves operate through the model's state variables. It has two exogenous state variables, TFP and the labor force, and one endogenous state variable, the capital stock.

9.1 Exogenous Persistence

In the model, each shock to the labor force has permanent effects. Where the shocks operate through births and immigration, this property is a reasonable approximation. Whether the substantial decline in the labor-force participation of existing individuals that occurred during the slump will ultimately reverse itself is an open question. As of early 2016, there was no sign that the return to essentially normal conditions in the labor market would result in a restoration of any of the large decline that accompanied the slump. Fig. 18 shows the path of the labor force as a percent of its initial normal value in the hypothetical crisis slump studied in the previous section. With four consecutive large incremental shortfalls in the labor force during years 1–4, the cumulative shortfall in the labor force in year 4 is about 6%. Though the labor-force growth rate returns to normal in year 5, the cumulative shortfall continues to become larger, because the growth process is multiplicative and is always at a lower base, post-crisis.

9.2 Endogenous Persistence

Endogenous persistence occurs through the capital stock. The effect of the capital discount is concentrated on investment, as shown in Table 12. An increase in that discount

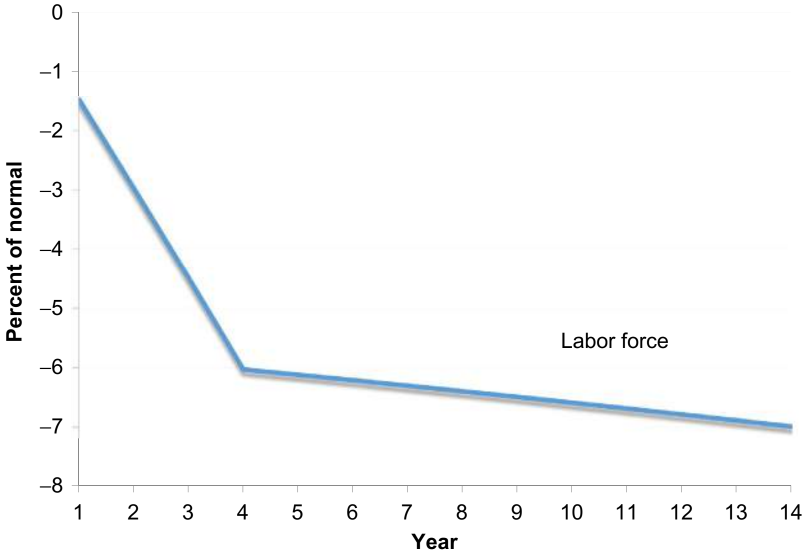


Fig. 18 Persistence of a labor-force shock.

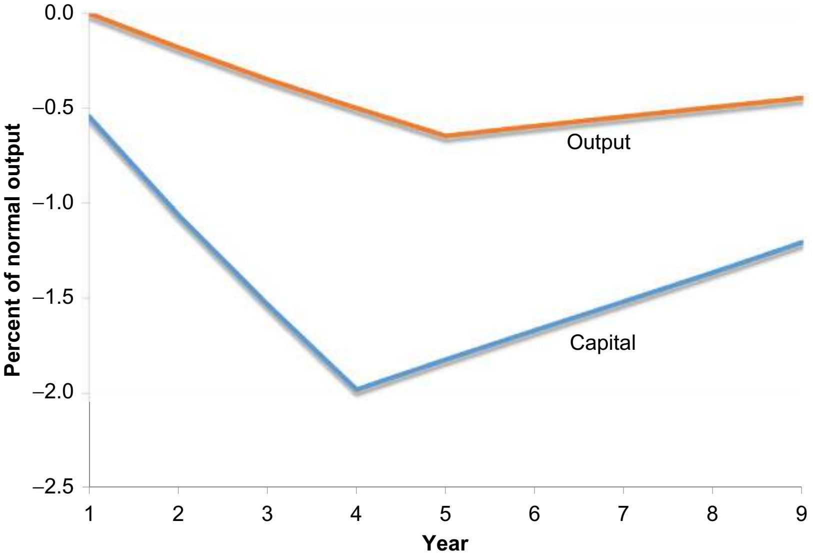


Fig. 19 Persistent effects of an elevation of the capital discount.

causes businesses to place a lower value on the future payoff to capital formation, so capital falls further and further below its normal growth path during a period of higher discount. The effects on output and other variables persist beyond the time when the discount declines back to normal. The capital stock returns only gradually to its normal growth path. Fig. 19 shows the effects of the 4-year period of increased capital discount described in Table 11 on the capital stock and on output. The figure shows the difference between the expected values of those variables conditional on the crisis values of the capital discount and the expected values with normal, noncrisis values of the discount. The effects on both variables cumulate during the 4 years with the higher crisis discount and then begin to return toward zero. Five years after the end of the crisis values, the effects remain strong.

Similar results apply to the other driving forces that have negative effects on investment in Table 11. These are the labor discount, which cuts investment by reducing the effective supply of labor, and the product-market wedge, which lowers the marginal revenue product of capital.

10. CONCLUDING REMARKS

This chapter is complementary to Ramey's chapter in this volume. Most of her discussion relates to empirical evidence from VARs and other econometric specifications, or to the properties of new Keynesian structural models, though she does also consider structural models more closely affiliated with the tradition of the real business-cycle model. She focuses extensively on monetary shocks—departures of monetary policy from its usual relation to current developments in the economy. No monetary shocks occur in the economy considered in this chapter. The central bank never pushes the short rate away from its equilibrium value to restore inflation to its target rate. In the context of the literature that includes monetary policy and monetary nonneutrality, the model here reveals the values of the interest rate and other variables that would prevail in the absence of sticky prices and wages. Both chapters consider government purchases as a driving force. In the empirical work Ramey considers, the focus is on the purchases multiplier, as revealed by the empirical relation between output and purchases. She finds that the multiplier is around one but with considerable dispersion across studies. In this chapter, Table 12 shows a multiplier of 0.53, the value in the row for output and the column for the spending shift. The lower value may be the result of the model's assumption of full monetary response to government purchases, letting the interest rate track the change in its equilibrium value. The sample period for the model includes times when, for example, monetary policy kept the interest rate constant in the face of an increase in purchases, which would considerably amplify the response of output. On the tax side of fiscal policy, Ramey considers taxes as explicit driving forces. Taxes have a role in this chapter because they are one of the sources of historical shifts in the capital discount. But I do not consider tax changes as special driving forces of the post-crisis slump. Ramey's chapter includes a detailed treatment of the measurement of technology shifts and their effects, as measured in empirical work. To measure TFP growth, she concludes in favor of measures with utilization adjustments. This chapter uses Fernald's measure without that adjustment. She also discusses, in detail, measures of technological change apart from TFP, relating to investment. She briefly mentions oil-price changes, credit conditions, policy uncertainty, fluctuations in the labor force, and the labor wedge as additional driving forces. She does not mention the product-market wedge as a driving force.

The importance of total factor productivity as a determinant of medium-term growth and economic performance is widely agreed among macroeconomists today, and is confirmed in the results of this chapter—Table 9 shows that, historically, movements of TFP by themselves would account for a standard deviation of output around 75% of the total of all driving forces. A decline in productivity growth occurred during the slump that began in 2008 and contributed to the shortfall in output, consumption, and capital formation during the slump. Whether the crisis of 2008 contributed to the decline in productivity growth is unresolved.

On the other hand, fluctuations in the size of the work force relative to the working-age population—the labor-force participation rate—are about as big as fluctuations in productivity and have similar effects. Research on medium-run fluctuations has neglected this driving force, even though research on participation itself has been extensive. The continuation during the recent slump of a major decline in participation that began in 2000—and is not the result of demographic shifts—worsened the slump. The evidence seems to point in the direction that the decline in participation was not the result of the crisis and resulting explosion of unemployment.

Evidence from financial markets appears to confirm the proposition that discounts applied to risky investments rose as a result of the crisis even as the safe rate fell to zero. In normal times, without the zero lower bound, higher discounts result in lower output and employment. There is an interesting unresolved question about the role of discount increases when the real rate is held fixed by the zero lower bound on the nominal interest rate and the immovability of inflation.

Models that attribute some of the depth and persistence of the response of the economy to financial shocks hold that the shocks cause increases in agency frictions within financial intermediaries or nonfinancial businesses. Financial wedges develop to ensure that managers deprived of equity still have continuation values sufficient to prevent misconduct. The evidence of widening wedges between the return to capital and the safe short rate is convincing, as is the sharp but transitory rise in the spreads between risky private bonds and Treasurys of the same maturity. The model in this chapter assigns a moderate but important role to financial frictions, as part of the driving force called the capital discount.

The new Keynesian model has called attention to the product-market wedge—the markup of price over cost—as the transmission mechanism of shocks to economic activity. With sticky prices, an increase in demand raises cost but not price, so the markup declines. The economy expands because the product-market wedge functions like a tax wedge in depressing activity and the decline in markups relieves that adverse effect. An interesting debate has yet to resolve the issue of the importance of the product-market wedge in the depth and persistence of slumps.

Finally, the model confirms earlier findings about the multiplier effects of shifts in product demand. As an important cause of declining consumption demand, household deleveraging has been assigned a major role in the recent slump and is an obvious candidate for explaining the persistence of the slump. In the model, an exogenous decline in product demand results in a large decrease in the interest rate, which stimulates consumption and investment. Rather than collapsing, the economy undergoes a large reallocation of resources. But with the zero lower bound in effect, the reallocation fails to occur. Instead, output and employment fall. As yet, the profession has not come forth with a well-founded model of that failure.

ACKNOWLEDGMENTS

Chapter prepared for the *Handbook of Macroeconomics*, John Taylor and Harald Uhlig, eds. The Hoover Institution supported this research. The research is also part of the National Bureau of Economic Research's Economic Fluctuations and Growth Program. I thank Gauti Eggertsson, Wouter den Haan, and Cosmin Ilut for illuminating conference discussions, and Valerie Ramey and the editors for comments. Complete backup for all of the calculations is available from my website, stanford.edu/~rehall

REFERENCES

Adrian, T., Colla, P., Shin, H.S., 2012. Which financial frictions? Parsing the evidence from the financial crisis of 2007–9. Working Paper 18335. National Bureau of Economic Research. http://www.nber.org/papers/w18335.

Aguiar, M., Hurst, E., Karabarbounis, L., 2013. Time use during the great recession. Am. Econ. Rev. 103, 1664–1696.

Angeletos, G.-M., Collard, F., Dellas, H., 2014. Quantifying confidence. Working Paper 20807. National Bureau of Economic Research. http://www.nber.org/papers/w20807.

Attanasio, O.P., Weber, G., 1995. Is consumption growth consistent with intertemporal optimization? Evidence from the consumer expenditure survey. J. Polit. Econ. 103 (6), 1121–1157.

Autor, D.H., 2011. The unsustainable rise of the disability rolls in the united states: causes, consequences, and policy options. Working Paper 17697, National Bureau of Economic Research. http://www.nber.org/papers/w17697.

Ball, L., Mazumder, S., 2011. The evolution of inflation dynamics and the great recession. Brookings Papers Econ. Act. (1), 337–405.

Barnichon, R., Figura, A., 2012. The determinants of the cycles and trends in U.S. Unemployment. Federal Reserve Board.

Barro, R.J., Redlick, C.J., 2011. Macroeconomic effects from government purchases and taxes. Q. J. Econ. 126 (1), 51–102. http://dx.doi.org/10.1093/qje/qjq002. http://qje.oxfordjournals.org/content/126/1/51.full.pdf+html, http://qje.oxfordjournals.org/content/126/1/51.abstract.

Benigno, G., Fornaro, L., 2015. Stagnation Traps. London School of Economics, London.

Bernanke, B.S., Gertler, M., Gilchrist, S., 1999. The financial accelerator in a quantitative business cycle framework. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics. Elsevier, North Holland, pp. 1341–1393 (Chapter 21).

Bhutta, N., 2012. Mortgage debt and household deleveraging: accounting for the decline in mortgage debt using consumer credit record data. Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington, DC.

Bianchi, F., Ilut, C., Schneider, M., 2012. Uncertainty Shocks, Asset Supply and Pricing over the Business Cycle. Duke University, Department of Economics, Durham, NC.

Bils, M., Klenow, P.J., Malin, B.A., 2014. Resurrecting the role of the product market wedge in recessions. Working Paper 20555, National Bureau of Economic Research. http://www.nber.org/papers/w20555.

Blundell, R., Pistaferri, L., Preston, I., 2008. Consumption inequality and partial insurance. Am. Econ. Rev. 98 (5), 1887–1921.

Brunnermeier, M.K., Sannikov, Y., 2014. A macroeconomic model with a financial sector. Am. Econ. Rev. 104 (2), 379–421. http://dx.doi.org/10.1257/aer.104.2.379.

Brunnermeier, M.K., Eisenbach, T.M., Sannikov, Y., 2012. Macroeconomics with financial frictions: a survey. Working Paper 18102, National Bureau of Economic Research. http://www.nber.org/papers/w18102.

Campbell, J.Y., Shiller, R.J., 1988. The dividend-price ratio and expectations of future dividends and discount factors. Rev. Finan. Stud. 1 (3), 195–228. ISSN 0893-9454. http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=EP4233837&site=ehost-live&scope=site.

Chari, V.V., Kehoe, P.J., McGrattan, E.R., 2007. Business cycle accounting. Econometrica 75 (3), 781–836. http://ideas.repec.org/a/ecm/emetrp/v75y2007i3p781-836.html.

Chodorow-Reich, G., 2014. The employment effects of credit market disruptions: firm-level evidence from the 2008–9 financial crisis. Q. J. Econ. 129 (1), 1–59. http://dx.doi.org/10.1093/qje/qjt031. http://qje.oxfordjournals.org/content/129/1/1.full.pdf+html. http://qje.oxfordjournals.org/content/129/1/1.abstract.

Chodorow-Reich, G., Karabarbounis, L., 2015. The cyclicality of the opportunity cost of employment. J. Polit. Econ. Working Paper 19678, forthcoming. http://www.nber.org/papers/w19678.

Christiano, L.J., Trabandt, M., Walentin, K., 2010. DSGE models for monetary policy analysis. Working Paper 16074. National Bureau of Economic Research. http://www.nber.org/papers/w16074.

Christiano, L.J., Eichenbaum, M., Rebelo, S., 2011. When is the government spending multiplier large? J. Polit. Econ. 119 (1), 78–121.

Christiano, L.J., Eichenbaum, M.S., Trabandt, M., 2016. Unemployment and business cycles. Econometrica, forthcoming.

Cochrane, J.H., 2011. Presidential address: discount rates. J. Finan. 66 (4), 1047–1108. ISSN 0022-1082. http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=1254835&site=ehost-live&scope=site.

Cochrane, J.H., 2014. The New-Keynesian Liquidity Trap. Booth School of Business, University of Chicago, Chicago.

Coenen, G., Erceg, C.J., Freedman, C., Furceri, D., Kumhof, M., Lalonde, R., Laxton, D., Lind, J., Mourougane, A., Muir, D., Mursula, S., de Resende, C., Roberts, J., Roeger, W., Snudden, S., Trabandt, M., in't Veld, J., 2012. Effects of fiscal stimulus in structural models. Am. Econ. J. Macroecon. 4 (1), 22–68. http://dx.doi.org/10.1257/mac.4.1.22.

Cole, H., Rogerson, R., 1999. Can the Mortensen-Pissarides matching model match the business cycle facts? Int. Econ. Rev. 40 (4), 933–960.

Comin, D., Gertler, M., 2006. Medium-term business cycles. Am. Econ. Rev. 96 (3), 523–551. http://dx.doi.org/10.1257/aer.96.3.523.

Correia, I., Farhi, E., Nicolini, J.P., Teles, P., 2010. Policy at the Zero Bound. Banco de Portugal.

Cúrdia, V., Woodford, M., 2015. Credit frictions and optimal monetary policy. Working Paper 21820, National Bureau of Economic Research. http://www.nber.org/papers/w21820.

Daly, M., Hobijn, B., Şahin, A., Valletta, R., 2011a. A rising natural rate of unemployment: transitory or permanent? Working Paper 2011-05, Federal Reserve Bank of San Francisco.

Daly, M.C., Hobijn, B., Valletta, R.G., 2011b. The recent evolution of the natural rate of unemployment. IZA Discussion Paper No. 5832.

Daly, M.C., Hobijn, B., Şahin, A., Valletta, R.G., 2012. A search and matching approach to labor markets: did the natural rate of unemployment rise. J. Econ. Perspect. 26 (3), 3–26.

Davis, S.J., Haltiwanger, J., 2014. Labor market fluidity and economic performance. Proceedings of the Jackson Hole Symposium, Federal Reserve Bank of Kansas, pp. 17–107.

Davis, S.J., von Wachter, T., 2011. Recessions and the costs of job loss. Brookings Papers Econ. Act. (2), 1–55.

Davis, S.J., Faberman, R.J., Haltiwanger, J.C., 2012. Recruiting intensity during and after the great recession: national and industry evidence. Am. Econ. Rev. Papers Proc. 102 (3), 584–588. http://dx.doi.org/10.1257/aer.102.3.584.

De Nardi, M., French, E., Benson, D., 2011. Consumption and the great recession. Working Paper, National Bureau of Economic Research. http://www.nber.org/papers/w17688.

Dynan, K., 2012. Is a household debt overhang holding back consumption? Brookings Papers Econ. Act. Spring, 299–362. ISSN 0007-2303. http://www.jstor.org/stable/23287219.

Eckstein, Z., Setty, O., Weiss, D., 2015. Financial Risk and Unemployment. Tel Aviv University, Tel Aviv, Israel.

Eggertsson, G.B., Krugman, P., 2012. Debt, deleveraging, and the liquidity trap: a Fisher-Minsky-Koo approach. Q. J. Econ. 127 (3), 1469–1513. ISSN 0033-5533. http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=1323598&site=ehost-live&scope=site.

Eggertsson, G.B., Mehrotra, N.R., 2014. A model of secular stagnation. Working Paper 20574, National Bureau of Economic Research. http://www.nber.org/papers/w20574.

Elsby, M.W., Hobijn, B., Şahin, A., Valletta, R.G., 2011. The labor market in the great recession: an update to September 2011. Brookings Papers Econ. Act. Fall, 353–384. ISSN 0007-2303. http://www.jstor.org/stable/41473602.

Elsby, M.W., Hobijn, B., Şahin, A., 2013. On the importance of the participation margin for market fluctuations. Working Paper 2013-05, Federal Reserve Bank of San Francisco.

Estevão, M., Tsounta, E., 2011. Has the Great Recession Raised U.S. Structural Unemployment? International Monetary Fund.

Farber, H.S., 2015. Job loss in the great recession and its aftermath: U.S. evidence from the displaced workers survey. Working Paper 21216, National Bureau of Economic Research. http://www.nber.org/papers/w21216.

Farber, H.S., Valletta, R.G., 2013. Do extended unemployment benefits lengthen unemployment spells? Evidence from recent cycles in the U.S. labor market. Working Paper 19048, National Bureau of Economic Research.

Farhi, E., Werning, I., 2013. A Theory of Macroprudential Policies in the Presence of Nominal Rigidities. Department of Economics, Harvard University, Cambridge, MA.

Farmer, R.E., 2012. The stock market crash of 2008 caused the great recession: theory and evidence. J. Econ. Dyn. Control 36 (5), 693–707. ISSN 0165-1889, http://dx.doi.org/10.1016/j.jedc.2012.02.003. http://www.sciencedirect.com/science/article/pii/S0165188912000401.

Fernald, J.G., 2012. A quarterly, utilization-adjusted series on total factor productivity. 2012–19, Federal Reserve Bank of San Francisco. Updated regularly on Fernald's FRBSF website.

Fernald, J., 2014. Productivity and potential output before, during, and after the great recession. NBER Macroecon. Annu. 29, 1–51.

Fleckenstein, M., Longstaff, F.A., Lustig, H., 2013. Deflation Risk. Anderson School, UCLA, Los Angeles.

Friedman, M., 1968. Presidential address. Am. Econ. Rev. 58 (1), 1–15.

Fujita, S., 2011. Effects of extended unemployment insurance benefits: evidence from the monthly CPS. Federal Reserve Bank of Philadelphia.

Fujita, S., Moscarini, G., 2013. Recall and unemployment. Working Paper 19640, National Bureau of Economic Research. http://www.nber.org/papers/w19640.

Gertler, M., Karadi, P., 2011. A model of unconventional monetary policy. J. Monetary Econ. 58 (1), 17–34. ISSN 0304-3932. http://dx.doi.org/10.1016/j.jmoneco.2010.10.004. Carnegie-Rochester Conference Series on Public Policy: The Future of Central Banking April 16–17, 2010. http://www.sciencedirect.com/science/article/pii/S0304393210001261.

Gertler, M., Kiyotaki, N., 2011. Financial intermediation and credit policy in business cycle analysis. In: Friedman, B., Woodford, M. (Eds.), Handbook of Monetary Economics. Elsevier, North Holland, pp. 547–599.

Gertler, M., Sala, L., Trigari, A., 2008. An estimated monetary DSGE model with unemployment and staggered nominal wage bargaining. J. Money Credit Bank. 40 (8), 1713–1764. ISSN 1538-4616. http://dx.doi.org/10.1111/j.1538-4616.2008.00180.x.

Gilchrist, S., Schoenle, R., Sim, J.W., Zakrašek, E., 2014a. Inflation Dynamics During the Financial Crisis. Department of Economics, Boston University, Boston.

Gilchrist, S., Sim, J.W., Zakrajek, E., 2014b. Uncertainty, financial frictions, and investment dynamics. National Bureau of Economic Research. Working Paper 20038, http://www.nber.org/papers/w20038.

Gomme, P., Ravikumar, B., Rupert, P., 2011. The return to capital and the business cycle. Rev. Econ. Dyn. 14 (2), 262–278. ISSN 1094-2025. http://dx.doi.org/10.1016/j.red.2010.11.004. http://www.sciencedirect.com/science/article/pii/S1094202510000591.

Gourio, F., 2012. Disaster risk and business cycles. Am. Econ. Rev. 102 (6), 2734–2766. http://dx.doi.org/10.1257/aer.102.6.2734.

Gourio, F., Rudanko, L., 2014. Customer capital. Rev. Econ. Stud. 81 (3), 1102–1136. http://dx.doi.org/10.1093/restud/rdu007. http://restud.oxfordjournals.org/content/81/3/1102.abstract.

Hagedorn, M., Karahan, F., Manovskii, I., Mitman, K., 2013. Unemployment benefits and unemployment in the great recession: the role of macro effects. Working Paper 19499. National Bureau of Economic Research. http://www.nber.org/papers/w19499.

Hall, R.E., 1995. Lost jobs. Brookings Papers Econ. Act. (1), 221–273.

Hall, R.E., 2004. Measuring factor adjustment costs. Q. J. Econ. 119 (3), 899–927.

Hall, R.E., 2005. Separating the business cycle from other economic fluctuations. The Greenspan era: lessons for the Future, Proceedings of the Federal Reserve Bank of Kansas City, 133–179.

Hall, R.E., 2009. By how much does GDP rise if the government buys more output? Brookings Papers Econ. Act. (2), 183–231.

Hall, R.E., 2011a. The high sensitivity of economic activity to financial frictions. Econ. J. 121, 351–378.

Hall, R.E., 2011b. The long slump. Am. Econ. Rev. 101 (2), 431–469. http://dx.doi.org/10.1257/aer.101.2.431. 2011 AEA Presidential Address.

Hall, R.E., 2012. How the financial crisis caused persistent unemployment. In: Wright, I.J., Ohanian, L.E., Taylor, J.B. (Eds.), Government Policies and the Delayed Economic Recovery. Hoover Institution Press, Stanford, pp. 57–83.

Hall, R.E., 2013. The routes into and out of the zero lower bound. Proceedings of the Jackson Hole Symposium, Federal Reserve Bank of Kansas City, pp. 1–35.

Hall, R.E., 2014. Quantifying the lasting harm to the U.S. economy from the financial crisis. NBER Macroecon. Annu. 29 (1), 71–128.

Hall, R.E., 2015. High Discounts and High Unemployment. Hoover Institution, Stanford University, Stanford.

Hall, R.E., 2016. Search-and-matching analysis of high unemployment caused by the zero lower bound. Rev. Econ. Dyn. 19, 210–217.

Hall, R.E., Schulhofer-Wohl, S., 2015. Measuring job-finding rates and matching efficiency with heterogeneous jobseekers. Working Paper 20939, National Bureau of Economic Research. http://www.nber.org/papers/w20939.

He, Z., Krishnamurthy, A., 2013. Intermediary asset pricing. Am. Econ. Rev. 103 (2), 732–770. http://dx.doi.org/10.1257/aer.103.2.732.

He, Z., Krishnamurthy, A., 2015. A Macroeconomic Framework for Quantifying Systemic Risk. Graduate School of Business, Stanford University, Stanford.

Herz, B., van Rens, T., 2011. Structural Unemployment. Universitat Pompeu Fabra, Barcelona.

Hyatt, H.R., Spletzer, J.R., 2013. The Recent Decline in Employment Dynamics. Center for Economic Studies, US Census Bureau, Washington, DC.

Jarosch, G., 2014. Searching for Job Security and the Consequences of Job Loss. Department of Economics, University of Chicago, Chicago.

Kaplan, G., Menzio, G., 2016. Shopping externalities and self-fulfilling unemployment fluctuations. J. Polit. Econ. forthcoming.

Kaplan, G., Violante, G.L., 2014. A model of the consumption response to fiscal stimulus payments. Econometrica 82 (4), 1199–1239. ISSN 1468-0262. http://dx.doi.org/10.3982/ECTA10528.

Karabarbounis, L., Neiman, B., 2014. The global decline of the labor share. Q. J. Econ. 129 (1), 61–103. http://dx.doi.org/10.1093/qje/qjt032. http://qje.oxfordjournals.org/content/129/1/61.abstract.

Kiyotaki, N., Moore, J., 2012. Liquidity, business cycles, and monetary policy. Working Paper 17934, National Bureau of Economic Research. http://www.nber.org/papers/w17934.

Kocherlakota, N.R., 2013. Impact of a land price fall when labor markets are incomplete. Federal Reserve Bank of Minneapolis.

Korinek, A., Simsek, A., 2014. Liquidity trap and excessive leverage. Working Paper 19970, National Bureau of Economic Research. http://www.nber.org/papers/w19970.

Kozlowski, J., Veldkamp, L., Venkateswaran, V., 2015. The tail that wags the economy: belief-driven business cycles and persistent stagnation. Working Paper 21719, National Bureau of Economic Research. http://www.nber.org/papers/w21719.

Krishnamurthy, A., Vissing-Jorgensen, A., 2013. Short-Term Debt and Financial Crises: What We Can Learn from U.S. Treasury Supply. Kellogg School, Northwestern University, Chicago.

Krueger, A.B., Mueller, A.I., 2011. Job search, emotional well-being, and job finding in a period of mass unemployment: evidence from high-frequency longitudinal data. Brookings Papers Econ. Act. (1), 1–70.

Krueger, A.B., Cramer, J., Cho, D., 2014. Are the long-term unemployed on the margins of the labor market? Brookings Papers Econ. Act. Spring, 229–299.

Krugman, P.R., 1998. It's Baaack: Japan's slump and the return of the liquidity trap. Brookings Papers Econ. Act. (2), 137–205.

Kuehn, L.A., Petrosky-Nadeau, N., Zhang, L., 2013. An Equilibrium Asset Pricing Model with Labor Market Search. Carnegie Mellon University, Tepper School of Business, Pittsburgh.

Ludvigson, S.C., Ma, S., Ng, S., 2015. Uncertainty and business cycles: exogenous impulse or endogenous response? Working Paper 21803, National Bureau of Economic Research. http://www.nber.org/papers/w21803.

Lustig, H., Verdelhan, A., 2012. Business cycle variation in the risk-return trade-off. J. Monetary Econ. 0304-3932. 59, S35–S49. http://dx.doi.org/10.1016/j.jmoneco.2012.11.003. http://www.sciencedirect.com/science/article/pii/S0304393212001511.

Lustig, H., Van Nieuwerburgh, S., Verdelhan, A., 2013. The wealth-consumption ratio. Rev. Asset Pricing Stud. 3 (1), 38–94. http://dx.doi.org/10.1093/rapstu/rat002. http://raps.oxfordjournals.org/content/3/1/38.abstract.

Mian, A., Sufi, A., 2010. The great recession: lessons from microeconomic data. Am. Econ. Rev. 100 (2), 51–56. ISSN 0002-8282. http://www.jstor.org/stable/27804962.

Mian, A.R., Sufi, A., 2012. What explains high unemployment? The aggregate demand channel. Working Paper 17830, National Bureau of Economic Research. http://www.nber.org/papers/w17830.

Mian, A., Rao, K., Sufi, A., 2013. Household balance sheets, consumption, and the economic slump. Q. J. Econ. 128 (4), 1687–1726. http://dx.doi.org/10.1093/qje/qjt020. http://qje.oxfordjournals.org/content/128/4/1687.abstract.

Michaillat, P., Saez, E., 2014. An economical business-cycle model. Working Paper 19777, National Bureau of Economic Research. http://www.nber.org/papers/w19777.

Mortensen, D.T., 2011. Comments on Hall's Clashing Theories of Unemployment. Department of Economics, Northwestern University, Chicago.

Mulligan, C.B., 2012. Do welfare policies matter for labor market aggregates? Quantifying safety net work incentives since 2007. Working Paper 18088. National Bureau of Economic Research. http://www.nber.org/papers/w18088.

Mulligan, C.B., 2012. The Redistribution Recession: How Labor Market Distortions Contracted the Economy. Oxford University Press, New York.

Nekarda, C.J., Ramey, V.A., 2013. The cyclical behavior of the price-cost markup. Working Paper 19099, National Bureau of Economic Research. http://www.nber.org/papers/w19099.

Parker, J.A., Souleles, N.S., Johnson, D.S., McClelland, R., 2011. Consumer spending and the economic stimulus payments of 2008. Working Paper 16684, National Bureau of Economic Research. http://www.nber.org/papers/w16684.

Petev, I., Pistaferri, L., Eksten, I.S., 2012. Consumption and the great recession: an analysis of trends, perceptions, and distributional effects. In: Grusky, D.B., Western, B., Wimer, C. (Eds.), The Great Recession. Russell Sage Foundation, New York.

Petrosky-Nadeau, N., Wasmer, E., 2013. The cyclical volatility of labor markets under frictional financial markets. Am. Econ. J. Macroecon. 5 (1), 193–221. http://dx.doi.org/10.1257/mac.5.1.193. http://www.ingentaconnect.com/content/aea/aejma/2013/00000005/00000001/art00007.

Petrosky-Nadeau, N., Wasmer, E., 2015. Macroeconomic dynamics in a model of goods, labor, and credit market frictions. J. Monetary Econ. 72 May, 97–113. ISSN 0304-3932. http://dx.doi.org/10.1016/j.jmoneco.2015.01.006. http://www.sciencedirect.com/science/article/pii/S0304393215000161.

Petrosky-Nadeau, N., Zhang, L., 2013. Unemployment crises. Working Paper 19207, National Bureau of Economic Research. http://www.nber.org/papers/w19207.

Phelps, E.S., Winter, S.G., 1970. Optimal price policy under atomistic competition. In: Phelps, E.S. et al. (Eds.), Microeconomic Foundations of Employment and Inflation Theory. Norton, New York, pp. 309–337.

Philippon, T., 2009. The bond market's q. Q. J. Econ. 124 (3), 1011–1056. http://ideas.repec.org/a/tpr/qjecon/v124y2009i3p1011-1056.html.

Ramey, V.A., 2011a. Can government purchases stimulate the economy? J. Econ. Liter. 49 (3), 673–685. http://dx.doi.org/10.1257/jel.49.3.673.

Ramey, V.A., 2011b. Identifying government spending shocks: it's all in the timing. Q. J. Econ. 126 Feb., 1–50. http://dx.doi.org/10.1093/qje/qjq008.. http://qje.oxfordjournals.org/content/early/2011/03/21/qje.qjq008.abstract

Ravn, M.O., Sterk, V., 2012. Job Uncertainty and Deep Recessions. University College London, London.

Reifschneider, D., Wascher, W., Wilcox, D., 2013. Aggregate supply in the United States: recent developments and implications for the conduct of monetary policy. Technical Report, Finance and Economics Discussion Series Divisions of Research & Statistics and Monetary Affairs Federal Reserve Board, Washington, DC.

Restrepo, P., 2015. Skill Mismatch and Structural Unemployment. Massachusetts Institute of Technology, Cambridge, MA.

Rognlie, M., 2015. Deciphering the fall and rise in the net capital share. Broookings Papers Econ. Act. 50 (1 (Spring)), 1–69.

Rotemberg, J.J., Woodford, M., 1999. The cyclical behavior of prices and costs. In: Taylor, Woodford, (Eds.), Handbook of Macroeconomics. Elsevier, North Holland, pp. 1051–1135 (Chapter 16).

Rothstein, J., 2011. Unemployment insurance and job search in the great recession. Brookings Papers Econ. Act. 43 (2 (Fall)), 143–213.

Sahin, A., Song, J., Topa, G., Violante, G.L., 2012. Mismatch unemployment. Working Paper 18265, National Bureau of Economic Research. http://www.nber.org/papers/w18265.

Shapiro, M.D., Slemrod, J., 2009. Did the 2008 tax rebates stimulate spending? Am. Econ. Rev. Papers Proc. 99 (2), 374–379.

Shimer, R., 2005. The cyclical behavior of equilibrium unemployment and vacancies. Am. Econ. Rev. 95 (1), 24–49.

Shimer, R., 2008. Convergence in macroeconomics: the labor wedge. Am. Econ. J. Macroecon. 1 (1), 280–297.

Spilimbergo, A., Symansky, S., Schindler, M., 2009. Fiscal multipliers, IMF Staff Position Note 2009.

Stock, J.H., Watson, M.W., 2010. Modeling inflation after the crisis. Proceedings of the Economic Policy Symposium, Federal Reserve Bank of Kansas City Working Paper. pp. 172–220. http://www.nber.org/papers/w16488.

Valletta, R., Kuang, K., 2010a. Extended unemployment and UI benefits. Federal Reserve Bank of San Francisco Economic Letter, pp. 1–4.

Valletta, R., Kuang, K., 2010b. Is structural unemployment on the rise? Federal Reserve Bank of San Francisco Economic Letter, pp. 1–5.

Walsh, C.E., 2003. Labor market search and monetary shocks. In: Altug, S., Chadha, J., Nolan, C. (Eds.), Elements of Dynamic Macroeconomic Analysis, Cambridge University Press, Cambridge, UK, pp. 451–486.

SECTION 5

Macroeconomic Policy

CHAPTER 28

Challenges for Central Banks' Macro Models

J. Lindé[*,†,‡], **F. Smets**[§,¶,‡], **R. Wouters**[||,‡]
[*] Sveriges Riksbank, Stockholm, Sweden
[†] Stockholm School of Economics, Stockholm, Sweden
[‡] CEPR, London, United Kingdom
[§] ECB, Frankfurt, Germany
[¶] KU Leuven, Leuven, Belgium
[||] National Bank of Belgium, Brussels, Belgium

Contents

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.04.009

Abstract

In this chapter, we discuss a number of challenges for structural macroeconomic models in the light of the Great Recession and its aftermath. It shows that a benchmark DSGE model that shares many features with models currently used by central banks and large international institutions has difficulty explaining both the depth and the slow recovery of the Great Recession. In order to better account for these observations, the chapter analyses three extensions of the benchmark model. First, we estimate the model allowing explicitly for the zero lower bound constraint on nominal interest rates. Second, we introduce time variation in the volatility of the exogenous disturbances to account for the non-Gaussian nature of some of the shocks. Third and finally, we extend the model with a financial accelerator and allow for time variation in the endogenous propagation of financial shocks. All three extensions require that we go beyond the linear Gaussian assumptions that are standard in most policy models. We conclude that these extensions go some way in accounting for features of the Great Recession and its aftermath, but they do not suffice to address some of the major policy challenges associated with the use of nonstandard monetary policy and macroprudential policies.

Keywords

Monetary policy, DSGE, and VAR models, Regime switching, Zero lower bound, Financial frictions, Great recession, Macroprudential policy, Open economy

JEL Classification Codes

E52, E58

1. INTRODUCTION

In this chapter, we discuss new challenges for structural macroeconomic models used at central banks in light of the Great Recession in United States and other advanced economies. This recession has had widespread implications for economic policy and economic performance, with historically low nominal interest rates and elevated unemployment levels in its aftermath. The fact that the intensification of the crisis in the fall of 2008 was largely unexpected and much deeper than central banks predicted and that the subsequent recovery was much slower, has raised many questions about the design of macroeconomic models at use in these institutions. Specifically, the models have been criticized for omitting key financial mechanisms and shocks stemming from the financial sector.

We start by analyzing the performance of a benchmark macroeconomic model during the Great Recession. The model we use—the well-known Smets and Wouters (2007) model—shares many features with the models currently used by central banks. When we analyze this model estimated over the precrisis period we find, confirming previous results in Del Negro and Schorfheide (2013), that actual GDP growth was outside the predictive density of the model during the most acute phase of the recession. To account for the depth of the recession, the model needs a cocktail of extremely unlikely shocks that mainly affect the intertemporal decision of households and firms to consume or invest such as risk-premium and investment-specific technology shocks. We then proceed to document that these shocks are non-Gaussian, and strongly related to observable financial variables such as the Baa-Aaa and term spread, suggesting the importance of including financial shocks and frictions to account for large recessions. Moreover, in order to account for the slow recovery, restrictive monetary policy shocks reflecting a binding lower bound on the nominal interest rate, negative investment shocks, and positive price mark-up shocks are needed. This configuration of shocks explains the slow recovery and the missing disinflation following the great recession.

To try to better account for these observations, we proceed to amend the benchmark model along three dimensions. First, we take the zero lower bound (ZLB henceforth) explicitly into account when estimating the model over the full sample. We do this using two alternative approaches. First, we implement the ZLB as a binding constraint on the policy rule with an expected duration that is determined endogenously by the model in each period. Second, we impose the expected duration of the ZLB spells during the recession to be consistent with external information derived from overnight index swap rates. Importantly, we find that the variants of the model estimated subject to the ZLB constraint typically feature a substantially higher degree of nominal stickiness in both prices and wages which helps to understand the inflation dynamics during the recession period and the subsequent slow recovery. In addition, an important characteristic of these variants of the model is a substantially higher response coefficient on the output gap in the policy rule. Incorporating the ZLB in the estimation and simulation of the model does not materially affect the median forecast of output and inflation in 2008Q3 as the probability of hitting the lower bound is estimated to be low before the crisis. It does, however, tilt the balance of risks towards the downside in the subsequent periods as the likelihood of monetary policy being constrained increases.

Second, in order to account for the non-Gaussian nature of the shocks driving most recessions, we allow for time-varying volatility in some of the shocks. In line with the previous literature, we find that the empirical performance of the model improves a lot when two regime change processes are allowed in the variance of the shocks. One of those regime switches captures the great moderation period from the mid-1980s to the mid-2000s, when overall macroeconomic volatility was much lower than both before and after this period. The other regime switching process captures the higher volatility of the risk-premium, the monetary policy, and the investment-specific technology shocks in recession periods. This regime switching process can account for the non-Gaussian nature of those shocks and also helps widening the predictive density of output growth at the end of 2008 as the probability of a financial recession increases.

Finally, we proceed to examine how the performance and properties of the basic model can be improved by introducing a financial accelerator mechanism and explicit shocks stemming from the financial sector. This exercise is initiated by embedding a variant of the Bernanke et al. (1999) financial accelerator into the workhorse model and estimating it under the standard assumption that the financial sector excerpts a time-invariant influence on business cycles: that is, we follow, eg, Christiano et al. (2003a), De Graeve (2008) and Queijo von Heideken (2009), and assume that the parameters characterizing the financial frictions are constant and that shocks stemming from the financial bloc are Gaussian. In this specification, we do not find that the financial accelerator adds much propagation of other macroeconomic shocks, and that movements in the Baa-Aaa spread we add as observable is mostly explained by the exogenous shock stemming from the financial sector. Driven by this result, and because of the non-Gaussian features of the smoothed shocks in the benchmark model, we examine if the performance of this augmented model can be improved by allowing for regime switching in the sensitivity of the external finance premium to the leverage ratio, which one may think of as risk-on/risk-off behavior in the financial sector. We find that allowing for regime switching in the sensitivity of external finance premium to the leverage ratio introduces a high degree of skewness in the predictive density of the spread and makes the model put nonzero probability in the predictive density on the observed 2008Q4 output growth outcome. Moreover, when we follow Del Negro and Schorfheide (2013) and condition on the actual spread outcome during the fourth quarter of 2008—which is reasonable since the spread reached its quarterly mean in the beginning of October—the model's ability to account for the severe growth outcome further improves. This result indicates that if we appropriately could integrate the nonlinear accelerator dynamics from financial frictions in our models, we may obtain a more realistic predictive density in line with reduced form time-varying volatility models.

The three extensions discussed in this chapter go some way to address some of the challenges faced by the benchmark DSGE model in accounting for the Great Recession and its aftermath. They all involve going beyond the linear Gaussian-modeling framework. However, they do not suffice to fully address some of the major empirical policy challenges. These new challenges stem from the fact that, following the crisis and hitting the zero lower bound, central banks have implemented a panoply of nonstandard monetary policy measures such a Large-Scale-Asset-Purchases and other credit easing policies. Basic extensions of the benchmark model with financial frictions (such as a financial accelerator) are not sufficient to be able to fully analyze the effectiveness of those policies and their interaction with the standard interest rate policy. Similarly, the financial crisis has given rise to the new macroprudential policy domain that aims at containing systemic risk and preserving financial stability. Current extensions of the benchmark model are often not rich enough to analyze the interaction between monetary and macroprudential policy. Being able to do so will require incorporating of a richer description of both solvency (default) and liquidity (bank runs) dynamics with greater complexity in terms of both nonlinearities and heterogeneity.

The rest of the chapter is structured as follows. Section 2 provides an incomplete survey of the macroeconomic models used by central banks and other international organizations. Following this survey, Section 3 presents the prototype model—the estimated model of Smets and Wouters (2003). This model shares many features of models in use by central banks. The section also discusses the data and the estimation of this model on precrisis data. In Section 4, we use this model estimated on precrisis data to analyze the crisis episode, which gives us valuable insights into the workings of the model. We also compare the performance of our structural model to a reduced-form benchmark VAR, which is estimated with Bayesian priors. As this analysis points to some important shortcomings of the benchmark model, we augment the baseline model in Section 5 along the three dimensions discussed earlier.

Finally, Section 6 sums up by discussing some other new and old challenges for structural macro models used in policy analysis and presents some conclusions. Appendices contain some technical details on the model, methods, and the data used in the analysis.

2. COMMON FEATURES OF CENTRAL BANK MODELS

In this section, we provide an incomplete survey of the key policy models currently in use at central banks and other key policy institutions like the IMF, European Commission, and the OECD. We aim at determining the similarity between models, and assess if—and how—they have been changed in response to the recession and developments since then.

A good starting point for the discussion is the paper by Coenen et al. (2012). Wieland et al. (2012) provides a complementary and very useful overview of policy models in use at central banks. An additional advantage with the paper by Wieland et al. is that they have pulled together an archive with well-know estimated macroeconomic models (both policy and academic) that can conveniently be used to run and compare various diagnostic shocks using a Matlab graphical user interface.[a] We nevertheless base our discussion on Coenen et al., as they focus exclusively on models in use at policy institutions. Coenen et al. studies the effects of monetary and fiscal shocks in the key policy models in use at the Bank of Canada (BoC-GEM), the Board of Governors of the Federal Reserve System (with two models, FRB-US and SIGMA), the European Central Bank (NAWM), the European Commission (QUEST), the International Monetary Fund (GIMF), and the OECD (OECD Fiscal). Out of the seven models, six are dynamic

[a] Taylor and Wieland (2012) use the database to compare the responses to monetary policy shocks. Wieland and Wolters (2013) study the forecasting behavior for a large set of models in the database.

stochastic general equilibrium (DSGE) models, while one–the FRB-US—is based on the polynomial adjustment cost (PAC) framework. Hence, an overwhelming majority of key policy institutions today use DSGE models as the core policy tool.[b] The switch from traditional backward-looking macroeconometric models (see, eg, Rudebusch and Svensson, 1999) to DSGEs occurred amid the forceful critique by Lucas (1976) and Sims (1980) of such models, and was made feasible due to the progress in the solution and estimation of such models (see, eg, Blanchard and Kahn, 1980 and Fair and Taylor, 1983) as well as the contribution of Christiano et al. (2005) who showed that such models, carefully specified, could feature a realistic monetary policy transmission mechanism. As pointed out by Clarida et al. (1999), Woodford (2003) and Galí (2008), these models assigns an important role to expectations for macroeconomic stabilization, and this view was embraced by policy makers at central banks. However, although macroeconomic models have been used in scenario analysis and affected policy making more generally, it is probably fair to say that the models impact on the short- and medium-term economic projections have been limited, see, eg, Iversen et al. (2016).

As outlined in detail in tables 1 and 2 by Coenen et al. (2012), the DSGE models share many similarities to the seminal models of Christiano et al. (2005) (CEE henceforth) and Smets and Wouters (2003, 2007). They typically feature imperfect competition in product and labor markets as vehicles to introduce sticky prices and wages. They also include important real rigidities like habit formation, costs of adjusting investment and variable capital utilization. Monetary policy is generally determined by a simple Taylor-type policy rule which allows for interest rate smoothing, but although they share many similarities with the academic benchmark models of CEE and Smets and Wouters (2007) (SW07 henceforth), policy models often embed some additional features. One such important feature is that they have a significant share of financially constrained households, ranging between 20% and 50%. In some models these are hand-to-mouth households, who take their labor income as given and determine consumption residually from a period-by-period budget constraint. In other models these are liquidity-constrained households, who face the same period-by-period budget constraint but solve an intertemporal decision problem between consumption and work effort. An additional difference between the policy models and the academic style ones is that the former generally has a much more detailed fiscal sector with many distortionary taxes, types of government spending and various transfers from the government to the households.[c]

[b] Other prominent institutions that have adopted estimated DSGE model as their core policy tool include Bank of England (COMPASS, see Burgess et al., 2013), Norges Bank (NEMO, see Brubakk et al., 2006), Sveriges Riksbank (RAMSES, see Adolfson et al., 2013), Federal Reserve Bank of New York (Del Negro et al., 2013), and the Federal Reserve Bank of Chicago (Brave et al., 2012).

[c] These results are broadly in line with the findings of Wieland et al. (2012).

Another interesting observation is that neither CEE nor SW07 include frictions in financial markets or a detailed banking sector in their models.[d] Four of the seven policy models included financial frictions prior to the crisis. By asking the policy institutions that were part of this study about their development efforts since then, it is clear that efforts have been made towards better integration of financial markets, with a focus on the interaction between banks and the firms in the economy. For instance, following the crisis, financial frictions following the approach of Bernanke et al. (1999) have been introduced in (at least) two of the three models that did not feature them before.[e]

The key lesson we draw from this is that while the crisis has had some impact on improving the modeling of the financial sector in DSGE models, it has not so far had a material impact on the type of models used at key policy institutions, which still share many features of the basic model developed by CEE.

3. A BENCHMARK MODEL

In this section, we show the benchmark model environment, which is the model of Smets and Wouters (2007). The SW07-model builds on the workhorse model by CEE, but allows for a richer set of stochastic shocks. In Section 3.4, we describe how we estimate it using aggregate times series for the United States.

3.1 Firms and Price Setting

3.1.1 Final Goods Production

The single final output good Y_t is produced using a continuum of differentiated intermediate goods $Y_t(f)$. Following Kimball (1995), the technology for transforming these intermediate goods into the final output good is

$$\int_0^1 G_Y\left(\frac{Y_t(f)}{Y_t}\right) df = 1. \tag{1}$$

As in Dotsey and King (2005), we assume that $G_Y(\cdot)$ is given by a strictly concave and increasing function:

[d] The CEE, but not the SW07-model, includes a working capital—or cost channel—of monetary policy whereby firms have to borrow at the policy rate to finance the wage bill. This channel allows the CEE model to account for the "Price-puzzle" (ie, that inflation rises on impact following a hike in the policy rate) that often emerges for monetary policy shocks in identified VAR models.

[e] We are grateful to Günter Coenen (ECB) and John Roberts (Federal Reserve Board) for providing very helpful responses to our questionnaire.

$$G_Y\left(\frac{Y_t(f)}{Y_t}\right)=\frac{\phi_t^p}{1-(\phi_t^p-1)\epsilon_p}\left[\left(\frac{\phi_t^p+(1-\phi_t^p)\epsilon_p}{\phi_t^p}\right)\frac{Y_t(f)}{Y_t}+\frac{(\phi_t^p-1)\epsilon_p}{\phi_t^p}\right]^{\frac{1-(\phi_t^p-1)\epsilon_p}{\phi_t^p-(\phi_t^p-1)\epsilon_p}} \\ +\left[1-\frac{\phi_t^p}{1-(\phi_t^p-1)\epsilon_p}\right], \tag{2}$$

where $\phi_t^p \geq 1$ denotes the gross markup of the intermediate firms. The parameter ϵ_p governs the degree of curvature of the intermediate firm's demand curve. When $\epsilon_p=0$, the demand curve exhibits constant elasticity as with the standard Dixit–Stiglitz aggregator. When ϵ_p is positive the firms instead face a quasi-kinked demand curve, implying that a drop in the good's relative price only stimulates a small increase in demand. On the other hand, a rise in its relative price generates a large fall in demand. Relative to the standard Dixit–Stiglitz aggregator, this introduces more strategic complementary in price setting which causes intermediate firms to adjust prices less to a given change in marginal cost. Finally, notice that $G_Y(1)=1$, implying constant returns to scale when all intermediate firms produce the same amount of the good.

Firms that produce the final output good are perfectly competitive in both product and factor markets. Thus, final goods producers minimize the cost of producing a given quantity of the output index Y_t, taking the price $P_t(f)$ of each intermediate good $Y_t(f)$ as given. Moreover, final goods producers sell the final output good at a price P_t, and hence solve the following problem:

$$\max_{\{Y_t,\, Y_t(f)\}} P_tY_t-\int_0^1 P_t(f)\,Y_t(f)df, \tag{3}$$

subject to the constraint in (1). The first order conditions (FOCs) for this problem can be written

$$\begin{aligned}\frac{Y_t(f)}{Y_t}&=\frac{\phi_t^p}{\phi_t^p-(\phi_t^p-1)\epsilon_p}\left(\left[\frac{P_t(f)}{P_t}\frac{1}{\Lambda_t^p}\right]^{-\frac{\phi_t^p-(\phi_p-1)\epsilon_p}{\phi_t^p-1}}+\frac{(1-\phi_t^p)\epsilon_p}{\phi_t^p}\right)\\ P_t\Lambda_t^p&=\left[\int P_t(f)^{-\frac{1-(\phi_t^p-1)\epsilon_p}{\phi_t^p-1}}df\right]^{-\frac{\phi_t^p-1}{1-(\phi_t^p-1)\epsilon_p}}\\ \Lambda_t^p&=1+\frac{(1-\phi_t^p)\epsilon_p}{\phi_p}-\frac{(1-\phi_t^p)\epsilon_p}{\phi_t^p}\int\frac{P_t(f)}{P_t}df,\end{aligned} \tag{4}$$

where Λ_t^p denotes the Lagrange multiplier on the aggregator constraint in (1). Note that when $\epsilon_p=0$, it follows from the last of these conditions that $\Lambda_t^p=1$ in each period t, and the demand and pricing equations collapse to the usual Dixit–Stiglitz expressions, ie,

$$\frac{Y_t(f)}{Y_t} = \left[\frac{P_t(f)}{P_t}\right]^{-\frac{\phi_t^p}{\phi_t^p-1}}, P_t = \left[\int P_t(f)^{\frac{1}{1-\phi_t^p}} df\right]^{1-\phi_t^p}.$$

3.1.2 Intermediate Goods Production

A continuum of intermediate goods $Y_t(f)$ for $f \in [0, 1]$ is produced by monopolistic competitive firms, each of which produces a single differentiated good. Each intermediate goods producer faces the demand schedule in Eq. (4) from the final goods firms through the solution to the problem in (3), which varies inversely with its output price $P_t(f)$ and directly with aggregate demand Y_t.

Each intermediate goods producer utilizes capital services $K_t(f)$ and a labor index $L_t(f)$ (defined later) to produce its respective output good. The form of the production function is Cobb–Douglas:

$$Y_t(f) = \varepsilon_t^a K_t(f)^\alpha [\gamma^t L_t(f)]^{1-\alpha} - \gamma^t \Phi,$$

where γ^t represents the labor-augmenting deterministic growth rate in the economy, Φ denotes the fixed cost (which is related to the gross markup ϕ_t^p so that profits are zero in the steady state), and ε_t^a is a total productivity factor which follows a Kydland and Prescott (1982) style process:

$$\ln \varepsilon_t^a = \rho_a \ln \varepsilon_{t-1}^a + \eta_t^a, \eta_t^a \sim N(0, \sigma_a). \tag{5}$$

Firms face perfectly competitive factor markets for renting capital and hiring labor. Thus, each firm chooses $K_t(f)$ and $L_t(f)$, taking as given both the rental price of capital R_{Kt} and the aggregate wage index W_t (defined later). Firms can without costs adjust either factor of production, thus, the standard static first-order conditions for cost minimization implies that all firms have identical marginal costs per unit of output.

The prices of the intermediate goods are determined by nominal contracts in Calvo (1983) and Yun (1996) staggered style nominal contracts. In each period, each firm f faces a constant probability, $1 - \xi_p$, of being able to reoptimize the price $P_t(f)$ of the good. The probability that any firm receives a signal to reoptimize the price is assumed to be independent of the time that it last reset its price. If a firm is not allowed to optimize its price in a given period, this is adjusted by a weighted combination of the lagged and steady state rate of inflation, ie, $P_t(f) = (1+\pi_{t-1})^{\iota_p}(1+\pi)^{1-\iota_p} P_{t-1}(f)$ where $0 \le \iota_p \le 1$ and π_{t-1} denotes net inflation in period $t - 1$, and π the steady state net inflation rate. A positive value of the indexation parameter ι_p introduces structural inertia into the inflation process. All told, this leads to the following optimization problem for the intermediate firms

$$\max_{\tilde{P}_t(f)} \mathrm{E}_t \sum_{j=0}^{\infty} (\beta\xi_p)^j \frac{\Xi_{t+j} P_t}{\Xi_t P_{t+j}} \left[\tilde{P}_t(f)\left(\Pi_{s=1}^j (1+\pi_{t+s-1})^{\iota_p} (1+\pi)^{1-\iota_p}\right) - MC_{t+j}\right] Y_{t+j}(f),$$

where $\widetilde{P}_t(f)$ is the newly set price and $\beta^j \frac{\Xi_{t+j} P_t}{\Xi_t P_{t+j}}$ the stochastic discount factor. Notice that given our assumptions, all firms that reoptimize their prices actually set the same price.

As noted previously, we assume that the gross price-markup is time varying and given by $\phi_t^p = \phi^p \varepsilon_t^p$, for which the exogenous component ε_t^p is given by an exogenous ARMA (1,1) process:

$$\ln \varepsilon_t^p = \rho_p \ln \varepsilon_{t-1}^p + \eta_t^p - \vartheta_p \eta_{t-1}^p, \eta_t^p \sim N(0, \sigma_p). \tag{6}$$

3.2 Households and Wage Setting

Following Erceg et al. (2000), we assume a continuum of monopolistic competitive households (indexed on the unit interval), each of which supplies a differentiated labor service to the production sector; that is, goods-producing firms regard each household's labor services $L_t(h)$, $h \in [0, 1]$, as imperfect substitutes for the labor services of other households. It is convenient to assume that a representative labor aggregator combines households' labor hours in the same proportions as firms would choose. Thus, the aggregator's demand for each household's labor is equal to the sum of firms' demands. The aggregated labor index L_t has the Kimball (1995) form:

$$L_t = \int_0^1 G_L\left(\frac{L_t(h)}{L_t}\right) dh = 1, \tag{7}$$

where the function $G_L(\cdot)$ has the same functional form as does (2), but is characterized by the corresponding parameters ϵ_w (governing convexity of labor demand by the aggregator) and a time-varying gross wage markup ϕ_t^w. The aggregator minimizes the cost of producing a given amount of the aggregate labor index L_t, taking each household's wage rate $W_t(h)$ as given, and then sells units of the labor index to the intermediate goods sector at unit cost W_t, which can naturally be interpreted as the aggregate wage rate. From the FOCs, the aggregator's demand for the labor hours of household h—or equivalently, the total demand for this household's labor by all goods-producing firms—is given by

$$\frac{L_t(h)}{L_t} = G_L'^{-1}\left[\frac{W_t(h)}{W_t} \int_0^1 G_L'\left(\frac{L_t(h)}{L_t}\right) \frac{L_t(h)}{L_t} dh\right], \tag{8}$$

where $G_L'(\cdot)$ denotes the derivative of the $G_L(\cdot)$ function in Eq. (7).

The utility function of a typical member of household h is

$$\mathrm{E}_t \sum_{j=0}^{\infty} \beta^j \left[\frac{1}{1-\sigma_c}\left(C_{t+j}(h) - \varkappa C_{t+j-1}\right)\right]^{1-\sigma_c} \exp\left(\frac{\sigma_c - 1}{1+\sigma_l} L_{t+j}(h)^{1+\sigma_l}\right), \tag{9}$$

where the discount factor β satisfies $0 < \beta < 1$. The period utility function depends on household h's current consumption $C_t(h)$, as well as lagged aggregate consumption per capita, to allow for external habit persistence (captured by the parameter $\varkappa$). The period utility function also depends inversely on hours worked $L_t(h)$.

Household h's budget constraint in period t states that expenditure on goods and net purchases of financial assets must equal to the disposable income:

$$\begin{aligned} &P_t C_t(h) + P_t I_t(h) + \frac{B_{t+1}(h)}{\varepsilon_t^b R_t} + \int_s \xi_{t,t+1} B_{D,t+1}(h) - B_{D,t}(h) \\ &= B_t(h) + W_t(h)L_t(h) + R_t^k Z_t(h) K_t^p(h) - a(Z_t(h))K_t^p(h) + \Gamma_t(h) - T_t(h). \end{aligned} \tag{10}$$

Thus, the household purchases part of the final output good (at a price of P_t), which is chosen to be consumed $C_t(h)$ or invest $I_t(h)$ in physical capital. Following Christiano et al. (2005), investment augments the household's (end-of-period) physical capital stock $K_{t+1}^p(h)$ according to

$$K_{t+1}^p(h) = (1-\delta)K_t^p(h) + \varepsilon_t^i \left[1 - S\left(\frac{I_t(h)}{I_{t-1}(h)}\right)\right] I_t(h). \tag{11}$$

The extent to which investment by each household turns into physical capital is assumed to depend on an exogenous shock ε_t^i and how rapidly the household changes its rate of investment according to the function $S\left(\frac{I_t(h)}{I_{t-1}(h)}\right)$, which we assume satisfies $S(\gamma) = 0, S'(\gamma) = 0$ and $S''(\gamma) = \varphi$ where γ is the steady state gross growth rate of the economy. The stationary investment-specific shock ε_t^i follows the process:

$$\ln \varepsilon_t^i = \rho_i \ln \varepsilon_{t-1}^i + \eta_t^i, \eta_t^i \sim N(0, \sigma_i).$$

In addition to accumulating physical capital, households may augment their financial assets through increasing their nominal bond holdings (B_{t+1}), from which they earn an interest rate of R_t. The return on these bonds is also subject to a risk-shock, ε_t^b, which follows

$$\ln \varepsilon_t^b = \rho_b \ln \varepsilon_{t-1}^b + \eta_t^b, \eta_t^b \sim N(0, \sigma_b). \tag{12}$$

Fisher (2015) shows that this shock can be given a structural interpretation.

We assume that agents can engage in friction-less trading of a complete set of contingent claims to diversify away idiosyncratic risk. The term $\int_s \xi_{t,t+1} B_{D,t+1}(h) - B_{D,t}(h)$ represents net purchases of these state-contingent domestic bonds, with $\xi_{t,t+1}$ denoting the state-dependent price, and $B_{D,t+1}(h)$ the quantity of such claims purchased at time t.

On the income side, each member of household h earns labor income $W_t(h)L_t(h)$, capital rental income of $R_t^k Z_t(h) K_t^p(h)$, and pays a utilization cost of the physical capital equal to $a(Z_t(h))K_t^p(h)$ where $Z_t(h)$ is the capital utilization rate. The capital services provided by household h, $K_t(h)$ thereby equals $Z_t(h)K_t^p(h)$. The capital utilization adjustment function $a(Z_t(h))$ is assumed to satisfy $a(1) = 0$, $a'(1) = r^k$, and $a''(1) = \psi/(1-\psi) > 0$, where $\psi \in [0, 1)$ and a higher value of ψ implies a higher cost of changing the utilization rate. Finally, each member also receives an aliquot share $\Gamma_t(h)$ of the profits of all firms, and pays a lump-sum tax of $T_t(h)$ (regarded as taxes net of any transfers).

In every period t, each member of household h maximizes the utility function in (9) with respect to consumption, investment, (end-of-period) physical capital stock, capital utilization rate, bond holdings, and holdings of contingent claims, subject to the labor demand function (8), budget constraint (10), and transition equation for capital (11).

Households also set nominal wages in Calvo-style staggered contracts that are generally similar to the price contracts described previously. Thus, the probability that a household receives a signal to reoptimize its wage contract in a given period is denoted by $1 - \xi_w$. In addition, SW07 specify the following dynamic indexation scheme for the adjustment of wages for those households that do not get a signal to reoptimize: $W_t(h) = \gamma(1+\pi_{t-1})^{\iota_w}(1+\pi)^{1-\iota_w} W_{t-1}(h)$. All told, this leads to the following optimization problem for the households

$$\max_{\tilde{W}_t(h)} \mathrm{E}_t \sum_{j=0}^{\infty} (\beta\xi_w)^j \frac{\Xi_{t+j}P_t}{\Xi_t P_{t+j}} \left[\tilde{W}_t(h)\left(\Pi_{s=1}^{j}\gamma(1+\pi_{t+s-1})^{\iota_w}(1+\pi)^{1-\iota_w}\right) - W_{t+j}\right] L_{t+j}(h),$$

where $\tilde{W}_t(h)$ is the newly set wage and $L_{t+j}(h)$ is determined by Eq. (7). Notice that with our assumptions all households that reoptimize their wages will actually set the same wage.

Following the same approach as with the intermediate-goods firms, we introduce a shock ε_t^w to the time-varying gross markup, $\phi_t^w = \phi^w \varepsilon_t^w$, where ε_t^w is assumed being given by an exogenous ARMA(1,1) process:

$$\ln \varepsilon_t^w = \rho_w \ln \varepsilon_{t-1}^w + \eta_t^w - \vartheta_w \eta_{t-1}^w, \eta_t^w \sim N(0, \sigma_w). \tag{13}$$

3.3 Market Clearing Conditions and Monetary Policy

Government purchases G_t are exogenous, and the process for government spending relative to trend output in natural logs, ie, $g_t = G_t/(\gamma^t Y)$, is given by the following exogenous AR(1) process:

$$\ln g_t = \left(1-\rho_g\right) \ln g + \rho_g\left(\ln g_{t-1} - \rho_{ga} \ln \varepsilon_{t-1}^a\right) + \eta_t^g, \eta_t^g \sim N\left(0, \sigma_g\right).$$

Government purchases neither have any effects on the marginal utility of private consumption, nor do they serve as an input into goods production. The consolidated government sector budget constraint is

$$\frac{B_{t+1}}{R_t} = G_t - T_t + B_t,$$

where T_t are lump-sum taxes. By comparing the debt terms in the household budget constraint in Eq. (10) with the equation earlier, one can see that receipts from the risk shock are subject to iceberg costs, and hence do not add any income to the government.[f] We acknowledge that this is an extremely simplistic modeling of the fiscal behavior of the government relative to typical policy models, and there might be important feedback effects between fiscal and monetary policies that our model does not allow for.[g] As discussed by Benigno and Nisticó (2015) and Del Negro and Sims (2014), the fiscal links between governments and central banks may be especially important today when central banks have employed unconventional tools in monetary policy. Nevertheless, we maintain our simplistic modeling of fiscal policy throughout the chapter, as it allows us to examine the partial implications of amending the benchmark model with more elaborate financial markets modeling and the zero lower bound constraint more directly.

The conduct of monetary policy is assumed to be approximated by a Taylor-type policy rule (here stated in nonlinearized form)

$$R_t \;= \max\left(0, R^{1-\rho_R} R_{t-1}^{\rho_R} \left(\frac{\Pi_t}{\Pi}\right)^{r_\pi(1-\rho_R)} \left(\frac{Y_t}{Y_t^{pot}}\right)^{r_y(1-\rho_R)} \left(\frac{Y_t}{Y_t^{pot}} \Big/ \frac{Y_{t-1}}{Y_{t-1}^{pot}}\right)^{r_{\Delta y}(1-\rho_R)} \varepsilon_t^r\right), \tag{14}$$

where Π_t denotes the is gross inflation rate, Y_t^{pot} is the level of output that would prevail if prices and wages were flexible, and variables without subscripts denote steady state values. The policy shock ε_t^r is supposed to follow an AR(1) process in natural logs:

$$\ln \varepsilon_t^r = \rho_r \ln \varepsilon_{t-1}^r + \eta_t^r, \eta_t^r \sim N(0, \sigma_r). \tag{15}$$

Total output of the final goods sector is used as follows:

$$Y_t = C_t + I_t + G_t + a(Z_t) - K_t,$$

where $a(Z_t) - K_t$ is the capital utilization adjustment cost.

3.4 Estimation on Precrisis Data

We now proceed to discuss how the model is estimated. To begin with, we limit the sample to the period 1965Q1–2007Q4 to see how a model estimated on precrisis data fares during the recession. Subsequently, we will estimate the model on data spanning the crisis.

[f] But even if they did, it would not matter as the government is assumed to balance its expenditures each period through lump-sum taxes, $T_t = G_t + B_t - B_{t+1}/R_t$, so that government debt $B_t = 0$ in equilibrium. Furthermore, as Ricardian equivalence (see Barro, 1974) holds in the model, it does not matter for equilibrium allocations whether the government balances its debt or not in each period.

[g] See, eg, Leeper and Leith (2016) and Leeper et al. (2015).

3.4.1 Solving the Model

Before estimating the model, we log-linearize all the equations of the model. The log-linearized representation is provided in Appendix A. To solve the system of log-linearized equations, we use the code packages Dynare (see Adjemian et al. (2011) and RISE (see Maih (2015)) which provides an efficient and reliable implementation of the method proposed by Blanchard and Kahn (1980).

3.4.2 Data

We use seven key macroeconomic quarterly US time series as observable variables: the log difference of real GDP, real consumption, real investment and the real wage, log hours worked, the log difference of the GDP deflator and the federal funds rate. A full description of the data used is given in Appendix C. The solid blue line in Fig. 1 shows the data for the full sample, which spans 1965Q1–2014Q2.[h] From the figure, we see the extraordinary large fall in private consumption, which exceeded the fall during the recession in the early 1980s. The strains in the labor market are also evident, with hours worked per capita falling to a postwar bottom low in early 2010. Finally, we see that the Federal reserve cut the federal funds rate to near zero in 2009Q1 (the FFR is measured as an average of daily observations in each quarter). Evidently, the zero bound was perceived as an effective lower bound by the FOMC committee, and they kept it as this level during the crisis and adopted alternative tools to make monetary policy more accommodating (see, eg, Bernanke, 2013). Meanwhile, inflation fell to record lows and into deflationary territory by late 2009. Since then, inflation has rebounded close to the new target of 2% announced by the Federal Reserve in January 2012.

The measurement equation, relating the variables in the model to the various variables we match in the data, is given by:

$$Y_t^{obs} = \begin{bmatrix} \Delta\ \ln GDP_t \\ \Delta\ \ln CONS_t \\ \Delta\ \ln INVE_t \\ \Delta\ \ln W_t^{real} \\ \ln HOURS_t \\ \Delta\ \ln PGDP_t \\ FFR_t \end{bmatrix} = \begin{bmatrix} \ln Y_t - \ \ln Y_{t-1} \\ \ln C_t - \ \ln C_{t-1} \\ \ln I_t - \ \ln I_{t-1} \\ \ln(W/P)_t - \ \ln(W/P)_{t-1} \\ \ln L_t \\ \ln \Pi_t \\ \ln R_t \end{bmatrix} \approx \begin{bmatrix} \overline{\gamma} \\ \overline{\gamma} \\ \overline{\gamma} \\ \overline{\gamma} \\ \overline{l} \\ \overline{\pi} \\ \overline{r} \end{bmatrix} + \begin{bmatrix} \widehat{y}_t - \widehat{y}_{t-1} \\ \widehat{c}_t - \widehat{c}_{t-1} \\ \widehat{\imath}_t - \widehat{\imath}_{t-1} \\ \widehat{w}_t^{real} - \widehat{w}_{t-1}^{real} \\ l_t \\ \pi_t \\ \widehat{R}_t \end{bmatrix} \tag{16}$$

where ln and Δ ln stand for log and log-difference, respectively, $\overline{\gamma} = 100(\gamma - 1)$ is the common quarterly trend growth rate to real GDP, consumption, investment and wages, $\overline{\pi} = 100\pi$ is the quarterly steady state inflation rate and $r = 100\left(\beta^{-1}\gamma^{\sigma_c}(1+\pi) - 1\right)$ is the

[h] The figure also includes a red-dashed line, whose interpretation will be discussed in further detail within Section 4.

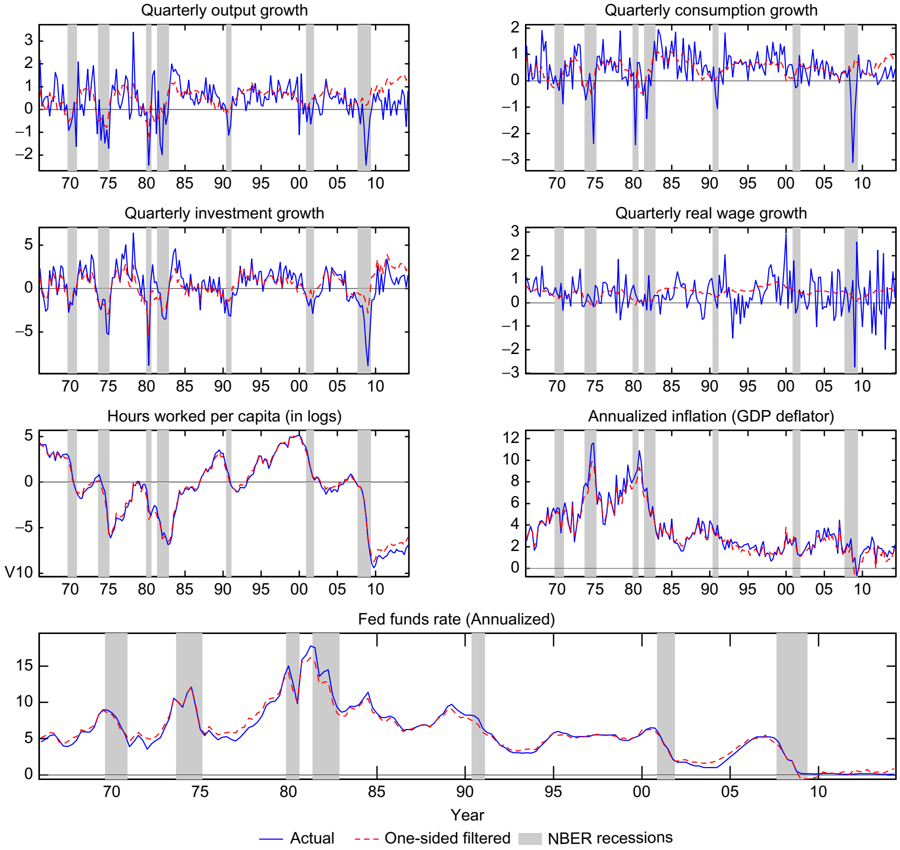


Fig. 1 Actual and filtered data in model estimated on precrisis data.

steady state nominal interest rate. Given the estimates of the trend growth rate and the steady state inflation rate, the latter will be determined by the estimated discount rate. Finally, $\bar{l}$ is steady state hours worked, which is normalized to be equal to zero.

Structural models impose important restrictions on the dynamic cross-correlation between the variables but also on the long run ratios between the macroaggregates. Our transformations in (16) impose a common deterministic growth component for all quantities and the real wage, whereas hours worked per capita, the real interest rate and the inflation rate are assumed to have a constant mean. These assumptions are not necessarily in line with the properties of the data and may have important implications for the estimation results. Some prominent papers in the literature assume real quantities to follow a stochastic trend, see, eg, Altig et al. (2011). Fisher (2006) argues that there is a stochastic trend in the relative price of investment and examines to what extent shocks that can explain this trend matter for business cycles. There is also an ongoing debate on whether hours worked per capita should be treated as stationary or not, see, eg, Christiano et al. (2003b), Galí and Pau (2004), and Boppart and Krusell (2015). Within the context of policy models, it is probably fair to say that less attention and resources have been spent to mitigate possible gaps in the low frequency properties of models and data, presumably partly because the jury is still out on the deficiencies of the benchmark specification, but also partly because the focus is on the near-term behavior of the models (ie, monetary transmission mechanism, forecasting performance, and historical decomposition) and these shortcomings do not seriously impair the model's behavior in this dimension.

3.4.3 Estimation Methodology

Following SW07, Bayesian techniques are adopted to estimate the parameters using the seven US macroeconomic variables in Eq. (16) during the period 1965Q1–2007Q4. Bayesian inference starts out from a prior distribution that describes the available information prior to observing the data used in the estimation. The observed data is subsequently used to update the prior, via Bayes' theorem, to the posterior distribution of the model's parameters which can be summarized in the usual measures of location (eg, mode or mean) and spread (eg, standard deviation and probability intervals).[i]

Some of the parameters in the model are kept fixed throughout the estimation procedure (ie, having infinitely strict priors). We choose to calibrate the parameters we think are weakly identified by the variables included in $\widetilde{Y}_t$ in (16). In Table 1, we report the parameters we have chosen to calibrate. These parameters are calibrated to the same values as had SW07.

The remaining 36 parameters, which mostly pertain to the nominal and real frictions in the model as well as the exogenous shock processes, are estimated. The first three columns in Table 2 shows the assumptions for the prior distribution of the estimated parameters. The location of the prior distribution is identical to that of SW07. We use the beta distribution for all parameters bounded between 0 and 1. For parameters assumed to be positive, we use the inverse gamma distribution, and for the unbounded parameters,

Table 1 Calibrated parameters

Parameter	**Description**	**Calibrated value**
δ	Depreciation rate	0.025
ϕ_w	Gross wage markup	1.50
g_y	Government G/Y ss-ratio	0.18
ϵ_p	Kimball curvature GM	10
ϵ_w	Kimball curvature LM	10

Note: The calibrated parameters are adapted from SW07.

[i] We refer the reader to Smets and Wouters (2003) for a more detailed description of the estimation procedure.

Table 2 Prior and posterior distributions: 1966Q1–2007Q4

		Prior distribution			Posterior distribution					
					Optimization		Metropolis chain			
Parameter		**Type**	**Mean**	**Std.dev. /df**	**Mode**	**Std.dev. Hess.**	**Mean**	**5%**	**95%**	**SW07 results Posterior mode**
Calvo prob. wages	ξ_w	Beta	0.50	0.10	0.79	0.055	0.75	0.61	0.82	0.73
Calvo prob. prices	ξ_p	Beta	0.50	0.10	0.69	0.051	0.69	0.60	0.76	0.65
Indexation wages	ι_w	Beta	0.50	0.15	0.63	0.136	0.58	0.36	0.79	0.59
Indexation prices	ι_p	Beta	0.50	0.15	0.23	0.093	0.26	0.13	0.44	0.22
Gross price markup	ϕ_p	Normal	1.25	0.12	1.64	0.076	1.64	1.52	1.77	1.61
Capital production share	α	Normal	0.30	0.05	0.21	0.018	0.20	0.18	0.24	0.19
Capital utilization cost	ψ	Beta	0.50	0.15	0.60	0.100	0.59	0.43	0.75	0.54
Investment adj. cost	φ	Normal	4.00	1.50	5.50	1.019	5.69	4.23	7.65	5.48
Habit formation	$\varkappa$	Beta	0.70	0.10	0.67	0.042	0.69	0.62	0.76	0.71
Inv subs. elast. of cons.	σ_c	Normal	1.50	0.37	1.53	0.138	1.44	1.23	1.69	1.59
Labor supply elast.	σ_l	Normal	2.00	0.75	2.15	0.584	2.03	1.13	2.99	1.92
Log hours worked in S.S.	$\bar{l}$	Normal	0.00	2.00	1.56	0.985	1.15	−0.56	2.72	−0.10
Discount factor	$100(\beta^{-1}-1)$	Gamma	0.25	0.10	0.13	0.052	0.16	0.08	0.25	0.16
Quarterly growth in S.S.	$\bar{\gamma}$	Normal	0.40	0.10	0.43	0.014	0.43	0.41	0.45	0.43
Stationary tech. shock	ρ_a	Beta	0.50	0.20	0.96	0.008	0.96	0.93	0.97	0.95
Risk premium shock	ρ_b	Beta	0.50	0.20	0.18	0.081	0.22	0.10	0.38	0.18
Invest. spec. tech. shock	ρ_i	Beta	0.50	0.20	0.71	0.053	0.71	0.61	0.80	0.71
Gov't cons. shock	ρ_g	Beta	0.50	0.20	0.97	0.008	0.97	0.96	0.98	0.97
Price markup shock	ρ_p	Beta	0.50	0.20	0.90	0.038	0.89	0.80	0.95	0.90
Wage markup shock	ρ_w	Beta	0.50	0.20	0.98	0.010	0.97	0.94	0.98	0.97
Response of g_t to ε_t^a	ρ_{ga}	Beta	0.50	0.20	0.52	0.086	0.49	0.38	0.67	0.52

Continued

Table 2 Prior and posterior distributions: 1966Q1–2007Q4—cont'd

		Prior distribution			Posterior distribution					SW07 results
					Optimization		Metropolis chain			
Parameter		**Type**	**Mean**	**Std.dev. /df**	**Mode**	**Std.dev. Hess.**	**Mean**	**5%**	**95%**	**Posterior mode**
Stationary tech. shock	σ_a	Invgamma	0.10	2.00	0.44	0.026	0.45	0.40	0.49	0.45
Risk premium shock	σ_b	Invgamma	0.10	2.00	0.24	0.022	0.24	0.19	0.27	0.24
Invest. spec. tech. shock	σ_i	Invgamma	0.10	2.00	0.41	0.041	0.41	0.34	0.48	0.45
Gov't cons. shock	σ_g	Invgamma	0.10	2.00	0.50	0.028	0.51	0.46	0.57	0.52
Price markup shock	σ_p	Invgamma	0.10	2.00	0.12	0.015	0.13	0.10	0.15	0.14
MA(1) price markup shock	ϑ_p	Beta	0.50	0.20	0.74	0.080	0.72	0.46	0.83	0.74
Wage markup shock	σ_w	Invgamma	0.10	2.00	0.31	0.025	0.30	0.25	0.34	0.24
MA(1) wage markup shock	ϑ_w	Beta	0.50	0.20	0.95	0.030	0.92	0.77	0.95	0.88
Quarterly infl. rate. in S.S.	$\overline{\pi}$	Gamma	0.62	0.10	0.79	0.114	0.82	0.65	1.01	0.81
Inflation response	r_π	Normal	1.50	0.25	2.01	0.174	2.07	1.75	2.33	2.03
Output gap response	r_y	Normal	0.12	0.05	0.10	0.023	0.10	0.05	0.13	0.08
Diff. output gap response	$r_{\Delta y}$	Normal	0.12	0.05	0.23	0.026	0.23	0.18	0.27	0.22
Mon. pol. shock std	σ_r	Invgamma	0.10	2.00	0.23	0.014	0.24	0.21	0.26	0.24
Mon. pol. shock pers.	ρ_r	Beta	0.50	0.20	0.12	0.062	0.15			0.12
Interest rate smoothing	ρ_R	Beta	0.75	0.10	0.82	0.022	0.82			0.81
Log marginal likelihood					Laplace	−961.81	MCMC	−960.72		

Note: Data for 1965Q1–1965Q4 are used as presample to form a prior for 1966Q1, and the log-likelihood is evaluated for the period 1966Q1–2007Q4. A posterior sample of 250, 000 postburn-in draws was generated in the Metropolis-Hastings chain. Convergence was checked using standard diagnostics such as CUSUM plots and the potential scale reduction factor on parallel simulation sequences. The MCMC marginal likelihood was numerically computed from the posterior draws using the modified harmonic estimator of Geweke (1999).

we use the normal distribution. The exact location and uncertainty of the prior can be seen in Table 2, but for a more comprehensive discussion of our choices regarding the prior distributions we refer the reader to SW07.

3.4.4 Posterior Distributions of the Estimated Parameters

Given these calibrated parameters in Table 1, we obtain the joint posterior distribution mode for the estimated parameters in Table 2 on precrisis data in two steps. First, the posterior mode and an approximate covariance matrix, based on the inverse Hessian matrix evaluated at the mode, is obtained by numerical optimization on the log posterior density. Second, the posterior distribution is subsequently explored by generating draws using the Metropolis-Hastings algorithm. The proposal distribution is taken to be the multivariate normal density centered at the previous draw with a covariance matrix proportional to the inverse Hessian at the posterior mode; see Schorfheide (2000) and Smets and Wouters (2003) for further details. The results in Table 2 shows the posterior mode of all the parameters along with the approximate posterior standard deviation obtained from the inverse Hessian at the posterior mode. In addition, it shows the mean along with the 5th and 95th percentiles of the posterior distribution, and finally, the last column reports the posterior mode in the SW07 paper.

There two important features to notice with regards to the posterior parameters in Table 2. First, the policy- and deep-parameters are generally very similar to those estimated by SW07, reflecting a largely overlapping estimation sample (SW07 used data for the 1965Q1–2004Q4 period to estimate the model). The only noticeable difference relative to SW07 is that the estimated degree of wage and price stickiness is somewhat more pronounced (posterior mode for ξ_w is 0.79 instead of 0.73 in SW07, and the mode for ξ_p has increased from 0.65 (SW07) to 0.69). The tendency of an increased degree of price and wage stickiness in the extended sample is supported by Del Negro et al. (2015b), who argue that a New Keynesian model similar to ours augmented with financial frictions points towards a high degree of price and wage stickiness to fit the behavior of inflation during the Great Recession. Second, the estimated variances of the shocks are somewhat lower (apart from the wage markup shock). Given that SW07 ended their estimation in 2004, and the so-called "Great Moderation" was still in effect from 2005 into the first half of 2007, the finding of reduced shock variances is not surprising.

4. EMPIRICAL PERFORMANCE OF BENCHMARK MODELS DURING THE GREAT RECESSION

We will now assess the performance of our benchmark DSGE model during the great recession in a number of dimensions. First and foremost, we study the forecasting performance of the model during the most intense phase of the recession, ie, the third and fourth quarters of 2008. In addition, we look into what the model has to say about the speed of recovery in the economy during the postcrisis period. In this exercise, we benchmark the performance of the DSGE model against a standard Bayesian VAR, which includes the same set of variables.

Second, we examine how the model interprets the "Great Recession", and assess the plausibility of the shocks the model needs to explain it. We do this from both a statistical and economic viewpoint.

4.1 Forecasting Performance of Benchmark Models During the Recession

We now use the DSGE model estimated on data up to 2007Q4 to forecast for the out-of-sample data. We start to make forecasts for 1, 2, …, 12 quarters ahead in the third and fourth quarter of 2008, conditional on observing data up to and including 2008Q3 and 2008Q4, respectively. Forecasts starting in these quarters are of particular interest as output plummeted in 2008Q4 (about -9.75% at an annualized quarterly rate) and in 2009Q1 (roughly -5.75% at an annualized rate). To provide a benchmark for the DSGE forecasts, we also report the forecasts of a Bayesian vector autoregressive (BVAR) model estimated on the same sample. While both models have been estimated for the same time series stated in Equation (16), we only show results for a subset of variables; the federal funds rate, output growth and price inflation (where inflation and output growth have been transformed into yearly rates by taking four-quarter averages). Warne et al. (2015) study how the predictive likelihood can be estimated, by means of marginalization, for any subset of the observables in linear Gaussian state-space models. Our exposition later is less formal and focuses on the univariate densities.[j]

The BVAR uses the standard Doan–Litterman–Sims (Doan et al., 1984) prior on the dynamics and an informative prior on the steady state following the procedure outlined in Villani (2009). We select the priors on the steady state in the BVAR to be consistent with those used in the DSGE model, which facilitates comparison between the two models. In both the DSGE and the BVAR, the median projections and 50%, 90%, and 95% uncertainty bands are based on 10, 000 simulations of respective model in which we allow for both shock and parameter uncertainty.[k]

In Fig. 2, the left column shows the forecasts in the DSGE conditional on observing data up to 2008Q3. As can be seen in the upper left panel, the endogenous DSGE model forecast predicted yearly GDP growth (four quarter change of log-output) to be about unchanged, whereas actual economic activity fell dramatically in the fourth quarter. Moreover, the 95% uncertainty band suggests that the large drop in output was

[j] We perform these forecasts on *ex post* data, collected on September 25th 2014 (see Appendix C).

[k] For an extensive comparison of the forecasting performance of the Smets and Wouters model along with a comparison to a BVAR and Greenbook forecasts on real-time data, see Edge and Gürkaynak (2010) and Wieland and Wolters (2013). Adolfson et al. (2007a, d) examine the forecasting properties of an open economy DSGE model on Swedish data.

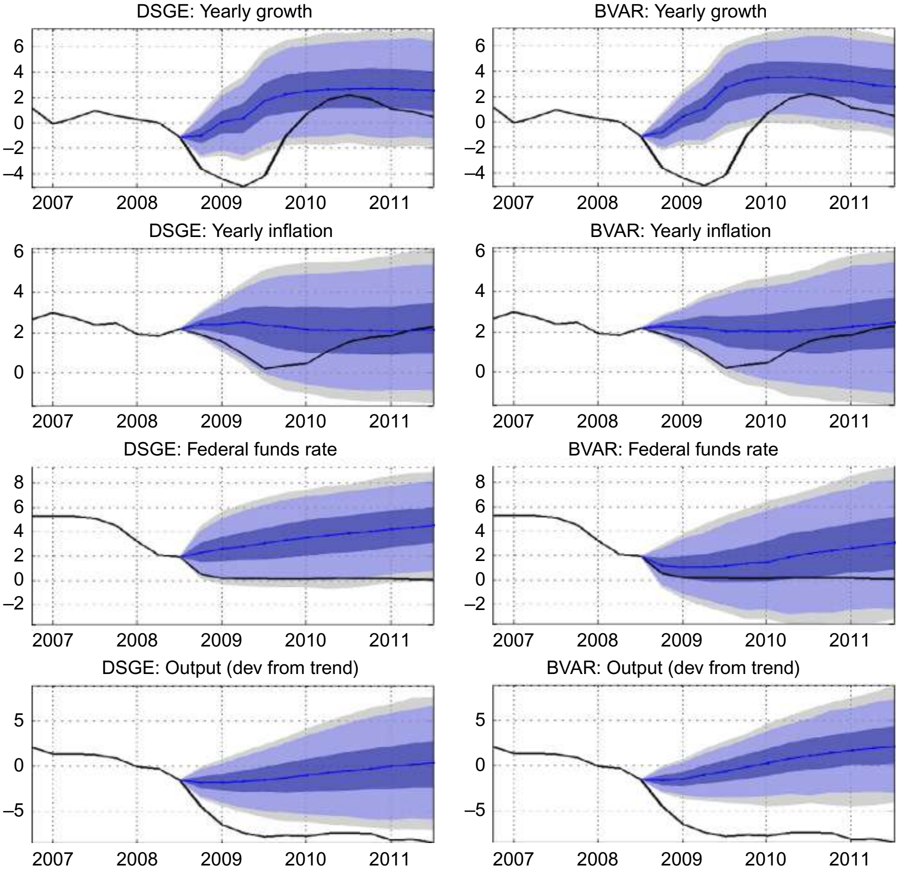


Fig. 2 Forecast 2008Q4–2011Q3 conditional on state in 2008Q3.

completely unexpected from the point of the view of the DSGE model. Thus, in line with Del Negro and Schorfheide (2013), our estimated model carries the implication that the "Great Recession" as late as of observing the outcome in 2008Q3 was a highly unlikely tail event. Turning to yearly inflation and the federal funds rate in the middle and bottom left panels, we also see that they fell considerably more than predicted by the model, but their decline are within or close to the 95% uncertainty bands of the linearized DSGE model and hence, cannot be considered as tail events to the same extent as the Great Recession.

Turning to the results for the BVAR, which are reported in the right column in Fig. 2, we see that the forecast distribution in the BVAR for yearly GDP growth is both quantitatively and qualitatively very similar to that in the DSGE model. Hence, the Great Recession was also a highly unlikely tail event according to the BVAR model. Given that the BVAR and the DSGE are both linearized models, the relatively high degree of similarity of the two model forecasts is not completely surprising. We also see that the uncertainty bands for the output roughly are equally sized in the DSGE as those in the BVAR model. This finding is neither obvious nor trivial as the DSGE model does not have a short-lag BVAR representation. The BVAR, on the other hand, does not impose nearly as many cross-restrictions on the parameter space as the DSGE model. Hence, allowing for parameter uncertainty will tend to increase the uncertainty bands considerably more in the BVAR relative to the DSGE model (the BVAR has around 190 free parameters, while the DSGE has 36). On net, these two forces appear to cancel each other out.

Moreover, as is clear from Fig. 3, the high degree of coherence between the DSGE and BVAR output growth forecasts also holds up when conditioning on the state in 2008Q4 and using the estimated models to make predictions for 2009Q1, 2009Q2, …, 2011Q4. For yearly inflation and the federal funds rate, the forecasts conditional on the state in 2008Q3 are very similar, as can be seen in the middle rows in Fig. 2.

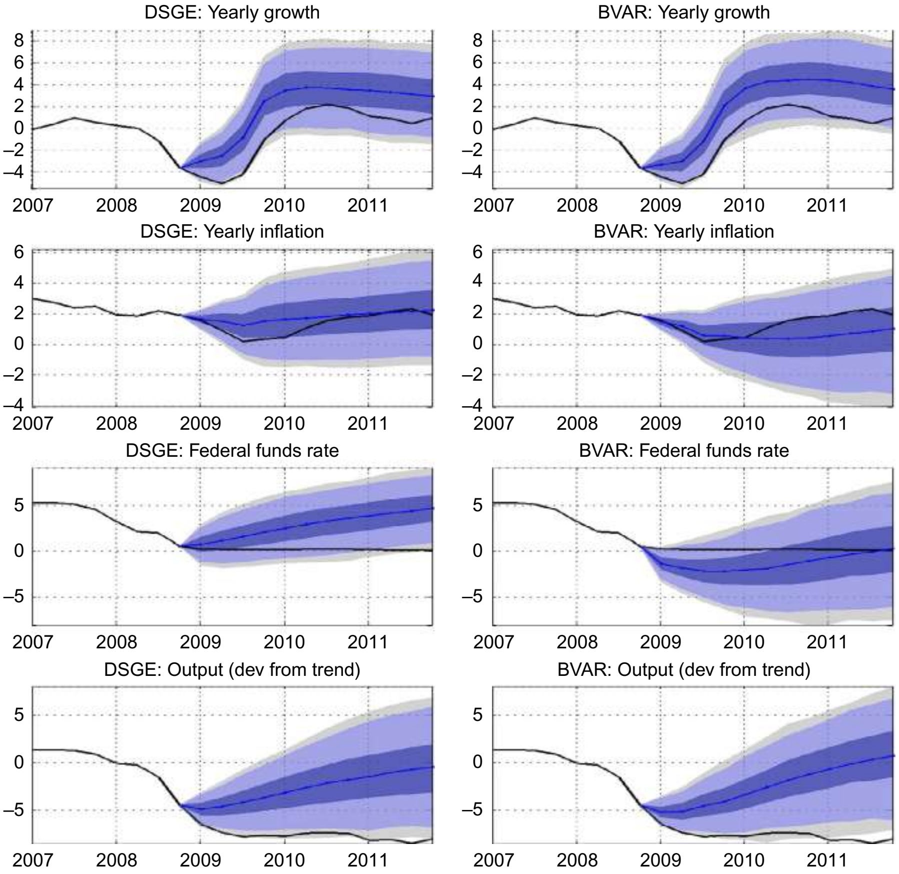


Fig. 3 Forecast 2009Q1–2011Q4 conditional on state in 2008Q4.

However, for the forecast made conditional on the state in 2008Q4 (Fig. 3), the DSGE and BVAR forecasts differ substantially, at least qualitatively. In this period, the BVAR predicts a prolonged period with near-zero inflation and a federal funds rate well below zero for 2 years, whereas the modal outlook in the DSGE model is that inflation would quickly return to near 2% and that the federal funds rate should therefore be increased steadily throughout the forecast horizon. The zero lower bound is not much of a concern in the DSGE model, while the BVAR suggests that it should be a binding constraint longer than 2 years.

Apart from failing to predict the crisis in the first place, both the BVAR and the DSGE model also have a clear tendency to forecast a quick recovery. For the benchmark DSGE model, this feature is evident already from Fig. 1. In this figure, the red-dotted line shows the one-sided filtered Kalman projections of the observed variables; that is, the projection for period t given all available information in period $t-1$. By comparing the one-sided filtered Kalman projections against the outcome (the blue-solid line) it is evident that the benchmark DSGE model predicts that growth in output, consumption and investment would pick up much quicker than they did following the recession. Hence, consistent with the findings in Chung et al. (2012), the benchmark DSGE model consistently suggests a V-shaped recovery and that better times were just around the corner, whereas the outcome is consistent with a much more slower recovery out of the recession as is evident from Figs. 2 and 3. Fig. 4 shows sequential BVAR forecasts 1, 2, …, 12 quarters ahead for the period 2008Q3–2014Q1 conditional on observing the state up to the date in which the forecasts start. In line with the results for the DSGE model, the results in this figure indicate that the BVAR also tends to predict a quick recovery of economic activity. Consistent with this reasoning, the forecasts for the level of output (as deviation from the deterministic trend), shown in the bottom row in Figs. 2 and 3, display that both the DSGE and the BVAR models overestimate the speed of recovery out of the recession.[l]

The slow recovery following the recession is consistent with the work by Reinhart and Rogoff (2009) and Jordà et al. (2012), who suggest that recoveries from financial crises are slower than recoveries from other recessions. The empirical observation by Reinhart and Rogoff has also been corroborated in subsequent theoretical work by Queralto (2013) and Anzoategui et al. (2015).[m] As our benchmark equilibrium model does not include the mechanisms of Queralto, it has a hard time accounting for the slow recovery following the recession, both in terms of the level and the growth rate of GDP. Our benchmark models—both the DSGE and the BVAR—rely on significant influence of adverse

[l] For both the BVAR and the DSGE model, the series for detrended output is the smoothed estimate from the DSGE model. When we construct the forecast of detrended output in the BVAR, we accumulate the projected quarterly growth rate of output after subtracting the estimated steady state growth rate in each period.

[m] Notwithstanding these results, Howard et al. (2011) argue out that the finding pertains to the level of economic activity, and not the growth rate (which is what we focused on in Fig. 1).

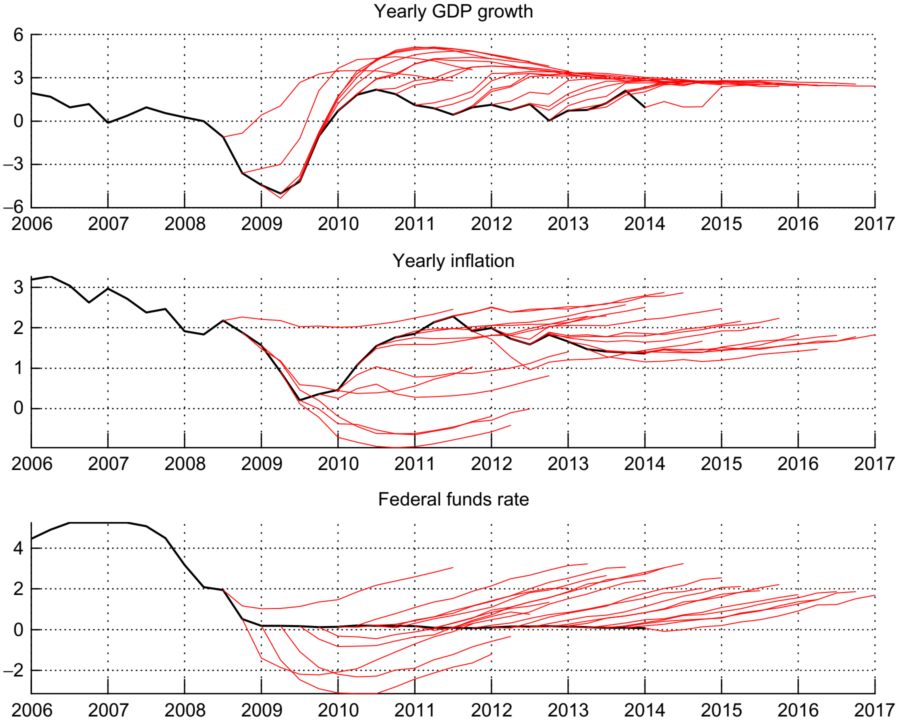


Fig. 4 Sequential BVAR forecasts 2008Q3–2014Q1.

exogenous shocks which weighs on economic activity during the recovery. While this might be deemed to be a significant weakness of these models, it should be noted that some major negative events may have contributed to hold back the recovery; eg, the European debt crisis which intensified in May 2010, and the showdown between the Republicans and democrats in the congress which created significant uncertainty in the US economy according to estimates by Fernández-Villaverde et al. (2011). With these events in mind, it is not entirely implausible that the models need some adverse shocks to account for the slow recovery.

4.2 Economic Interpretation of the Recession

As indicated in the previous section, both the DSGE and BVAR models are dependent on major adverse shocks to account for the recession. In this section, we examine what shocks are filtered out as the drivers of the recession and its aftermath. We will focus entirely on the benchmark DSGE, as it would be hard to identify all the shocks in the BVAR model. We extract the smoothed shocks through the Kalman filter by using the model estimated on the precrisis period for the full sample (without reestimating the parameters).

In Fig. 5, the left column shows the two-sided smoothed Kalman filtered innovations—eg, η_t^a for the technology shock in Eq. (5)—for the seven shock processes in

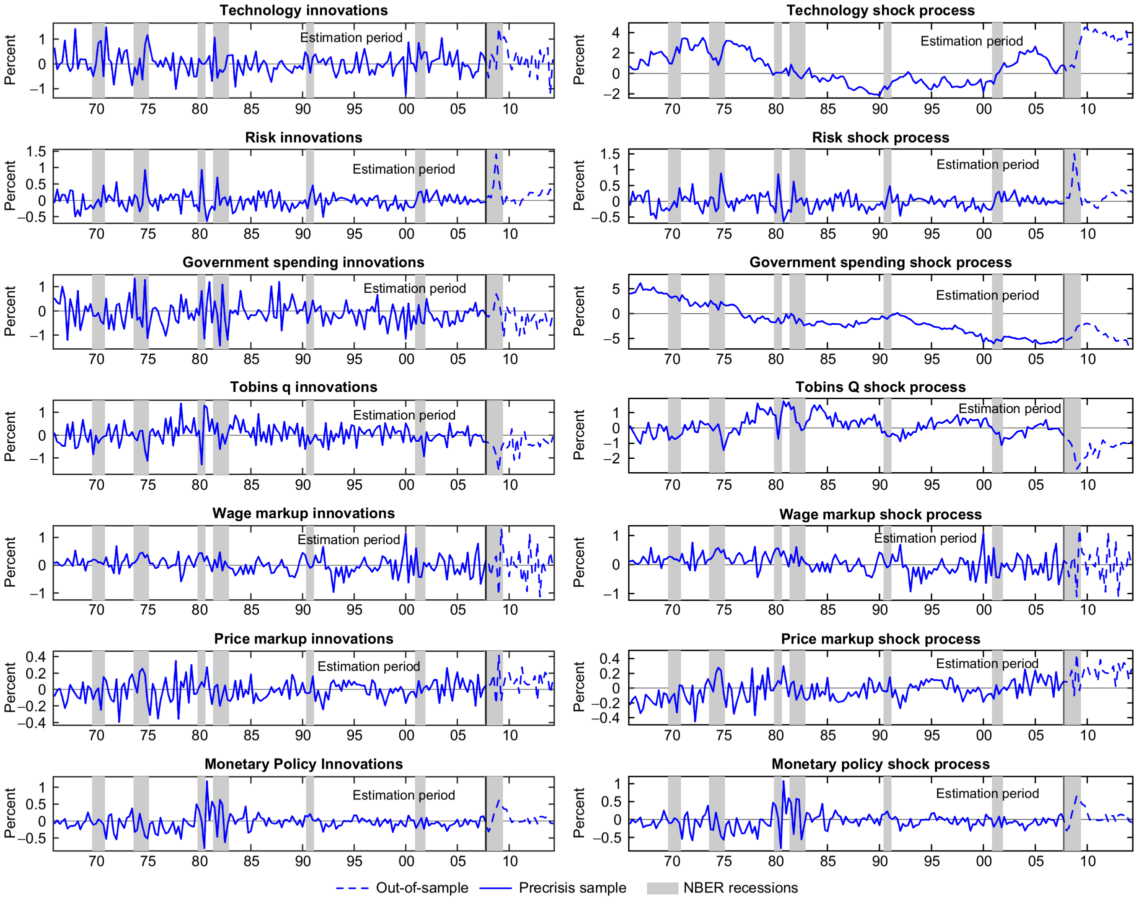


Fig. 5 Smoothed innovations and shocks in model estimated on precrisis data.

the model using the posterior mode parameters. In the right column, we show the two-sided smoothed shock processes in levels—eg, ε_t^a for the technology shock in Eq. (5). The blue solid-line indicates the in-sample period, and the blue-dotted line the out-of-sample period. The grey bars are NBER dated recessions.

Before analyzing the role various shocks played during the crisis and its aftermath, it is insightful to discuss if there are any signs in the precrisis shocks about what events might have been causal for the crisis itself. As is clear from the left column in the figure, there is nothing that stands out in the innovations between 2000 and the burst of the crisis. There were a string of positive innovations to technology during 2003–2005, which led to a run-up in technology (right upper panel) during this period. To the extent that households and firms expected this positive development to continue and were taken off-guard by the adverse outcomes 2006 and onward, this could have been a contributing factor to the crisis. Christiano et al. (2010b) argue that over-optimistic expectations of future technology have been associated with credit cycles that have contributed to boom-bust cycles in the real US economy in a model with a more elaborate financial sector.[n] Our benchmark model does not include a financial sector and thus, cannot be used to assess this possibility explicitly. Loose monetary policy have also been argued as a possible driver for the crisis, see, eg, Taylor (2007). Our estimated model lend some, but limited, support to this view; although the estimated policy rule suggest that monetary policy was on average expansionary between 2002 and 2006, the magnitude of the deviations are not very large though, as seen from the lower panels in Fig. 5. Based on the shock decomposition, it is therefore hard to argue that the Fed's conduct of monetary policy was causal for the crisis.[o]

With this discussion in mind, we now turn to the crisis and its aftermath. As is clear from Fig. 5, the key innovations happened to technology, investment specific technology (the Tobin's Q-shock), and the risk-premium shock during the most intense phase of the recession. More specifically, the model filters out a very large positive shock to technology (about 1.5% as shown in the upper left panel, which corresponds to a 3.4 standard error shock) in 2009Q1. In 2008Q4 and 2009Q1, the model also filters out two negative investment specific technology shocks (about −1 and −1.5%—or 2.0 and 3.7 standard errors—respectively). The model moreover filters out a large positive risk shocks in

[n] The focus of Christiano et al. is what monetary policy should do to mitigate the inefficient boom-bust cycle. They do not consider the role macroprudential regulation could play to mitigate the cycle.

[o] The main reason why our policy shocks are much smaller in magnitude than those computed by Taylor is that we consider a more elaborate policy rule with considerable interest rate smoothing ($\rho_R = 0.82$, see Table 2). One could argue about whether one should allow for interest rate smoothing or whether this persistence should be attributed to the exogenous monetary policy shock (ie, a higher ρ_r in the process for ε_t^r in Eq. (15). In our estimated model, however, the log marginal likelihood strongly favors a high degree of interest rate smoothing and low persistence of the exogenous policy shocks (ie, a combination of high ρ_R and low ρ_r).

2008Q3–Q4, and in 2009Q1 (0.5%, 1.5%, and 0.5%, respectively, equivalent to 1.9, 6.0, and 2.8 standard errors). These smoothed shocks account for the bulk of the sharp decline in output, consumption and investment during the acute phase of the crisis at the end of 2008 and the beginning of 2009. Our finding of a large positive technology shock in the first quarter of 2009 may at first glance be puzzling, but can be understood from Figs. 1 and 3. In these, we see that output (as deviation from trend) fell less during the recession than did hours worked per capita. Hence, labor productivity rose sharply during the most acute phase of the recession. The model replicates this feature of the data by filtering out a sequence of positive technology shocks. These technology shocks will stimulate for output, consumption and investment. The model thus needs some really adverse shocks that depresses these quantities even more and causes hours worked per capita to fall, and this is where the positive risk premium and investment specific technology shocks come into play. These shocks cause consumption (risk premium) and investment (investment specific)—and thereby GDP—to fall. Lower consumption and investment also causes firms to hire less labor, resulting in hours worked per capita to fall.

Another shock that helps account for the collapse in activity at the end of 2008 is the smoothed monetary policy shock shown in the bottom left panel (expressed at a quarterly rate). This shock becomes quite positive in 2008Q4 and 2009Q1; in annualized terms it equals roughly 150 (1.6 standard errors) and 250 (2.8 standard errors) basis points in each of these quarters, respectively. As the actual observations for the annualized federal funds rate is about 50 and 20 basis points, these sizable policy shocks suggests that the zero lower bound is likely to have been a binding constraint, at least in these quarters. This finding is somewhat different from those of Del Negro and Schorfheide (2013) and Del Negro et al. (2015b), who argued that the zero lower bound was not a binding constraint in their estimated models.

The large smoothed innovations translate into very persistent movements in some of the smoothed shock processes, reported in the right column in Fig. 5. For the simple AR(1) shock processes, the degree of persistence is governed by the posterior for ρ. As can be seen from Table 2, the posterior for $\rho_a(\rho_b)$ is very high (low), whereas the posterior for ρ_i is somewhere in between. It is therefore not surprising that the technology process is almost permanently higher following the crisis, whereas the risk shock process quickly recedes towards steady state. Our finding of a very persistent rise in the exogenous component of total factor productivity (TFP) is seemingly at odds with Christiano et al. (2015), who reports that TFP fell in the aftermath of the recession. Christiano et al. (2015) and Gust et al. (2012) also report negative innovations to technology in 2008 (see fig. 5 in their paper). While a closer examination behind the differences in the results would take us too far, we note that our findings aligns very well with Fernald (2012). Specifically, our smoothed innovations to technology are highly correlated with the two TFP measures computed by Fernald (2012), as can be seen from Table 3. The table shows the correlations between our technology innovations η_t^a, shown in the left column

Table 3 Correlations between smoothed and actual TFP shocks

	Sample period	
TFP measure	**Precrisis: 66Q1–07Q4**	**Full: 66Q1–14Q2**
$\text{Corr}\left(\Delta\text{Raw}, \eta_t^a\right)$	0.483	0.522
$\text{Corr}\left(\Delta\text{Corrected}, \eta_t^a\right)$	0.602	0.608

Note: "ΔRaw" denotes the first difference of the quarterly unadjusted measure in Fernald (2012), while "ΔCorrected" is the first difference Fernald's capacity utilization adjusted TFP measure. In the model, the smoothed estimates of the innovations η_t^a (see Eq. (5)) are used. This series is depicted in the upper left column of Fig. 5.

in Fig. 5, and the period-by-period change in the raw and utilization-corrected measure of TFP by Fernald. From the first column in the table, we learn that the correlation between our innovations and his raw measure is almost 0.5 for the estimation sample period. As we are studying first differences and innovations, this correlation must be considered quite high. Even more reassuring for our model is that the correlation between our smoothed innovation series and Fernald's utilization adjusted series is as high as 0.6. When extending the sample to include the crisis and postcrisis period, we see that these correlations remain high; if anything, they become slightly higher. We believe this lends support for our basic result that weak TFP growth was not a key contributing factor to the crisis.

For the two markup shocks, we notice that they are not nearly as highly correlated as the technology shock although the estimated AR(1) coefficients for these processes are quite high (0.89 for the price markup shock, and 0.97 for the wage markup shock, see Table 2). The reason why their correlation is so low is the estimated MA(1) coefficients, ϑ_p and ϑ_w in Eqs. (6) and (13) are rather high, ie, 0.72 and 0.92, respectively. Despite the generally low correlation of the price shock process during the precrisis period, we see that its outcome is driven by a sequence of positive innovations during the crisis period. This finding is in line with Fratto and Uhlig (2014), who found that price markup shocks played an important role to avoid an even larger fall in inflation during the crisis, and contributed to the slow decline in employment during the postcrisis recovery.[p] The wage markup shock process does not display any clear pattern after the precrisis period, but it is clear that its variance has increased since the end of the 1990s suggesting that the model provides a less accurate description of wage-setting behavior in the US labor market since

[p] The prominent role of the price and wage markup for explaining inflation and behavior of real wages in the SW07-model have been criticized by Chari et al. (2009) as implausibly large. Galí et al. (2011), however, shows that the size of the markup shocks can be reduced substantially by allowing for preference shocks to household preferences.

then. However, it should be kept in mind that this finding may not necessarily remain if alternative wage series are used.[q]

The historical decompositions in Fig. 6 summarizes the impact of the various shocks on the output growth, inflation, federal funds rate and output as deviation from a trend during 2007Q1–2014Q2 in the benchmark model estimated on data up to 2007Q4 (see Table 2). Notice that the scale on the left- and right-axes are not the same (except for the two-sided-smoothed output as deviation from trend): the left axis shows the

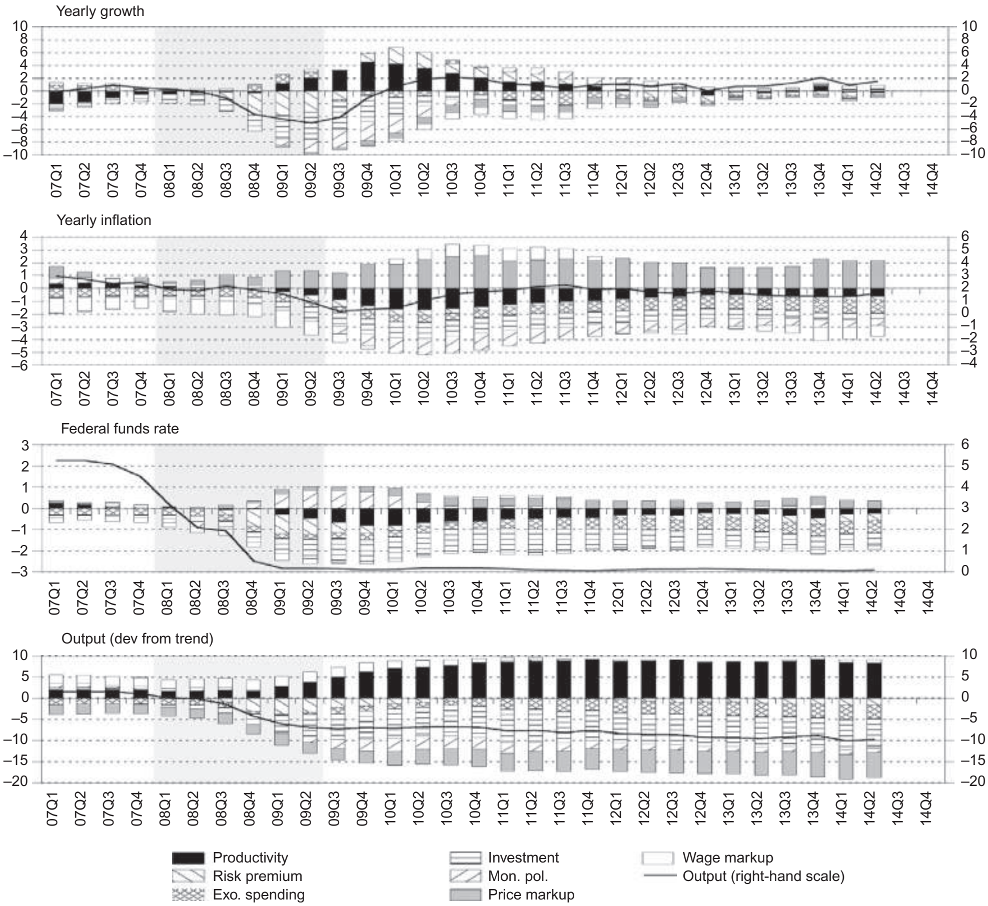


Fig. 6 Historical decompositions of yearly output growth (four-quarter change), yearly inflation (four-quarter change), fed funds rate, and output (deviation from trend). Left axis shows the contributions of the shocks (bars) to fluctuations around the steady state and right axis shows actual outcomes (in levels).

[q] Because of potential measurement problems pertaining to Galí et al. (2011) and Justiniano et al. (2013b) use two series for real wage growth when estimating their DSGE model.

contributions of the various shocks to fluctuations around the steady state, whereas the right axis shows evolution of each variable in levels. Thus, for each period the sum of the bars on the left axis plus the steady state value for each variable (not shown) equals the actual outcome (thin line). For output as deviation from trend, the steady state value is nil, why the sum of the bars directly equals the smoothed values.

As seen from the figure, the risk premium, the investment specific technology and the monetary policy shocks are the key drivers behind the decline in output during the recession period, whereas TFP as discussed earlier had some offsetting impact on output. However, all four shocks contributed to the gradual decline in inflation. The nominal interest rate would clearly have dropped below zero in absence of the zero bound constraint. The slow recovery is attributed to the persistence of the shocks that were responsible for the recession, but also captures new unexpected headwinds along with positive innovations to markups in prices and wages. Interestingly, the negative impact of the risk premium shock is relatively short lived. To a large extent this of course reflects that the model is not rich enough to propagate financial shocks sufficiently, but it is also conceivable that this partly captures the stimulus coming from the nonconventional monetary policy actions. The continuously low interest rate is consistent with the weak state of the economy during this period; output (as deviation from trend) is well below its precrisis trend and inflation persistently below its targeted rate, and sustained subpar growth (slow or nonexistent recovery in output as deviation from trend). As the precrisis model features a moderate degree of price and wage stickiness, inflation would have fallen persistently into negative territory in the absence of other shocks. This is counter-factual relative to the data, and the missing deflation in the model estimated on precrisis data is accounted for by inflationary markup shocks.

While the smoothed shocks—that the model needs to explain the crisis period—are not too surprising given the model's specification, it is nevertheless clear that the benchmark model needs a highly unlikely combination of adverse shocks in 2008Q4 and 2009Q1 to account for the most intense phase of the recession. Therefore, we now discuss the statistical properties of the shocks and examine if they correlate with some key observable financial variables not included in our set of observables.

4.3 Statistical Properties of the Innovations and Their Relation to Financial Indicators

Table 4 provides an overview of the statistical properties of the estimated structural shocks and of the forecast errors for the seven observed macro variables. Most of the forecast errors display a significant amount of kurtosis, a feature that they inherit from the underlying macro variables. For the structural shocks, the problems are mostly concentrated in two shocks—the monetary policy and the risk premium shock—that display highly significant deviations from the underlying Gaussian assumption. The structural innovations in the policy rate and the risk premium are characterized by a highly skewed

Table 4 Statistical distribution of innovations

	Sample period							
	Precrisis: 66Q1–07Q4				Full Sample: 66Q1–14Q2			
Innovations in	**Mean**	**Std**	**Skew**	**Kurt**	**Mean**	**Std**	**Skew**	**Kurt**
Technology	0.04	0.44	0.43*	4.09*	0.04	0.46	0.32	3.76
Risk premium	0.00	0.24	0.74**	5.12**	0.00	0.19	1.03**	7.08**
Inv. spec. techn.	0.02	0.42	0.09	3.95*	0.02	0.37	0.09	3.73
Exog. spending	−0.07	0.50	0.30	3.66	−0.07	0.49	0.25	3.65
Price markup	0.00	0.12	−0.14	3.49	0.00	0.12	0.01	3.62
Wage markup	0.01	0.31	0.10	3.89	0.01	0.37	0.03	4.48**
Monetary policy	−0.03	0.23	0.76**	8.09**	−0.04	0.23	0.80**	8.45**
Forecast errors in								
Output growth	−0.04	0.66	0.38*	5.05**	0.01	0.69	0.12	5.10**
Consumption growth	0.01	0.56	−0.42*	4.50**	0.08	0.62	−0.89**	6.77**
Investment growth	0.25	1.62	0.14	5.24**	0.25	1.73	−0.02	5.43**
Hours per capita	−0.04	0.53	0.03	4.25**	−0.02	0.55	−0.03	3.96*
Inflation	0.05	0.26	0.22	4.05*	0.04	0.25	0.30	4.14**
Real wage growth	−0.05	0.63	0.14	3.89	−0.04	0.73	−0.03	4.72**
Short rate	−0.01	0.24	1.29**	12.25**	−0.02	0.22	1.80**	15.31**

Note: *, ** indicate a significance at 5% and 1%, respectively.

and fat-tailed distribution.[r] We identified the large disturbances in these shocks already in the previous section as crucial drivers of the recent recession, but Table 4 illustrates that both processes were already affected by non-Gaussian innovations in the precrisis model as well. As observed in Fig. 5, these negative outliers occur mostly during the recession periods.

This feature implies that the predictive density of linear Gaussian DSGE models underestimates systematically the probability of these large recession events. This observation is important because it means that the model considers the strong economic downturns that we typically observe during recession periods as extremely unlikely tail events.[s] Linear Gaussian models may therefore be inappropriate instruments for analyzing policy questions related to risk scenario's or stress test exercises.

[r] The innovations in the structural shocks are also characterized by a significant ARCH effect illustrating the systematic time-varying volatility structures.

[s] This observation is consistent with the findings presented by Chung et al. (2012).

It is also interesting to note that the two structural shocks that generate most of the extreme events are directly related to the intertemporal decisions and to the developments in the monetary and the financial sector of the economy. The non-Gaussian nature of financial returns, spreads and risk premiums is widely documented in the financial literature. Therefore, it appears like a natural hypothesis to assume that the non-Gaussian shocks that are identified in our macro model reflect the influence—or the feedback—from financial disruptions to the rest of the economy. To support this argument, we calculate the correlations between our estimated structural innovations and a set of popular financial returns and spreads. We selected seven measures related to the different segments of the financial sector and for which long time series are available: the Baa-Aaa spread, the term spread, the Ted spread, the return on the S&P index, the return on the Fama-French financial sector portfolio, the change in the Shiller house price index and the VOX index. Table 5 summarizes the correlation between these seven financial indicators and our seven structural innovations. The strongest correlations in this table—exceeding 0.3 in absolute terms—are observed between our identified risk premium innovation and the Baa-Aaa and Term spreads, and between the monetary policy innovation and the Term and Ted spreads.

To see the strong linkages between some of the smoothed shocks and the financial variables in an alternative way, we regress the structural innovations on this set of financial

Table 5 Correlation between innovations and financial indicators

		Innovations						
Innovations in		σ_a	σ_b	σ_i	σ_g	σ_p	σ_w	σ_r
Technology	σ_a	1.00						
Risk premium	σ_b	−0.11	1.00					
Inv. spec. techn.	σ_i	−0.19	−0.08	1.00				
Exog. spending	σ_g	0.01	0.27	−0.06	1.00			
Price markup	σ_p	−0.03	0.18	0.05	0.13	1.00		
Wage markup	σ_w	0.00	−0.01	−0.07	−0.21	−0.09	1.00	
Monetary policy	σ_m	0.09	−0.17	−0.05	0.17	−0.05	−0.04	1.00

	Innovations						
Financial Indicator	σ_a	σ_b	σ_i	σ_g	σ_p	σ_w	σ_r
Baa-Aaa	−0.10	0.39	−0.21	0.28	0.04	−0.02	0.04
Term spread	0.11	0.33	−0.11	−0.04	−0.07	0.10	−0.46
Ted spread	−0.20	−0.13	0.13	0.18	0.14	−0.02	0.34
Return S&P	0.14	−0.24	0.18	−0.20	−0.13	0.02	−0.13
Return Fin	0.02	0.03	0.01	−0.05	−0.14	0.02	−0.10
Return HP	−0.07	−0.07	0.25	−0.06	0.00	0.02	−0.14
VOX	−0.12	0.10	0.03	0.13	0.09	0.01	−0.05

Note: The data sources are provided in Appendix C.

Table 6 Regression analysis of innovations and financial indicators

	Precrisis sample			Full sample		
Innovations in	σ_b	σ_i	σ_r	σ_b	σ_i	σ_r
Contemporaneous impact from financial indicator on innovations						
Baa-Aaa	0.29*	−0.57*	0.09	0.28*	−0.26*	−0.02
Term spread	0.10*	−0.05	−0.18*	0.09*	−0.02	−0.18*
Ted spread	−0.09*	0.16*	0.15*	−0.08*	0.12*	0.14*
Return S&P	−0.64	1.51*	−0.27	−0.70*	1.37*	−0.45
Return Fin	0.46*	−0.24	−0.05	0.33*	−0.22	−0.01
Return HP	1.35	5.41*	−3.52	0.10	4.67*	−2.63
VOX	0.38	0.00	−0.44	0.34	0.67	−0.61
F/p-value	7.00/0.00	4.45/0.00	15.80/0.00	11.79/0.00	4.86/0.00	14.31/0.00
Skew/kurt resid	0.04/2.97	0.17/3.22	0.60/4.39	0.15/3.11	0.14/3.15	0.57/4.08
Granger Causality regressions						
F/p-value	1.73/0.06	1.53/0.11	2.11/0.01	1.67/0.06	2.05/0.02	1.62/0.07

Note: * indicates significance at 5%. The financial indicators do not have a significant effect on the other nonreported innovations.

observables. The results of these multivariate regressions are shown in Table 6. In contemporaneous regressions, the significant coefficients are again only apparent in the risk premium, monetary policy and—at a slightly weaker significance level—for the investment specific technology innovation. The most interesting feature of the regression results is that the remaining unexplained variation (ie, the regression residuals) are basically normally distributed. Thus, shock outliers seem to coincide with periods of clear financial stress as measured by our observed financial indicators. Also noteworthy is that in Granger causality regression tests, none of the financial indicators carry significant predictive power for the structural innovations. Because financial variables can essentially be observed in real time; however, they can still provide timely indications of big structural innovations. Including these variables in our list of observables can therefore be very useful to improve the model now-cast and the conditional forecast performance.[t] Even so, this strategy will probably not improve the out-of-sample prediction performance of our linearized models *ex ante* to the observation of financial stress signals. It might also require non-Gaussian and nonlinear models to exploit this information from financial variables more efficiently in our macro models.

[t] See Del Negro and Schorfheide (2013) for strong evidence in this direction.

5. AUGMENTING THE BENCHMARK MODEL

As the analysis in Section 4 suggested that the benchmark model suffers from some important shortcomings, we study in this section to which extent its performance can be improved by allowing for zero lower bound on policy rates, time-varying volatility of the shocks, and by introducing financial frictions and a cost-channel into the model. The modeling of financial frictions follows the basic approach in the seminal work of Bernanke et al. (1999). In contrast to the analysis in Section 3.4, we estimate the different perturbations of the model on data including the crisis period in this section.

5.1 Assessing the Impact of the Zero Lower Bound

We assess the impact of imposing the zero lower bound (ZLB) in the estimation in two alternative ways. These procedures differ in the way the duration of the ZLB spells is determined. In our first approach, the incidence and duration of the ZLB spells are endogenous and consistent with the model expectations. In the second approach, we model them as "exogenous" and require the model to match information from the market-based overnight index swap rates following Del Negro et al. (2015b). In both approaches, we make use of the same linearized model equations (stated in Appendix A), except that we impose the nonnegativity constraint on the federal funds rate. To do this, we adopt the following policy rule for the federal funds rate

$$\begin{aligned}\widehat{R}_t^* &= \rho_R \widehat{R}_{t-1} + (1-\rho_R)(r_\pi \widehat{\pi}_t + r_y(\widehat{ygap}_t)) + r_{\Delta y}\Delta(\widehat{ygap}_t) \quad , \\ \widehat{R}_t &= \max\left(-\bar{r}, \widehat{R}_t^* + \widehat{\varepsilon}_t^r\right).\end{aligned} \tag{17}$$

The policy rule in (17) assumes that the interest rate set by the bank, $\widehat{R}_t$, equals $\widehat{R}_t^* + \widehat{\varepsilon}_t^r$ if unconstrained by the ZLB. $\widehat{R}_t^*$, in turn, is a shadow interest rate that is not subject to the policy shock $\widehat{\varepsilon}_t^r$. Note that $\widehat{R}_t$ in the policy rule (17) is measured as percentage point deviation of the federal funds rate from its quarterly steady state level ($\bar{r}$), so restricting $\widehat{R}_t$ not to fall below $-\bar{r}$ is equivalent to imposing the ZLB on the nominal policy rate.[u] In its setting of the shadow or notional rate we assume that the Fed is smoothing over the lagged actual interest rate, as opposed to the lagged notional rate $\widehat{R}_{t-1}^*$. We made this assumption to preserve the property that $\widehat{\varepsilon}_t^r$ is close to white noise. Smoothing over the notional rate in (17) would cause the policy shock to become highly persistent, with an AR(1) coefficient roughly equal to ρ_R.[v]

[u] See (16) for the definition of $\bar{r}$. If writing the policy rule in levels, the first part of (17) bee replaced by (14) (omitting the policy shock), and the ZLB part would bee $R_t = \max\left(1, R_t^* \varepsilon_t^r\right)$.

[v] To see this, replace $\hat{R}_{t-1}$ with $\hat{R}_{t-1}^*$ in the first equation in (17) and then substitute $\hat{R}_t = \hat{R}_t^* + \hat{\varepsilon}_t^r$ from the second equation to write the unconstrained policy rule with the actual policy rate $\hat{R}_t$. Then, the residual will be $\hat{u}_t^r \equiv \hat{\varepsilon}_t^r - \rho_R \hat{\varepsilon}_{t-1}^r$. Hence, the residual $\hat{u}_t^r$ will be roughly white noise in this case when $\hat{\varepsilon}_t^r$ has an AR (1)-root ρ_R.

To impose the policy rule (17) when we estimate the model, we use the method outlined in Hebden et al. (2010). This method is convenient because it is quick even when the model contains many state variables, and we provide further details about the algorithm in Appendix A.[w] In a nutshell, the algorithm imposes the nonlinear policy rule in Eq. (17) through current and anticipated shocks (add factors) to the policy rule. More specifically, if the projection of $\widehat{R}_{t+h}$ in (17) given the filtered state in period t in any of the periods $h = 0, 1, \ldots, T$ for some sufficiently large nonnegative integer T is below $-\bar{r}$, the algorithm adds a sequence of anticipated policy shocks $\widehat{\varepsilon}^{r}_{t+h|t}$ such that $\mathrm{E}_t\widehat{R}_{t+h} \geq 0$ for all $h = \tau_1, \tau_1 + 1, \ldots, \tau_2$. If the added policy shocks put enough downward pressure on the economic activity and inflation, the duration of the ZLB spell will be extended both backwards (τ_1 shrinks) and forwards (τ_2 increases) in time. Moreover, as we think about the ZLB as a constraint on monetary policy, we further require all current and anticipated policy shocks to be *positive* whenever $\widehat{R}^{*}_{t} < -\bar{r}$. Imposing that all policy shocks are strictly positive whenever the ZLB binds, amounts to think about these shocks as Lagrangian multipliers on the nonnegativity constraint on the interest rate, and implies that we should not necessarily be bothered by the fact that these shocks may not be normally distributed even when the ZLB binds for several consecutive periods $t, t + 1, \ldots, t + T$ with long expected spells each period (h large).

We will subsequently refer to this method as "Endogenous ZLB duration", as it implies that both the incidence and the duration of the ZLB is endogenous determined by the model subject to the criterion to maximize the log marginal likelihood. In this context, it is important to understand that the nonnegativity requirement on the current and anticipated policy shocks for each possible state and draw from the posterior, forces the posterior itself to move into a part of the parameter space where the model can account for long ZLB spells which are contractionary to the economy. Without this requirement, DSGE models with endogenous lagged state variables may experience sign switches for the policy shocks, so that the ZLB has a stimulative rather than contractionary impact on the economy even for fairly short ZLB spells as documented by Carlstrom et al. (2012).[x] As discussed in further detail in Hebden et al., the nonnegativity assumption for all states and draws from the posterior also mitigates the possibility of multiple equilibria (indeterminacy). Finally, it is important to point our that when the ZLB is not a binding constraint, we assume the contemporaneous policy shock $\widehat{\varepsilon}^{r}_{t}$ in Eq. (17) can be either negative or positive; in this case we do not use any anticipated policy shocks as monetary policy is unconstrained.

[w] Iacoviello and Guerrieri (2015) have subsequently shown how this method can be applied to solve DSGE models with other types of asymmetry constraints.

[x] This can be beneficial if we think that policy makers choose to let the policy rate remain at the ZLB although the policy rule dictated that the interest rate should be raised ($\hat{R}^{*}_{t}$ is above $-\bar{r}$). In the case of the United States, this possibility might be relevant in the aftermath of the crisis and we therefore subsequently use an alternative method which allows for this.

However, a potentially serious shortcoming of the method we adapt to assess the implications of the ZLB is that it relies on perfect foresight and hence does not explicitly account for the role of future shock uncertainty as in the work of Adam and Billi (2006) and Gust et al. (2012). Even so, we implicitly allow for parameter and shock uncertainty by requiring that the filtered current and anticipated policy shocks in each time point are positive for all parameter and shock draws from the posterior whenever the ZLB binds. More specifically, when we evaluate the likelihood function and find that $\mathrm{E}_t\widehat{R}_{t+h}<0$ in the modal outlook for some period t and horizon h conditional on the parameter draw and associated filtered state, we draw a large number of sequences of fundamental shocks for $h=0, 1, \ldots, 12$ and verify that the policy rule (17) can be implemented for all possible shock realizations through positive shocks only. For those parameter draws this is not feasible, we add a smooth penalty to the likelihood which is set large enough to ensure that the posterior will satisfy the constraint.[y] As we document below, the nonnegativity constraint on the anticipated policy shocks in the face of parameter and fundamental shock uncertainty has considerable implications for the estimation of the model, and shock and parameter uncertainty is therefore partly accounted for in our estimation procedure.[z]

To provide a reference point for the ZLB estimations we start out by estimating the model for the full sample period, but disregarding the existence of the ZLB. The posterior mode and standard deviation in this case are shown in the first two columns in Table 7, and labeled “No ZLB model”. The only difference between these results and those reported in Table 2 is that the sample period has been extended from 2007Q4 to 2014Q2. By comparing the results, a noteworthy difference is that the estimated degree of wage and price stickiness has increased even further relative to the precrisis sample. The posterior mode for the sticky wage parameter (ξ_w) has increased from 0.79 to 0.83, and the sticky price parameter (ξ_p) from 0.69 to 0.75. Relative to the SW07 posterior mode, ξ_w has increased from 0.73 to 0.83 and ξ_p from 0.65 to 0.75. These increases are substantial, considering that the sample has been expanded with less than 10 years and that these parameters affect the slope of the wage and price pricing curves in a nonlinear fashion, implying an even sharper reduction in the slope coefficients for the forcing variables

[y] For example, it turns out that the model in 2008Q4 implies that the ZLB would be a binding constraint in 2009Q1 through 2009Q3 in the modal outlook. For this period we generated 1000 shock realizations for 2009Q1,2009Q2, …, 2011Q4 and verified that we could implement the policy rule (17) for all forecast simulations of the model through nonnegative current and anticipated policy shocks. For the draws with adverse shocks, the duration of the ZLB was prolonged substantially during the forecast horizon, with expected ZLB spells close to 4 years occurring. We provide further details in Appendix B how the likelihood function is constructed when we impose the ZLB in the estimations.

[z] Alternatively, we could implement this type of restriction by using a stochastic filter in which the prediction is calculated by integrating over a simulated forecast distribution. Parameter values that generate explosive paths and positive outliers typical for sign reversal realizations would be punished automatically in the likelihood evaluation.

Table 7 Posterior distributions in SW07-Model 1966Q1–2014Q2

Parameter		No ZLB model		Endogenous ZLB duration		OIS-based ZLB duration	
		Posterior		Posterior		Posterior	
		Mode	Std.dev. Hess.	Mode	Std.dev. Hess.	Mode	Std.dev. Hess.
Calvo prob. wages	ξ_w	0.83	0.040	0.85	0.026	0.86	0.035
Calvo prob. prices	ξ_p	0.75	0.039	0.83	0.032	0.89	0.023
Indexation wages	ι_w	0.69	0.122	0.57	0.120	0.56	0.122
Indexation prices	ι_p	0.22	0.081	0.25	0.085	0.38	0.106
Gross price markup	ϕ_p	1.60	0.073	1.46	0.073	1.39	0.072
Capital production share	α	0.19	0.015	0.16	0.018	0.14	0.016
Capital utilization cost	ψ	0.80	0.075	0.73	0.094	0.60	0.120
Investment adj. cost	φ	4.58	0.941	4.61	0.61	5.84	1.095
Habit formation	$\varkappa$	0.62	0.054	0.62	0.031	0.68	0.041
Inv subs. elast. of cons.	σ_c	1.49	0.138	1.02	0.105	0.80	0.080
Labor supply elast.	σ_l	1.81	0.555	2.03	0.465	2.06	0.576
Hours worked in S.S.	$\bar{l}$	−0.40	1.178	−0.18	1.024	0.25	0.844
Discount factor	$100(\beta^{-1}-1)$	0.10	0.042	0.13	0.056	0.12	0.054
Quarterly growth in S.S.	$\bar{\gamma}$	0.41	0.014	0.42	0.026	0.43	0.016
Stationary tech. shock	ρ_a	0.96	0.008	0.97	0.014	0.97	0.018
Risk premium shock	ρ_b	0.40	0.104	0.85	0.055	0.97	0.008
Invest. spec. tech. shock	ρ_i	0.84	0.039	0.85	0.057	0.78	0.075
Gov't cons. shock	ρ_g	0.97	0.007	0.98	0.010	0.97	0.009
Price markup shock	ρ_p	0.92	0.030	0.88	0.046	0.86	0.048
Wage markup shock	ρ_w	0.97	0.010	0.98	0.024	0.99	0.005
Response of g_t to ε_t^a	ρ_{ga}	0.51	0.077	0.52	0.063	0.52	0.069
Stationary tech. shock	σ_a	0.46	0.025	0.48	0.032	0.50	0.029
Risk premium shock	σ_b	0.19	0.026	0.10	0.015	0.08	0.007
Invest. spec. tech. shock	σ_i	0.36	0.032	0.31	0.028	0.30	0.044
Gov't cons. shock	σ_g	0.49	0.025	0.48	0.026	0.48	0.025

Continued

Table 7 Posterior distributions in SW07-Model 1966Q1–2014Q2—cont'd

		No ZLB model		Endogenous ZLB duration		OIS-based ZLB duration	
		Posterior		Posterior		Posterior	
Parameter		**Mode**	**Std.dev. Hess.**	**Mode**	**Std.dev. Hess.**	**Mode**	**Std.dev. Hess.**
Price markup shock	σ_p	0.12	0.013	0.13	0.011	0.14	0.012
MA(1) price markup shock	ϑ_p	0.80	0.058	0.79	0.070	0.80	0.071
Wage markup shock	σ_w	0.37	0.022	0.36	0.020	0.36	0.021
MA(1) wage markup shock	ϑ_w	0.96	0.013	0.96	0.025	0.98	0.007
Quarterly infl. rate. in S.S..	$\overline{\pi}$	0.81	0.102	0.76	0.106	0.70	0.103
Inflation response	r_π	1.69	0.153	1.86	0.159	2.15	0.165
Output gap response	r_y	0.05	0.016	0.10	0.013	0.16	0.027
Diff. output gap response	$r_{\Delta y}$	0.24	0.027	0.24	0.020	0.23	0.024
Mon. pol. shock std	σ_r	0.23	0.013	0.22	0.012	0.21	0.011
Mon. pol. shock pers.	ρ_r	0.21	0.070	0.10	0.058	0.06	0.041
Interest rate smoothing	ρ_R	0.80	0.028	0.83	0.016	0.85	0.018
Log marginal likelihood		Laplace	−1146.69	Laplace	−1151.99	Laplace	−1175.24

Note: See notes to Table 2. The "No ZLB model" neglects the presence of the zero lower bound in the estimations, whereas the "Endogenous ZLB duration" allows the duration of the ZLB to be endogenous as described in the main text. Finally, the "OIS-based ZLB duration" imposes the duration of the ZLB in each point in time according to OIS rates for the federal funds rate between 2008Q4 and 2011Q2.

(wage markup and marginal costs, respectively) in the linearized price and wage equations. Evidently, the much higher degree of price and wage stickiness is only partly driven by the fact that prices and real wages fell modestly relative to output during the Great Recession (as can be seen in Fig. 1); even before the recession materialized there was already a strong trend in the data towards higher stickiness parameters, consistent with the findings by Del Negro et al. (2015b).[aa] Even so, we note that our estimated full sample model without the ZLB still features a much lower degree of price and wage stickiness than the policy model recently estimated by Brave et al. (2012).[ab]

In Fig. 7, we plot conditional forecast distributions for selected variables for the "No ZLB model" posterior in Table 7. In the left column, the forecast is conditional on the state in 2008Q3, whereas in the other two columns it is conditional on the filtered state in 2008Q4. Similarly to the results for the precrisis models in Fig. 2, the results in the left column shows that the severe drop in economic activity in 2008Q4 was outside the 95th percent uncertainty bands, even though the model is estimated on the full sample. This thus should be considered as an in-sample exercise. However, the median forecast conditional on the state in 2008Q4 is very accurate for yearly output growth and output (as deviation from trend) and the actual outcome is well within the uncertain bands for these variables, even disregarding the ZLB. For the federal funds rate, we see that the median forecast for the federal funds rate falls only slightly below nil for three quarters (2009Q1–2009Q3). This seemingly suggest that the ZLB was not much of a binding constraint during the Great Recession, consistent with the finding and interpretation in Del Negro et al. (2015b). This interpretation, however, ignores the fact that the forecast distribution for the federal funds rate has considerable mass below nil. Shifting this part of the distribution to 0 and above may therefore change the median outlook considerably.

To examine this possibility, the third column in Fig. 7 reports the forecast distribution when sampling parameters and shocks from the posterior distribution for the "No ZLB model" in Table 7, but with the unconstrained policy rule replaced by the policy rule in (17). This means that the actual and expected federal funds rate will respect the ZLB during the forecast horizon. Importantly, the 1000 different shock realizations used to construct the forecast distribution in the ZLB case are identical to those used to construct the unconstrained forecast distribution. Given the state in 2008Q4 the only difference between the results in the second and third column is that the federal funds rate is

[aa] This finding implies that the lower slope does not seem to be related to aggregate volatility, consistent with the findings by Vavra (2013).

[ab] As different models make alternative assumptions about strategic complements in price and wage setting, we have the reduced form coefficient for the wage and price markups in mind when comparing the degree of price and wage stickiness. In our benchmark model this coefficient equals 0.012 at the posterior mode for the New Keynesian Phillips curve which is similar to the estimate of Del Negro et al. (2013) (0.016). The estimate of Brave et al. (2012); however, the mode is as low as 0.002.

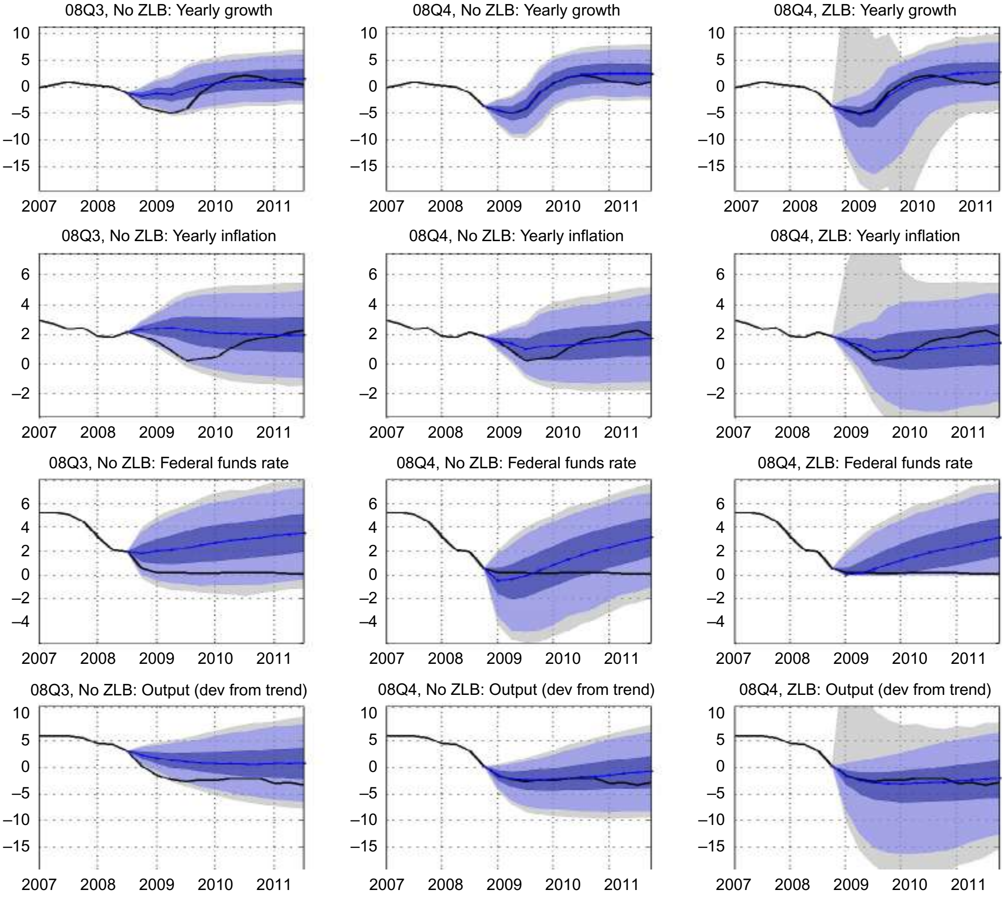


Fig. 7 Forecast 2009Q1–2011Q4 conditional on state in 2008Q4 for model estimated through 2014Q2 without imposing the ZLB.

constrained from falling below zero. As can be seen from the panels for output growth and output as deviation from trend, imposing the ZLB on the federal funds rate widens their uncertainty bands downwards quite notably. For output as deviation from trend, the lower 95th percentile shifts down from roughly −10% to nearly −20% in 2010. Hence, in the absence of unconventional monetary policies and coordination between monetary and fiscal policy (ie, fiscal stimulus when the economy enters a long-lived liquidity trap), the baseline model suggests that the ZLB may be associated with large economic costs.

On the other hand, the upper-95th percent bands for these variables are also much higher when the federal funds rate is constrained to fall below zero conditional on the given state in 2008Q4. For detrended output, the upper 95th percentile is above 10% in 2009. For yearly inflation, the upper 95th percentile is above 6%. Despite these elevated upper uncertainty bands for output growth, detrended output and inflation, the upper 95th percentile for the federal funds rate is lower than the corresponding percentile in the unconstrained policy rate distribution. This seemingly goes against the specification of the policy rule in (17) as the systematic part of the policy rule governing $\widehat{R}_t^*$ calls for a high policy rate whenever inflation, output growth and the output gap is high. The reason why this does not happen in the conditional ZLB distribution in Fig. 7 is that the model estimated without the imposed ZLB constraint needs large *negative* current and anticipated policy shocks $\widehat{\varepsilon}^r_{t+h|t}$ to satisfy $\mathrm{E}_t\widehat{R}_{t+h} \geq 0$. In essence, when the economy is hit by some really adverse shocks in these simulations and the policy rate is constrained to respond to these shocks for a sufficiently long period, inflation expectations and economic activity fall to such a large extent that a sequence of *negative* instead of *positive* policy shocks $\widehat{\varepsilon}^r_{t+h|t}$ for $h = 0, 1, \ldots, \tau_2$ are needed to prevent the federal funds rate to fall below nil. As discussed in Hebden et al. (2010) and Carlstrom et al. (2012), the switch in signs of the policy shocks only happens in the relatively few draws for which the policy rate is expected to be constrained by the lower bound for a very prolonged period of time (ie, τ_2 is large). This also explains why the upper 95th percentiles for inflation and output shifts up so much while the 90th percentile is roughly unchanged relative to the unconstrained distribution. The 90th percentile is associated with simulations of favorable fundamental shocks and parameter draws for which no large negative policy shocks are needed to prevent the policy rate to fall below nil.

We believe this result—that the ZLB can trigger adverse shocks to have sharply expansionary effects on the economy—is an unpalatable feature of the model. Therefore when we reestimate the model subject to the ZLB constraint on the federal funds rate, we believe it is crucial to impose the additional constraint—discussed in the beginning of this section—that the parameters of the economy have to be such that all current and expected policy shocks used to impose the policy rule in (17) are positive whenever the ZLB binds. By imposing this constraint, we ensure that the reestimated model does not feature any sign reversals of the policy shocks even for the most long-lived liquidity traps in our forecast distributions.

The estimation results for this variant of the model are reported in Table 7 and labeled "Endogenous ZLB duration". We use this label because both the incidence and duration of the ZLB spells are endogenous estimation outcomes in the model, and do not necessarily conform with other commonly used measures of the expected future path of the federal funds rate such as overnight index swap (OIS, henceforth) rates. By comparing the results with the "No ZLB model", we see that imposing the ZLB in the estimations have quite important implications for the posterior distribution. First of all the degree of price and wage stickiness is elevated even further, and the estimated parameters imply a slope of the New Keynesian Phillips curve of 0.006. This is somewhat lower than the median estimates of literature which cluster in the range of about 0.009–0.014, but well within standard confidence intervals provided by empirical studies (see, eg, Adolfson et al., 2005;

Altig et al., 2011; Galí and Gertler, 1999; Galí et al., 2001; and Lindé, 2005). In addition, the higher degree of nominal wage stickiness makes marginal costs even more sticky in the ZLB model. Together these features makes inflation and inflation expectations more slow to react to various shocks and therefore allow the model to cope with long spells at the ZLB without triggering indeterminacy problems (ie, switches in signs for the policy shocks). This finding is consistent with Erceg and Lindé (2010), who argue that a low slope of the Phillips curve is consistent with the development during the recent crisis where inflation and inflation expectations have fallen very moderately despite large contractions in output. It is also consistent with many recent papers which have estimated similar DSGE models, see, eg, Brave et al. (2012) and Del Negro et al. (2015b).

In addition to the higher degree of wage and price stickiness, there are two other important differences. Firstly, the coefficient on the output gap in the policy rule (Eq. (17)), r_y, is about twice as high as in the "No ZLB model". To the extent the output gap becomes significantly negative during the Great Recession, this will tend to push down the path of the federal funds rate and extend the duration of the ZLB. Secondly, the persistence coefficient in the risk premium shock process, ρ_b, increases sharply from 0.40 to 0.85. However, since the posterior mode for σ_b is reduced from 0.19 to 0.10, the unconditional variance for the risk-premium shock nevertheless falls slightly (from 0.044 to 0.039) in the ZLB model. Therefore the higher persistence does not imply a significantly larger role for the risk-premium shocks (apart from expectational effects). Even so, the likelihood prefers naturally more persistence in the shock process of the risk premium above a repeated set of positive innovations to explain the duration of the crisis and the slow pace of the recovery, but this shift in the posterior distribution of the parameters goes with a cost during the tranquil periods. This time variation in the role of the financial wedge over periods with more or less financial stress will be further discussed in Section 5.3.

Fig. 8 shows the forecast distribution (given the state in 2008Q4) in the "Endogenous ZLB duration" variant of the model. The left column gives the results when the ZLB is counterfactually neglected, whereas the right column shows the results when the ZLB is imposed. As expected, we see that the forecast distribution in the variant of the model which counterfactually neglects the ZLB features symmetric uncertainty bands around the modal outlook, and is a little bit too optimistic about the outlook for output relative to the model which imposes the ZLB (right column). More surprisingly is that the modal outlook for 2008Q4 in the model estimated and imposing ZLB constraint (right column in Fig. 8) differs very little to the modal outlook in the "No ZLB model" which completely neglects the ZLB (the middle column in Fig. 7). Obviously, a key difference is that the median path of the federal funds rate is constrained by the lower bound in 2009, but below nil in the unconstrained version of the model. Still, the quantitative difference for the median projection is small. The most noticeable difference between the No ZLB model and the model estimated under

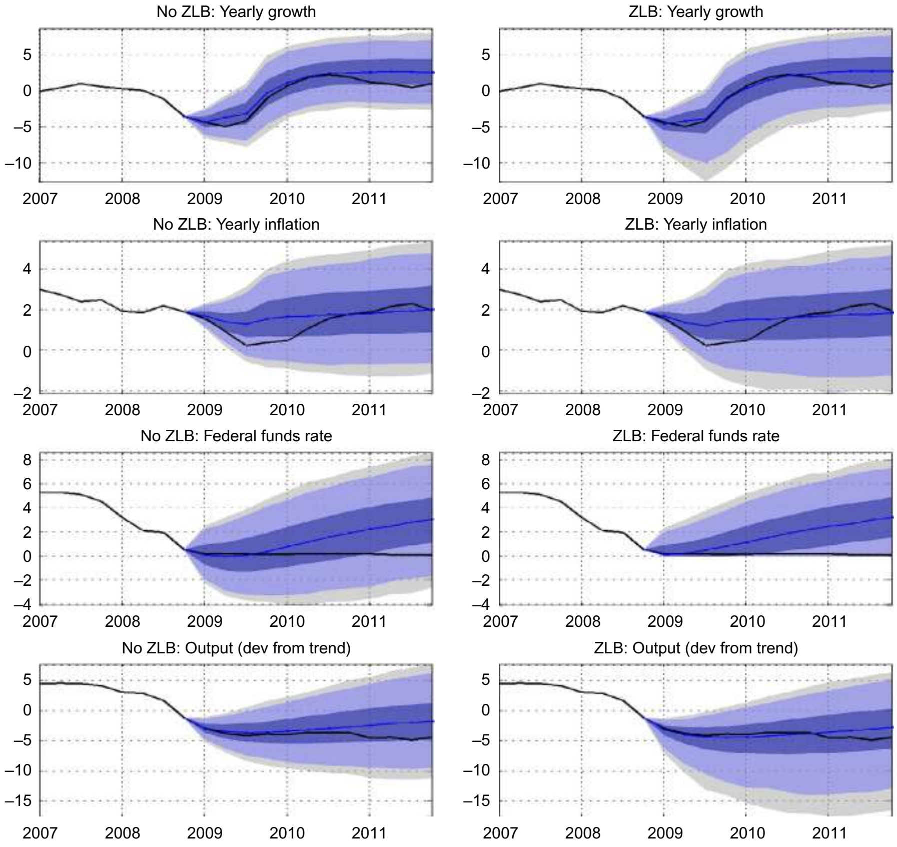


Fig. 8 Forecast 2009Q1–2011Q4 conditional on state in 2008Q4 for model estimated through 2014Q2 when imposing the ZLB.

the ZLB is the uncertainty bands: they are wider and downward skewed in the model that imposes the ZLB constraint (the right column of Fig. 8) compared to the No ZLB model that neglects the presence of the ZLB constraint.

However, the forecast distributions in the "No ZLB model"(the right column in Fig. 7)—which enforces the ZLB *ex post*—differs dramatically to the forecast distributions in the model estimated under the ZLB constraint (the right column in Fig. 8). The higher degree of wage and price stickiness in the model estimated under the ZLB constraint insulate the economy from the disaster scenarios and the indeterminate equilibria, and therefore shrink the uncertainty bands considerably. Overall this suggests that taking the ZLB into account in the estimation stage may be of key importance in assessing its economic consequences, and that it is not evident that models estimated on precrisis data can be useful for policy analysis when the economy enters into a long-lived liquidity trap. In such situations, the precrisis policy models may feature too much flexibility in price and wage setting, and, eg, yield implausibly large fiscal multipliers as noted by, eg, Erceg and Lindé (2010).

Another interesting feature of the model which neglects the ZLB and the variant of the model which is constrained to imposing Equation (17) through positive current and anticipated policy shocks is that the former has a higher log-marginal likelihood (-1146.7 vs -1152). This implies that imposing the ZLB on the model is somewhat costly in terms of data coherence. However, as suggested by the small differences in the conditional forecast distributions in Figs. 7 (middle column) and 8 (right column), it is not evident if this difference in log marginal likelihood is important from an economic viewpoint, although it is large enough to be sizable in terms of a Bayesian posterior odds ratio.

As the model is endogenously determining the incidence and duration of the ZLB spell, it is interesting to note that according to the model, the ZLB is expected in 2008Q4 to be a binding constraint from 2009Q1 to 2009Q3 in the modal outlook. The expected positive policy shocks we use to impose the ZLB partially substitute for the exceptionally huge risk premium shocks that drive the economy to the ZLB in the first place.[ac] The constraint is then expected to be binding during 2009 with a maximum duration of five quarters given the state in 2009Q1, and from 2010Q2 and onward, the model expects the interest rate to lift off already in the next quarter. The short duration of the ZLB spells is consistent with the findings of Chung et al. (2012). The fact that the federal funds rate has remained at the ZLB since then is by the model explained either as a result of expansionary monetary policy actions—forward guidance—or as standard policy reactions to unexpected headwinds. The filtered shocks suggest a dominant role for the second interpretation.

As noted previously, an alternative to letting the DSGE model determine the expected duration of the ZLB in each time period is to use OIS data for the federal funds rate as observables when estimating the model. By doing so, we follow Del Negro et al. (2015b) and require that the expected federal funds rate in the model matches the OIS data in each point in time when the ZLB is binding, ie, from 2008Q4 and onward. We use OIS data (acquired from the Federal Reserve Board) for 1, 2, …, 12 quarters' expected federal funds rates, and require the model to match those rates exactly through anticipated policy shocks following the general idea outlined by Maih (2010). The appealing feature of Maih's algorithm is that it does not require us to include standard

[ac] This high substitutability between anticipated monetary shocks that capture the effect of the ZLB on the one hand and the risk premium shock on the other hand, implies that it is very difficult to quantify accurately the precise impact of the ZLB on growth during the crisis. For instance, when a lagged shadow rate is used in the monetary policy rule instead of the lagged actual rate, the anticipated monetary policy shocks needed to impose the ZLB becomes much larger and more of the recession would then be attributed to the ZLB constraint while the contribution of the exogenous risk premium shock would decline significantly in the decomposition.

deviations for each of the anticipated policy shocks we use to fit the OIS data, and that the log-marginal likelihood can be compared to the models which does not condition on OIS data.

Before we turn to the results in Table 7, there are two additional important pieces of information. First, as we interpret the OIS data as expected means of future federal funds rates, we set them equal to nil in each point in time whenever they are lower than 50 basis points. We do this as our OIS estimation procedure does not explicitly account for future shock uncertainty, and the projected path of the interest rate from the model should therefore be viewed as a modal outlook (which will be lower than the mean of the forecast distribution when the ZLB binds). Second, because the Federal Reserve did not use explicit time-dependent forward guidance until August 2011, we restrict all anticipated policy shocks to be positive prior to this date. After this date we do not impose any signs on the anticipated policy shocks, because credible forward guidance—or a "lower for longer policy"—in the spirit of Reifschneider and Williams (2000) and Eggertsson and Woodford (2003), which extends the duration of the ZLB, is better viewed as expansionary than contractionary policy. Specifically, we allow the model to explain the sharp flattening of the OIS curve between the second and third quarter in 2011 with negative policy shocks, and do not impose this flattening to be associated with a noticeable deterioration in the economic outlook. According to the data, however, the magnitude of these expansionary "forward guidance" shocks are modest: interpreting the long ZLB spells as a deliberate "lower for longer" decision by the policy makers would further boost the predicted recovery by the model which goes against the observed slow and disappointing recovery in growth following the crisis.[ad]

The results when imposing the incidence and duration of the ZLB to adhere with OIS rates are shown in the left panel in Table 7, labeled "OIS-based ZLB duration." Relative to the posterior "Endogenous ZLB duration," for which the incidence and duration of the ZLB is determined endogenously in the model, we see that the degree of price stickiness is elevated further (from 0.83 to 0.89), and now implies a slope of the Phillips curve (ie, direct sensitivity of current inflation to marginal cost) of 0.003. This is substantially lower than, eg, the estimate in Altig et al. (2011), but still higher than Brave et al. (2012). To square this estimate with the microliterature is a challenge, and probably requires a combination of firm-specific capital (as in Altig et al., 2011),

[ad] There is a growing literature on the effectiveness of forward guidance. While Andrade et al. (2015) argue mainly on theoretical grounds that forward guidance may not be effective when agents have heterogeneous beliefs, Campbell et al. (2012), Williams (2014), and Del Negro et al. (2015a) argue on empirical grounds that forward guidance have had some positive impact. Even so, Del Negro, Giannoni and Patterson recognize that forward guidance may be too potent in a standard New Keynesian model relative to what the empirical evidence supports, and therefore integrate perpetual youth structure into the model to reduce its effectiveness. By and large, our estimated model produces results that are in line with their findings and suggests that forward guidance have had some, but limited, impact on the economy.

firm-specific labor (as in Woodford, 2003), and a higher sensitivity of demand to relative prices (ie, higher Kimball parameter ε_p). Apart from the higher stickiness we also see an elevated role for the risk-premium shock in this model (ρ_b rises sharply from 0.85 to 0.97, whereas the std of the innovations only falls moderately from 0.10 to 0.08), and that the degree of habit formation consumption ($\varkappa$) and investment adjustment costs (φ) rises somewhat. Finally, the response coefficient for the output gap in the policy rule is increased further, and is now 3× higher than in the model which neglects the presence of the ZLB.

The reason why these parameters are further changed relative to the "No ZLB model" is that the OIS data generally imposes longer-lived ZLB episodes than the model endogenously produces. In order to be able to explain those episodes with *positive* anticipated policy shocks through 2011Q2 the model needs to make dynamics more sluggish and explain the rebound in inflation during 2010 with temporary shocks. However, enforcing this sluggish dynamics on the model is rather costly in terms of log-marginal likelihood, which falls from -1152 in the model with endogenous ZLB duration to -1175.2 for the OIS-based ZLB duration. This is a sizable drop and a possible interpretation is that the SW07-model despite imposing the ZLB constraint, was more optimistic about the recovery than market participants during this episode.

There are of course other possibilities as to why the ZLB episodes in the model are short-lived relative to what OIS data suggest. They include that; (i) the model miss-measures the size and persistence of the relevant output gap, (ii) the model-consistent or rational expectation hypothesis fails to capture the stickiness and persistence in expectations that might be caused by learning dynamics or information filtering issues, (iii) the steady state natural real rate has fallen (eg, due to lower trend growth) and this has caused the (gross) steady state nominal interest rate R in Eq. (14) to fall; *ceteris paribus* this calls for an extended ZLB duration, and (iv), the Federal Reserve decided to respond more vigorously to the negative output gap (ie, r_y in Eq. (14) increased) from the outset of the Great Recession and thereafter.[ae] Yet other possibilities is that our model above misses out on time-varying volatility of the shocks and omits financial frictions and the cost channel of monetary policy. We explore these latter possibilities below.

5.2 Allowing for Time-Varying Volatility

As documented earlier, the prototype linear Gaussian model with constant volatility does not provide a realistic predictive density for the forecast, in particular around severe recession periods or periods of high financial and monetary stress. A large share

[ae] To the extent that these mechanisms are at work, they should be picked up in our estimated model as expansionary monetary policy shocks due to the presumption in our analysis that the Fed before and after the crisis (ie, upon exit from the ZLB) adheres to the same Taylor-type policy rule (Eq. (17)), and that agents form their expectations accordingly.

of the research effort on DSGE models since the financial crisis and the Great recession has tried to overcome these weaknesses of the basic DSGE setup. By now, most models used in academia and in policy institutions contain financial frictions and financial shocks in an effort to introduce stronger amplification mechanisms in the model. As we will discuss in the next section, however, to the extent that even the modified models adopt a Gaussian linear framework, they still depend on extremely large shocks to predict important recessions. The explicit modeling of the nonlinear macrofinance interactions is complex and ambitious and the research in that direction has not yet been integrated in empirical macro models. A technically feasible avenue to improve the predictive densities of the linear DSGE model is to allow for a more complicated stochastic structure. Here we illustrate this approach by considering a Markov Switching (MS) stochastic structure following Liu et al. (2013).[af] By allowing for such a shock structure, the hope is that the estimated model can capture the phenomena that the economic outlook is sometimes very uncertain (ie, the economy is filtered to be in the high volatility regime), without necessarily destroying its ability to provide reasonably narrow forecast uncertainty bands in normal times (ie, in the low volatility regime).

Low frequency changes in the shock variances have been analyzed by Fernández-Villaverde and Rubio-Ramírez (2007) and Justiniano and Primiceri (2008) via stochastic volatility processes. Chib and Ramamurthy (2014) and Curdia et al. (2014) show that a Student's *t*-distribution for the innovations is also strongly favored by the data as it allows for rare large shocks. The latter authors makes the point that the time variation in shock variances should contain both a low and a high frequency component.

To capture these insights, we consider a version of the benchmark model in which we allow for two independent Markov Switching processes in the shock variances. Each Markov process can switch between a low and a high volatility regime. One process affects the volatility of all the structural innovations with exception of the wage markup shock, based on the observation that the wage markup and the observed real wage variable has a completely different volatility profile compared to the other shocks and variables as shown in Figs. 1 and 5. The second Markov process is restricted to the non-Gaussian structural shocks as identified in Table 6 in Section 4.3: this process affects the volatility in the monetary policy, the risk premium and the investment specific innovations. The volatility in these three shocks is scaled by both the common (σ_c) and the monetary/financial volatility factor (σ_{mf}). The typical process for these three shocks is now written as follows:

$$\widehat{\varepsilon}_t = \rho\widehat{\varepsilon}_{t-1} + \sigma_{mf}\left(s_{mf}\right)\cdot\sigma_c(s_c)\cdot\sigma\cdot\eta_t,\ \eta_t \sim N(0,1).$$

[af] We use the RISE toolbox to implement this exercise, see Maih (2015).

The estimated transition probabilities are summarized by the following matrices:

$$Q_c\begin{pmatrix} low \\ high \end{pmatrix} = \begin{bmatrix} 0.95 & 0.07 \\ 0.05 & 0.93 \end{bmatrix} \quad Q_{mf}\begin{pmatrix} low \\ high \end{pmatrix} = \begin{bmatrix} 0.92 & 0.46 \\ 0.08 & 0.54 \end{bmatrix}.$$

The relative volatility of the two regimes are estimated as:

$$\sigma_c\begin{pmatrix} low \\ high \end{pmatrix} = \begin{bmatrix} 1 \\ 1.74 \end{bmatrix} \quad \text{and} \quad \sigma_{mf}\begin{pmatrix} low \\ high \end{pmatrix} = \begin{bmatrix} 1 \\ 2.33 \end{bmatrix}.$$

In Fig. 9, we plot the smoothed regime probabilities for the model estimated over the complete sample. A filtered probability near unity (zero) implies that the economy is filtered to be in the high (low) volatility regime.

The common volatility process captures the great moderation phenomena. The high volatility regime is typically preferred during most of the 1970s and the first half of the 1980s, while the low volatility regime is active during the great moderation and is interrupted by the financial crisis and the resulting Great Recession. Both regimes are estimated to be persistent and the relative volatility during the high volatility regime is almost twice as high as in the low volatility regime. The monetary/financial volatility process captures the increase in the volatility during most of the recession periods and in the late 1970s- and early 1980s-episode of increased monetary policy uncertainty. The expected duration of this high volatility/financial stress regime is relatively short lived with a quarterly transition probability of 0.46%. The estimated parameters that describe the regimes and the regime probabilities are very stable when estimating the model for the precrisis period or for the complete sample (not shown).

Table 8 shows that the estimated log marginal likelihood of our model with switching volatility outperforms the log marginal likelihood of the homoscedastic Gaussian models by far. In this sense, our results confirm the results in the literature based on stochastic

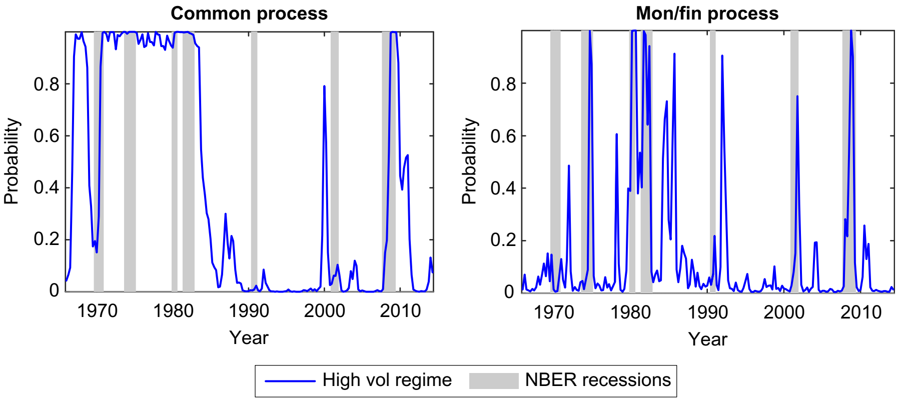


Fig. 9 Smoothed probabilities of the two volatility Markov processes.

Table 8 Log marginal likelihood of alternative regime switching specifications

	Sample period	
	Precrisis: 66Q1–07Q4	**Full sample: 66Q1–14Q2**
No regime switching (RS)	−961.8	−1146.7
RS in common process	−894.6	−1060.9
RS in mon/fin process	−911.8	−1082.1
RS in common and mon/fin process	−881.7	−1046.0

Note: None of the models in this table are estimated subject to the ZLB on policy rates.

volatility or *t*-distributed shocks. In contrast with Liu et al. (2013) and in support of the results of Curdia et al. (2014), we find strong evidence in favor of a setup that allows for multiple sources of volatility changes. The time-varying volatility structure requires sufficient flexibility to account for a common low frequency trend on the one hand, and a more cyclical high frequency process that controls mainly the monetary and financial shocks on the other hand.[ag]

Accounting for the non-Gaussian stochastic structure drastically improves the log marginal likelihood of our models, but leaves the estimated parameters, ie, the central forecasts and identified innovations, relatively unaffected. Most of the gains are realized because the predictive densities attribute appropriate probabilities to the extreme tail events: the large downturns in recessions and the corresponding sharp responses in policy rates. To illustrate this property, we consider the predictive forecast distribution with the precrisis model conditional on data up to 2008Q3, and we calculated the percentile interval that contains the 2008Q4 realized output growth observation (see Fig. 10). For our baseline precrisis model, the realized 2008Q4 growth rate falls completely outside of the simulated predictive densities based on 10,000 draws with parameter and shock volatility, as is clear from the left panel in the figure (see also Fig. 2). In contrast, in the model with Markov Switching volatility, almost 1% of the simulated forecasts fall below the 2008Q4 realization, as shown in the figure's right panel.[ah] The Markov

[ag] Our restrictive setup of two processes improve the log marginal likelihood by 115 for the complete sample. More flexible structures could easily improve this result but this goes with a cost because these setups are less robust, are computational much more intensive and lack an intuitive interpretation of the regimes. Curdia et al. report a gain of 154 in the log marginal likelihood for a setup that contains a combination of shock specific stochastic volatility and *t*-distributed innovations.

[ah] We want to emphasize that it is not the case that we are more content with this model just because it gives a positive probability that the great recession could indeed happen. As pointed out earlier, our rationale for going in this direction is that the models with regime switching in shock variances and the propagation of financial frictions (see analysis in the next section) improves the statistical properties of the model (as suggested by the strong improvement in log marginal likelihood) and makes sense from an economic viewpoint (supporting the widely held belief that financial frictions are key to understand the crisis).

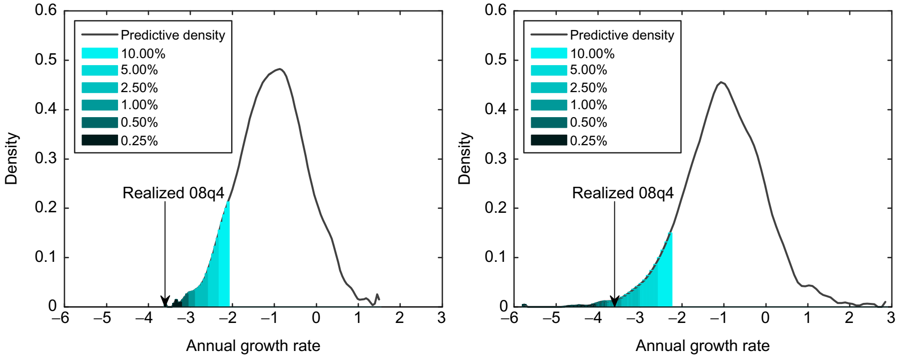


Fig. 10 Distributions for output growth (four-quarter change) in 2008Q4 given state in 2008Q3 in models with constant volatility (left panel) and time-varying volatility (right panel).

Switching volatility structure, by allowing for a mixture of normal distributions, gives more probability to the tails in general. In addition, the probability of the high volatility regimes in both the high and the low frequency Markov processes increased already by 2008Q3 because the magnitude of the realized shocks preceding the fourth quarter observation were relatively large.

5.3 Augmenting the Model with Financial Frictions and a Cost Channel

We incorporate a financial accelerator mechanism into the benchmark model in Section 3 following the basic approach of Bernanke et al. (1999). Thus, the intermediate goods producers rent capital services from entrepreneurs rather than directly from households. Entrepreneurs purchase physical capital from competitive capital goods producers (at price $\widehat{Q}_t^k$, and resell it back at the end of each period), with the latter employing the same technology to transform investment goods into finished capital goods as described by Eq. (11). To finance the acquisition of physical capital ($\widehat{\bar{k}}_t$), each entrepreneur combines his net worth ($\widehat{NW}_t^e$) with a loan from a bank, for which the entrepreneur must pay an external finance premium due to an agency problem. We follow Christiano et al. (2008) by assuming that the debt contract between entrepreneurs and banks is written in nominal terms (rather than real terms as in Bernanke et al., 1999). Banks, in turn, obtain funds to lend to the entrepreneurs by receiving deposits from households, with households bearing no credit risk (reflecting assumptions about free competition in banking and the ability of banks to diversify their portfolios). In equilibrium, shocks that affect entrepreneurial net worth—ie, the leverage of the corporate sector—induce fluctuations in the corporate finance premium.[ai]

[ai] For further details about the setup, see Bernanke, Gertler and Gilchrist, and Christiano, Motto and Rostagno. Excellent expositions are also provided by Christiano et al. (2007) and Gilchrist et al. (2009).

When estimating the model with the financial friction mechanism embedded, we add one more observable variable, the widely-used Baa-Aaa corporate credit spread (see Appendix C for exact definition and data sources). This spread plays a key role in the Bernanke–Gertler–Gilchrist framework. Since we also want to learn about the importance of shocks originating in the financial sector, and because we need as many shocks as observables to avoid stochastic singularity, we also add a "net worth" shock to the set of estimated shocks. We derive this shock by allowing the survival probability of the entrepreneurs to vary over time. Hence, this shock will enter in the accumulation equation for the entrepreneurs net worth. An alternative would have been to allow for a shock directly in the equation which relates the spread (or equivalently, the external finance premium) to the entrepreneurs leverage ratio following, eg, Del Negro and Schorfheide (2013) or Christiano et al. (2008). We preferred, however, not to add a shock directly in the spread equation in an attempt to elevate the endogenous propagation of the financial accelerator mechanism.[aj] Even so, the equation for the external finance premium,

$$\mathrm{E}_t\widehat{R}^e_{t+1} - \widehat{R}^b_t = \chi\left(\widehat{Q}^k_t + \widehat{\bar{k}}_t - \widehat{NW}^e_t\right), \tag{18}$$

still contains a shock because we assume that the financing rate of the banks, $\widehat{R}^b_t$, is not the risk-free rate set by the central bank, but rather the sum of the policy rate $\widehat{R}_t$ and the risk-premium shock $\widehat{\varepsilon}^b_t$.

As recent research by Christiano et al. (2015) and Gilchrist et al. (2015) emphasize the importance of firms financing conditions for their price setting behavior, we also embed a cost channel into the model. Specifically, we assume that firms have to borrow short to finance their wage bill following Christiano et al. (2005). As shown in the CEE paper, the working capital channel can cause inflation to rise following a tightening of monetary policy if firms financing costs rise sufficiently. To allow for sharp increases in firms financing costs, we assume that the relevant financing rate is the expected nominal return on capital for the entrepreneurs as opposed to the risk-free policy rate. However, instead of imposing that all firms borrow to finance their entire wage bills as in CEE, we estimate a parameter, ν, which determines the share of firms that are subject to working capital, so that the expression for log-linearized marginal costs becomes

$$\widehat{mc}_t = (1-\alpha)\ \left(\widehat{w}_t + \widehat{R}^f_t\right) + \alpha\ \ \widehat{r}^k_t - \widehat{\varepsilon}^a_t,$$

[aj] Christiano et al. (2008) embed a complete banking sector into their model and estimate it using 17× series and an equal number of shocks. A benefit, however, of our more modest perturbation of the model size and number of observables matched is that it allows for a straightforward comparison with the findings in the benchmark SW07-model.

where $\widehat{R}_t^f$ is the effective working capital interest rate given by

$$\widehat{R}_t^f = \frac{\nu R}{\nu R + 1 - \nu} \mathrm{E}_t \widehat{R}_{t+1}^e, \tag{19}$$

in which $\mathrm{E}_t \widehat{R}_{t+1}^e$ is the nominal expected return on capital for the entrepreneurs. From Eq. (19), we notice that $\widehat{R}_t^f = \mathrm{E}_t \widehat{R}_{t+1}^e$ when $\nu = 1$.

The SW07-model embedded with the financial friction mechanism and the cost-channel thus include five additional estimated parameters; ν, the two parameters for the AR(1) process for net worth (ρ_{nw} and σ_{nw}), the monitoring cost parameter μ which indirectly determines the sensitivity of the external finance premium to the entrepreneurs leverage ratio (χ in Eq. (18), and a constant ($\bar{c}_{sp}$) which captures the mean of the credit spread. Estimation results for three specifications of the model are provided in Table 9; first we have the "Precrisis sample" (sample 1966Q1–2007Q4 without the ZLB), second, the full sample (66Q1–14Q2) when imposing the ZLB constraint with endogenous duration, and third we study a variant of the model with the ZLB which allows the key parameter μ to switch stochastically between a high and low value. The adopted priors for the five new parameters are provided in the notes to the table. The priors for the other parameters are the same as before (and already stated in Table 2).

In the precrisis model, the external finance premium delivers only a very modest amplification of the standard shocks. The estimated elasticity of the spread to the net worth ratio is small (with $\mu = 0.033$ and χ in Eq. (18) equals 0.012, implying an annualized spread sensitivity of 0.048), a result that is in line with the estimates reported in Gilchrist et al. (2009). The exogenous risk-premium shock and—to a lower degree—the monetary policy shock are most impacted by the introduction of the FA mechanism because they have the biggest impact on the price of capital and net worth. The net worth channel tends to support the persistence in the response of investment to these shocks. The low sensitivity of the spread to the traditional shocks also implies that most of the fluctuations in the external finance premium are generated by the new exogenous shock that is assumed to hit directly the net worth of the entrepreneurs. This highly volatile shock explains up to 70% of the variance in the spread and one-third of the variance in investment. As such, the net worth shock substitutes for the exogenous risk premium and for the investment-specific technology shock. The latter also captures financial frictions as suggested by Justiniano et al. (2013a). Overall, the impact of the net worth shock on the macrodynamics remains modest and one important reason for this is that the net worth shock typically crowds out private consumption and this clashes with the observed strong comovement between consumption, and investment over the business cycle.[ak]

[ak] This crowding out problem is not present for our reduced form risk-premium shock ε_t^b in Eq. (12), see Fisher (2015) for a structural interpretation of this risk-premium shock.

Table 9 Posterior distributions in SW model with financial frictions

Parameter		Precrisis sample		Endogenous ZLB duration		Endog. ZLB dur. with regime switch	
		Mode	Std.dev.Hess.	Mode	Std.dev.Hess.	Mode	Std.dev.Hess.
Calvo prob. wages	ξ_w	0.72	0.082	0.83	0.009	0.86	0.017
Calvo prob. prices	ξ_p	0.68	0.045	0.84	0.024	0.83	0.029
Indexation wages	ι_w	0.67	0.129	0.63	0.125	0.60	0.130
Indexation prices	ι_p	0.21	0.084	0.23	0.081	0.23	0.085
Gross price markup	ϕ_p	1.61	0.077	1.45	0.062	1.43	0.063
Capital production share	α	0.21	0.018	0.17	0.016	0.17	0.016
Capital utilization cost	ψ	0.44	0.114	0.50	0.100	0.64	0.096
Investment adj. cost	φ	4.71	0.845	4.61	0.564	4.00	0.560
Habit formation	$\varkappa$	0.77	0.037	0.67	0.018	0.63	0.025
Inv subs. elast. of cons.	σ_c	1.27	0.110	0.97	0.100	1.04	0.084
Labor supply elast.	σ_l	1.50	0.565	1.58	0.437	1.85	0.459
Hours worked in S.S.	$\bar{l}$	0.85	1.082	−0.48	0.804	−0.23	0.768
Discount factor	$100(\beta^{-1} - 1)$	0.13	0.051	0.12	0.049	0.12	0.049
Quarterly growth in S.S.	$\bar{\gamma}$	0.43	0.015	0.42	0.015	0.42	0.017
Stationary tech. shock	ρ_a	0.96	0.011	0.96	0.012	0.97	0.012
Risk premium shock	ρ_b	0.26	0.083	0.83	0.022	0.85	0.029
Invest. spec. tech. shock	ρ_i	0.80	0.055	0.84	0.040	0.88	0.035
Gov't cons. shock	ρ_g	0.96	0.010	0.97	0.009	0.97	0.009
Price markup shock	ρ_p	0.92	0.034	0.89	0.039	0.89	0.040
Wage markup shock	ρ_w	0.98	0.013	0.98	0.007	0.97	0.001
Response of g_t to ε_t^a	ρ_{ga}	0.49	0.076	0.53	0.068	0.53	0.068
Stationary tech. shock	σ_a	0.47	0.029	0.49	0.027	0.49	0.027
Risk premium shock	σ_b	0.21	0.021	0.11	0.010	0.10	0.010
Invest. spec. tech. shock	σ_i	0.35	0.036	0.31	0.020	0.32	0.013
Gov't cons. shock	σ_g	0.47	0.029	0.47	0.024	0.47	0.024
Price markup shock	σ_p	0.12	0.015	0.12	0.013	0.13	0.013
MA(1) price markup shock	ϑ_p	0.75	0.079	0.79	0.070	0.79	0.071
Wage markup shock	σ_w	0.31	0.025	0.37	0.020	0.37	0.021
MA(1) wage markup shock	ϑ_w	0.92	0.049	0.96	0.008	0.96	0.001

Continued

Table 9 Posterior distributions in SW model with financial frictions—cont'd

Parameter		Precrisis sample		Endogenous ZLB duration		Endog. ZLB dur. with regime switch	
		Mode	Std.dev.Hess.	Mode	Std.dev.Hess.	Mode	Std.dev.Hess.
Quarterly infl. rate. in S.S.	$\overline{\pi}$	0.78	0.105	0.73	0.097	0.76	0.093
Inflation response	r_π	1.91	0.170	1.78	0.119	1.83	0.133
Output gap response	r_y	0.07	0.022	0.10	0.008	0.11	0.012
Diff. output gap response	$r_{\Delta y}$	0.24	0.028	0.24	0.014	0.24	0.015
Mon. pol. shock std	σ_r	0.23	0.014	0.22	0.012	0.22	0.011
Mon. pol. shock pers.	ρ_r	0.14	0.068	0.10	0.047	0.09	0.047
Interest rate smoothing	ρ_R	0.81	0.026	0.84	0.006	0.84	0.009
Net worth shock pers.	ρ_{nw}	0.25	0.080	0.30	0.088	0.30	0.084
Net worth shock std	σ_{nw}	0.27	0.031	0.19	0.024	0.23	0.032
Working capital share	ν	0.34	0.120	0.64	0.228	0.60	0.251
Credit spread in S.S.	$\bar{c}_{sp}$	1.51	0.292	1.28	0.285	0.97	0.059
Monitoring cost	μ	0.03	0.004	0.06	0.007		
Monitoring cost—Regime 1	μ_1					0.03	0.004
Monitoring cost—Regime 2	μ_2					0.08	0.011
Trans. Prob.—R1 to R2	p_{12}					0.04	0.015
Trans. Prob.—R2 to R1	p_{21}					0.16	0.055
Log marginal likelihood		Laplace	−897.80	Laplace	−1112.00	Laplace	−1063.00

Note: For the financial friction parameters, we use the same prior as for the other exogenous shocks (stated in Table 2). For μ and $\bar{c}_{sp}$, we use a normal distribution with means 0.25 and 1.00 and standard deviations 0.10 and 0.50, respectively. Finally, for ν we use a beta distribution with mean 0.50 and standard deviation 0.20. The "Precrisis sample" neglects the presence of the ZLB and is estimated on data up to 2007Q4, whereas the "Endogenous ZLB duration" imposes the ZLB as described in Section 5.1, and is estimated up to 2014Q2. "Endog. ZLB dur. with regime switch" also imposes the ZLB, but allows μ to vary stochastically between a low (μ_1) and high (μ_2) value. For μ_1 and μ_2, we use a normal distribution with means 0.025 and 0.25, and standard deviations 0.01 and 0.10, respectively. For the transition probabilities p_{12} and p_{21}, we use a beta distribution with means 0.10 and 0.30 and standard deviations 0.05 and 0.10, respectively.

The direct comparison of the marginal likelihood with the baseline model is complicated because the financial frictions model (FF model henceforth) has an additional observable in the form of the Baa–Aaa spread. When we estimate the FF-model without this additional observable, the log marginal likelihood improves by a factor of 10 when no additional shock is considered and by a factor of 20 when the net worth shock is retained. With a posterior mode for $\mu = 0.2$ in this variant of the model, the estimated sensitivity of the spread to the net worth ratio in this model is much higher, ie, 0.08, or 0.32 in annualized terms. This result is more supportive for an important endogenous amplification effect of the standard shocks through the net worth channel (see also De Graeve (2008) for a similar result). This observation suggests that the use of the Baa–Aaa spread as an observable for the external finance premium in the model can be too restrictive. Baa-Aaa spread is only one specific measure for default risk, and the cost of credit for firms is determined by various risks and constraints in the financial sector.[al]

Not surprisingly, when we evaluate the performance of the FF-model for the complete sample including the 2008Q4–2009Q1 crisis period, the monitoring cost parameter μ and the implied elasticity of the spread to the net-worth ratio doubles. Perhaps surprisingly, the standard error of the exogenous net-worth shock is substantially lower, 0.19 vs 0.27 in the model estimated on precrisis data. We interpret this finding to imply that the endogenous amplification becomes more important when including the crisis period in the estimation sample. As we also impose the ZLB constraint in the estimation of this model, the estimated nominal wage and price stickiness is again very high (0.83 and 0.84, respectively) so that all the expected policy shocks that are required for the model to respect the ZLB constraint are positive. It is also striking that in this full-sample model, the estimated fraction of the wage bill that requires external financing is substantially higher than in the precrisis version, supporting the argument in Christiano et al. (2015) that this channel was important during crisis. The magnitude of this cost channel increases from 0.33 to 0.64, but in both models the uncertainty in the posterior distribution for this parameter is very high. These two observations, the time variation in the role of financial frictions and the potential role of the cost channel for the inflation dynamics, are discussed in more detail below.

5.3.1 A Regime Switching Model with Occasionally More Severe Financial Frictions

Precrisis DSGE models typically neglected the role of financial frictions. This additional transmission mechanism was considered nonvital for forecasting output and inflation during the great moderation period, and by Occam's razor arguments this mechanism was typically left out. However, as our discussion of the in-sample innovations illustrated, there was already strong evidence in our estimated precrisis model for occasionally big

[al] Gilchrist et al. (2009) and Gilchrist and Zakrajsek (2012) present alternative indicators of the default spread that have a stronger predictive power for economic activity than the Baa–Aaa spread.

disturbances that seemed to be highly correlated with financial spreads and return indicators. When looking at these results from a broader perspective that also gives appropriate attention to the potential risks around the central banks forecast, these outliers should not be disregarded. A linear Gaussian approach is not the most efficient framework for handling these issues. The instability in the estimated parameters of our FF-model depending on the estimation sample clearly illustrates these limitations. To more efficiently capture the time-varying relevance of the financial frictions in our model, we therefore consider here a Markov switching setup in which the constraints from the financial frictions can become much more binding occasionally.

In our Regime Switching Financial Friction model (RS-FF), we allow for two possible regimes: one regime (high-FF) with a high monitoring costs—implying a high sensitivity of the spread to the net worth position—and another regime (low-FF) with a low monitoring costs and low sensitivity of spread to leverage.[am] The estimation results for this model is reported last in Table 9, and the data prefer this RS-FF setting compared to the linear FF-model as shown by the gain in the log marginal likelihood of more than 30 in the precrisis context (not shown) and around 50 in the sample with the recent crisis.[an] The transition probabilities and the regime-specific μ parameter are given by:

$$Q_{FF}\begin{pmatrix}\text{low}\\ \text{high}\end{pmatrix}=\begin{bmatrix}0.96 & 0.16\\ 0.04 & 0.84\end{bmatrix}\ \mu_{FF}\begin{pmatrix}\text{low}\\ \text{high}\end{pmatrix}=\begin{bmatrix}0.029\\ 0.084\end{bmatrix}.$$

The estimation results indicate that the elasticity of the spread to the leverage ratio varies between the two regimes by a factor of 2.7. As shown in Fig. 11, the high-FF regime is active mainly around the two recession periods in the 1970s, and its probability increases slightly during all recessions. When evaluated over the more recent period the probability of the high-FF regime starts to rise early in 2008 and remains active during the financial crisis in 2009, but quickly returns to the low-FF regime after 2009. The higher marginal likelihood is due to the time-varying volatility in the spread: in the high-FF regime, the financial friction is strongly binding and the spread reacts more than twice as strong to the leverage ratio. The impact of shocks on investment is also higher but the magnitude of the amplification is moderate up to a factor of 1.5 maximum. The expected period-by-period persistence of the high-FF regime is limited (0.84) and this reduces the impact of spread increases on the discounted value of future expected returns on investment.

As evidenced in Fig. 12, the central forecast of the single-regime precrisis FF-model, conditional on data up to 2008Q3 is completely missing the magnitude of the 2008Q4

[am] Christiano et al. (2014) focus instead on the distribution of the idiosyncratic productivity risk as the source for time-varying financial frictions. Levin et al. (2004) identify the time variation in the bankruptcy cost parameter, the equivalent of our monitoring cost, as the source for the counter-cyclical external premium behavior.

[an] Suh and Walker (2016) also finds support for time-variation in parameters governing financial frictions.

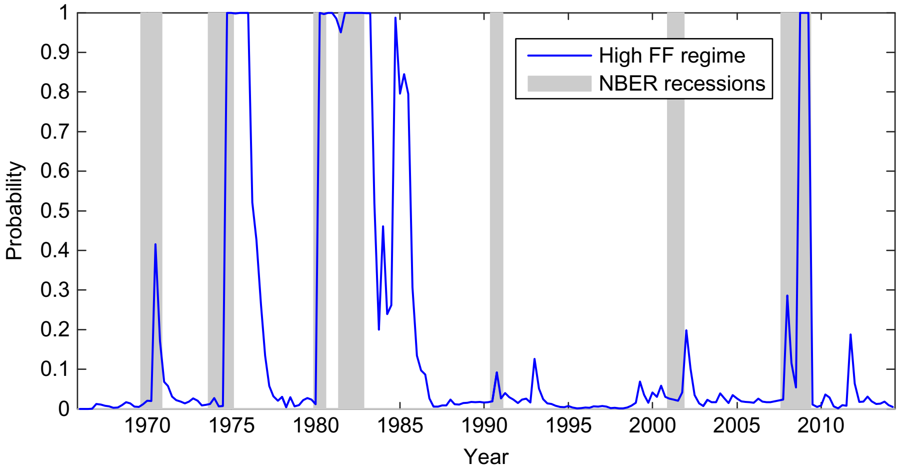


Fig. 11 Markov process in the financial frictions model.

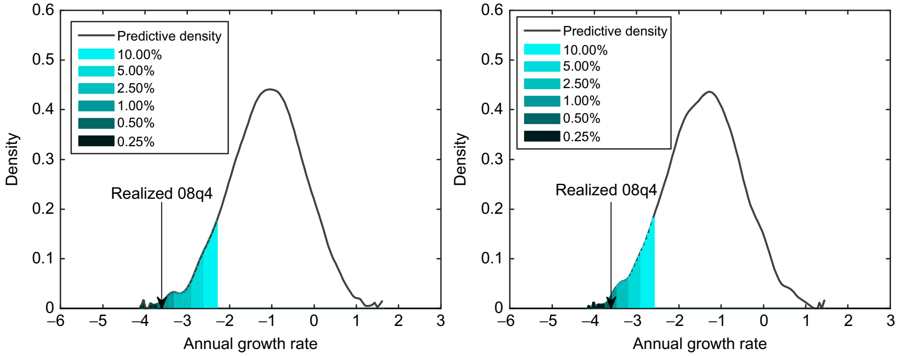


Fig. 12 Distributions for output growth (four-quarter change) in 2008Q4 given state in 2008Q3 in financial friction models with constant parameters (left panel) vs regime switching (right panel).

downturn just as the benchmark SW07-model without financial frictions. By comparing the no-financial friction model—left panel in Fig. 10—with the constant parameter FF-model—left panel in Fig. 12, we see that the distribution around the FF-forecast is more disperse due to the extra volatility that is generated by the spread and the additional net worth shock. As a result, the extreme negative output growth realization of 2008Q4 now falls within the 0.25% interval of the predictive density, which is some improvement relative to the baseline model. The precrisis RS-FF model, shown in the right panel in Fig. 12, further improves on this result because the probability of being in the high friction regime increased in 2008Q3 (56% against an unconditional probability of 20%) and this introduces a high degree of skewness in the predictive density of the spread. While the precrisis FF-model predicts a 1% upper tail for the expected spread above 2.3 percentage points in 2008Q4, this becomes as high as 3 percentage points in the RS-FF model. The probability of the observed 2008Q4 output growth outcome now lies around the 0.5% tail interval, which is still small but at least the *ex post* realized event obtains some nonzero probability in the predictive density. This result indicates that if we appropriately could integrate the nonlinear accelerator dynamics from financial frictions in our DSGE models we may obtain a more realistic predictive density that resembles these from the reduced form time-varying volatility models such as our RS-volatility example in Section 5.2.

Given the important role of the spread in the short run forecast, it is also informative to show how a conditional forecast, conditional on the timely observation of the spread, performs in the crisis period. Therefore, we make a forecast conditional on the 2008Q3 state of the economy as filtered by the precrisis FF-model but now we also provide the model with the information that the spread increased to the exceptionally high observed level of 3.02 percentage points in 2008Q4 (from 1.55 percentage points in 2008Q3). This conditioning is plausible in real time as the spread already in the beginning of the fourth quarter in 2008 (mid-October) had reached 3 percentage points. Fig. 13 shows the unconditional (left panel) and conditional (right panel) forecast distributions for GDP growth in 2008Q4. As seen from the figure, the forecast conditional on the timely information from the spread display a median prediction for annual GDP growth of −2.11% in 2008Q4 and −1.92% in 2009Q1 (not shown), which should to be compared to the observed −3.61% and −4.42% in the actual data and unconditional forecast of −1.05% (left panel in the figure) and 0.06% (not shown).

In the RS-FF model, the result depends very much on the regime in which the economy is finding itself in 2008Q3: the impact of conditioning on the spread is most disturbing when the economy is in the low friction regime. Extreme high spreads are very difficult to reconcile with the low friction regime, with its low elasticity of spread to leverage, and therefore the spreads are translated in huge negative shocks in net worth

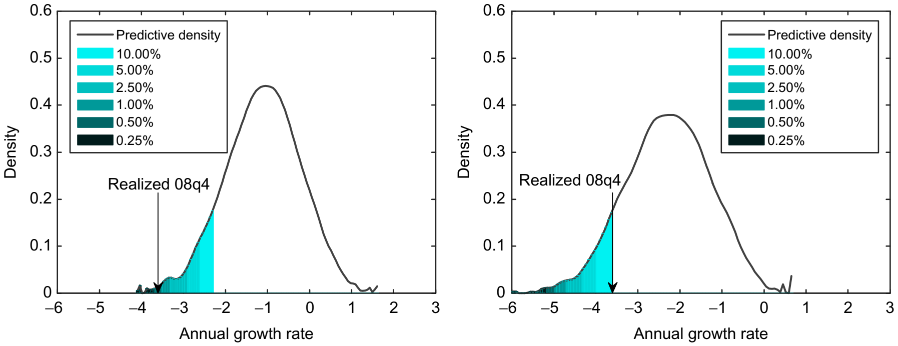


Fig. 13 Distributions for output growth (four-quarter change) in 2008Q4 in constant parameter financial friction model. Left panel is unconditional projection given state in 2008Q3, whereas the right panel is conditional on the spread in 2008Q4.

and/or risk premiums which then also result in worse output growth predictions of −2.53% and −3.01% in 2008Q4 and 2009Q1.[ao] The real-time information on the spread and the presence of the additional transmission mechanism allow the FF-model to considerably improve the accuracy of the central forecast in the crisis period. Our results confirm the findings of Del Negro and Schorfheide (2013), who also compare the predictive performance of a standard SW setup with an augmented SW-FF model. They observe that the relative performance of the two models changes over time. On average the model without financial frictions generates more accurate forecasts, but during the recent financial crisis a SW-FF model—that also exploits the timely information on spread and interest rate—produces better forecasts for output and inflation. Del Negro et al. (2014) built on these results and develop a new method for combining predictive densities from recursively estimated models using time-varying weights. As in our RS-approach, this dynamic linear prediction pooling relies on weights that follow an exogenous process. The next step in this research agenda would be to endogenize the occurrence of financial stress periods during which constraints are reinforced and additional feedback mechanisms are activated.[ap]

5.3.2 The Cost Channel of Financial Spreads and Inflation Dynamics

In Section 4.2, when we discussed the economic interpretation of the great recession through the lense of the baseline SW07-model, we observed that the model requires a series of positive mark up shocks to explain the maintained inflation rate during the period of slow recovery and persistent negative output gap. These positive mark up shocks are necessary despite the high estimate of nominal stickiness in price and wage setting. This trend towards more nominal stickiness was already present in the subsample estimates presented by SW07. The high nominal stickiness also plays a crucial role in the explanation of the recent inflation dynamics by Del Negro et al. (2015b) and Fratto and Uhlig (2014). These positive markup shocks disappear completely in our version of the SW07-model, in which we implement the ZLB, and that features an even higher degree of nominal stickiness. The question arises whether this estimated stickiness parameter should be interpreted effectively as a sign of pure nominal stickiness in the price setting practice or whether it reflects some other mechanism that lowered the responsiveness of inflation to the slack in production capacity.

[ao] This somewhat counter-intuitive result of the RS-FF model is related to the nature of the conditional forecast exercise: conditioning on a given spread observation has larger effects when that observation deviates more from the baseline unconditional forecast. The gain from the RS-FF model is precisely that the unconditional forecast will show larger dispersion in the high-FF regime and lower dispersion in the low-FF regime.

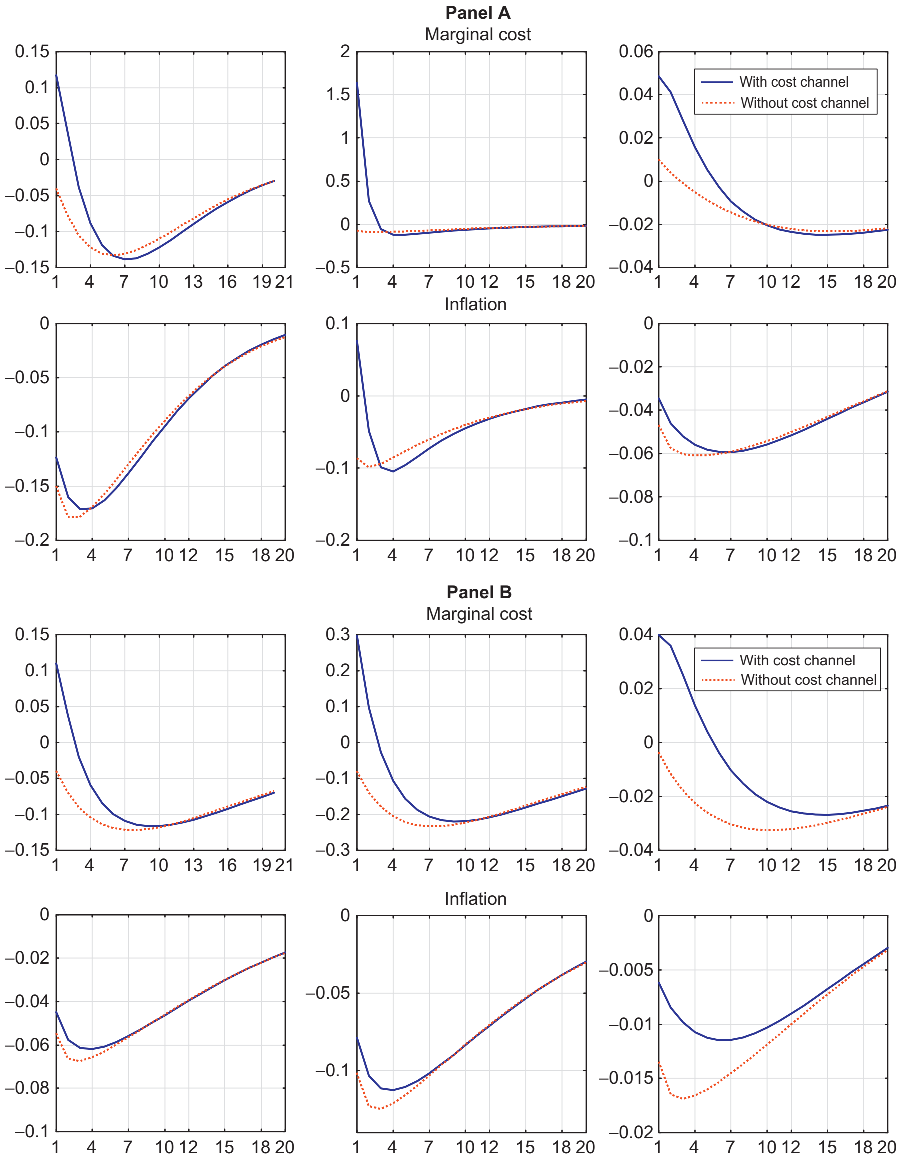
[ap] Various approaches have been developed in this context: Guerrieri and Iacoviello (2013) with occasionally binding constraints, Dewachter and Wouters (2014) with third order nonlinear approximations and Bocola (2013) with a combination of occasionally binding constraints and nonlinear risk premiums.

As noted by Christiano et al. (2015), one mechanism that might contribute to this inflation resilience, in particular during periods of increased financial constraints and high financing costs, is the cost channel. Firms that are financially constrained and that must finance their operations with expensive external capital can experience an increase in their marginal production costs if these financing costs dominate the influence of the other cost components. Related to this cost channel, firms can have other arguments to keep their prices high during periods of financial constraints: high markups can be necessary for firms to generate sufficient cash flow or firms might be forced by their financing constraints to give up on market share (see Gilchrist et al., 2015). Note that this cost channel also plays a crucial role in the explanation of the inflation inertia following a monetary policy shock in Christiano et al. (2005).

Our FF-model contains a parameter that controls the strength of the cost channel. This parameter reflects the fraction of the wage bill that firms have to finance with credit. In this setup, we assume that the external finance premium is also affecting the cost for these intertemporal loans of the firms. In the precrisis model, this fraction of the wage bill on which the financial cost applies is estimated to be quite low (0.33) and the posterior distribution has a large uncertainty margin around this mode. This parameter increases to 0.63 in the complete sample estimation, still with a large uncertainty, but at least there is some indication that the cost channel was more relevant during the recent crisis. To examine the potency of this channel in our model, Fig. 14 plots the impulse response functions of the three shocks that directly affect the external financing costs—the monetary policy shock, the exogenous risk premium shock, and the wealth shock—on the marginal cost and inflation for the two extreme values (zero and one) of the cost channel parameter. Given the large estimation uncertainty around the magnitude of the cost channel parameter, these two extreme values are not completely unlikely and their relevance can probably change depending on the nature of the financial shocks and the constraints. We plot the results for both the precrisis model, with a moderate degree of nominal stickiness, and the full sample ZLB model with a high degree of stickiness.

In both model versions and for all three shocks, it is obvious that marginal cost behaves quite different if the cost channel is fully active compared to a situation in which the cost channel is completely absent. The presence of the cost channel implies that the marginal cost increases at least during the first quarters following each of these shocks. The persistence of this positive effect depends on the type of shock and tends to be shorter for the risk-premium shock and most persistent for the net-worth shock.

The impact on inflation can differ substantially depending on the volatility of the cost shock and on the persistence of the shock relative to the degree of nominal stickiness which determines the degree of forward-lookingness in price setting. In the precrisis model, the exogenous risk-premium shock is highly volatile, but short lived. Combined with the moderate degree of stickiness the cost channel drastically changes the response of inflation to this shock. Inflation rises on impact due to the high risk-premium component



Panel A

Marginal cost

Inflation

Panel B

Marginal cost

Inflation

Fig. 14 The transmission of financial shocks: monetary policy (left column); risk premium (middle column) and net-worth (right column) shock. Panel A: Precrisis model. Panel B: Endogenous ZLB model.

in the financing costs, but the effect is very short lived. In the model with ZLB constraint—with more stickiness—the price setting is more forward looking and the persistence of the shock is crucial. In such a context, the smooth inflation process is dependent on the long-run expected marginal cost. In this case, only the net worth shock has a sufficiently persistent effect on the financing cost to exert a positive impact on inflation; the temporarily high risk free rate and risk premium shock are missing sufficient persistence to have a substantial impact on the inflation dynamics.

From this impulse response analysis, it follows that the cost channel can contribute to the slow response of inflation in a financial crisis context. When the external finance shock for firms are sufficiently high and/or sufficiently persistent, as it is the case for a net worth shock that is expected to have long lasting effects on the financing costs, this inflationary pressure from the cost channel can be quantitatively important. These results illustrate that the financial crisis should not necessarily be viewed as a purely negative aggregate demand shock without an impact on the supply side of the economy. With both aggregate demand and aggregate supply shifting inward by the financial shock, inflation should not necessarily be expected to react that much in a financial crisis situation.

6. STATE OF MACROECONOMIC MODELING: CRITICAL ASSESSMENT AND OUTLOOK

In this section, we conclude by discussing both "new" and "old" challenges for macroeconomic models. As evidenced earlier, the financial crisis has generated new challenges for macroeconomic models used at central banks. When the Great Recession and the financial crisis are included in the estimation sample, we must adjust the specification and empirical estimation strategy of our policy models. Our chapter provides some avenues for moving in that direction, and suggests that the gains of doing so may be considerable. Our suggested modifications have in common of moving away from the standard linear Gaussian setup by including time variation in exogenous and endogenous disturbances. An important short-cut, however, in our adopted Markow Switching framework is that the regime changes are modeled as exogenous events and hence, unrelated to the conduct of policy. At this stage we therefore consider our extensions as a shortcut for truly endogenous nonlinear and state dependent propagation mechanisms. Further progress on the specification of nonlinear methods, solution and filtering techniques, as well as computational techniques, are ongoing for analyzing nonlinear integrated macrofinance models. Together with a broader set of observable variables, these models should allow us to more efficiently identify the nature of shocks, their transmission, and their implications for policy. At this stage, it is important that different theoretical frameworks should be exploited to formulate and validate alternative model specifications.

There were also well-known challenges for central bank models prior to the financial crisis, and they have not been mitigated by the evidence brought forward by the crisis.[aq]

[aq] For instance, the influential work of Del Negro et al. (2007) suggested that workhorse closed economy DSGE models suffered from misspecification problems. Adolfson et al. (2008) confirmed this finding for a standard open economy model.

The balanced growth and the stationarity assumptions provide discipline to the model forecasts, but these long term restrictions often conflict with the observed stochastic trends in many important macro ratios. This mismatch between the theoretical assumptions and the empirical properties can result in overestimation of the persistence in the endogenous frictions and exogenous shocks. It may also be necessary to reevaluate the forecast implications of full information and rational expectations in the models with alternative assumptions about information and expectations formation building on the seminal work of Evans and Honkapohja (2001), Sims (2003, 2010), and Woodford (2014).

Macro models necessarily abstract of many sector details. Recently, a lot of effort have been devoted to model the financial sector. In the standard Smets and Wouters (2007) model analyzed in this chapter, the risk premium shock combines the impact of credit supply conditions, risk aversion, anticipations about future policy actions and the effect of quantitative easing (QE) policies targeting yield curve or risk spreads. Integrating the analysis of financial markets explicitly into general equilibrium is hence of first-order importance, both for firms (the focus of our chapter) and households, eg, along the lines suggested by Iacoviello (2005) and Liu et al. (2013). Other models incorporate an active role for financial institutions in the credit supply process or the asset pricing functions: Christiano et al. (2003a, 2008, 2014), Gerali et al. (2010), and Gertler and Kiyotaki (2010) are inspiring examples. Innovative new macrofinance models, as in, eg, Brunnermeier and Sannikov (2014), He and Krishnamurthy (2012), and Mendoza (2010) suggest that strong endogenous risk and feedback channels between the real and the financial sectors can go a long way in explaining the change in volatility and correlations between tranquil and stress periods. A more explicit recognition of default in both the financial and nonfinancial private sectors as in Clerc et al. (2015) is also an important avenue.

However, other sectors of the economy also have very similar problems in that the exogenous shocks represent a large range of influences that might call for different policy responses depending on the specific underlying distortion or inefficiency. One obvious example is the labor market with very diverging underlying trends in labor participation at intensive and extensive margins, and with shocks and distortions affecting both the labor supply and demand conditions. More work is needed to examine in which dimensions the labor market implications of the standard New Keynesian sticky wage model analyzed by Galí et al. (2011) fall short relative to the data, and if recent work with a more elaborate labor market modeling (see, eg, Gertler et al., 2008; Christiano et al., 2010a; and Christiano et al., 2016) can remedy those shortcomings. Some prominent economists, like Kocherlakota (2009), have recently reiterated that incomplete insurance and heterogeneity in labor and product markets is key for understanding the propagation and welfare costs of business cycles. Thus, the representative agents framework preserved by Gertler et al. and Christiano et al. may not be sufficient in the end, although it represents a clear step forward relative to current generation of policy models.

In a world increasingly integrated through trade of goods and services and more globalized financial markets, policy models also need to be able to account for the impact of foreign shocks. Two old challenges for open economy models is to account for the high degree of observed comovement between real quantities (see, eg, Backus et al., 1992 and Justiniano and Preston, 2010), and the relationship between interest rate differentials and exchange rate movements (ie, the uncovered interest rate parity condition, see, eg, Eichenbaum and Evans, 1995 and Chaboud and Wright, 2005). A voluminous literature deals with these issues, but there is yet no consensus on the "solutions" to these challenges.[ar]

Another key challenge posed for macro models at use in central banks following the crisis is that they have to provide a framework where topical questions can be addressed. First, they have to provide a framework where the central bank can use both conventional monetary policy (manipulating short rates) and unconventional policies (large scale asset purchases (LSAPs) and QE) to affect the economy. A serious treatment of unconventional monetary policy in policy models seems to imply that we have to tackle one old key challenge in macro modeling, namely the failure of the expectations hypothesis (see, eg, Campbell and Shiller, 1991), in favor of environments where the expectations hypothesis does not necessarily hold. One theoretical framework consistent with the idea that large scale asset purchases can reduce term premiums for different maturities and put downward pressure on long-term yields is the theory of preferred habit, see, eg, Andrés et al. (2004) and Vayanos and Vila (2009). Extensions in this direction appear crucial for evaluating the unconventional monetary policy measures during the crisis. Second, apart from analyzing unconventional policies during the crisis, the aftermath of the crisis have brought a renewed focus on financial stability issues, which implies that we need to be able to integrate financial stability considerations into macro models traditionally used for monetary policy analysis only. This involves stress testing exercises and the creation of an environment with an effective role for various macroprudential tools. This requires a more realistic modeling of the interbank market as the one by Boissay et al. (2015). The "3D model" developed by Clerc et al. (2015) and IMF's GIMF model with banks (see Andrle et al., 2015) represent important steps in this direction. Unconventional monetary policy and macroprudential instruments have important distributional effects and this calls for sufficient heterogeneity among agents that are affected by these measures. As mentioned before, the actual and potential budgetary implications of these measures require an explicit modeling of the systematic fiscal reaction function.

[ar] The estimated open economy DSGE model developed by Adolfson et al. (2007b, 2008, 2011) which early on was integrated into operational use at the Riksbank (see Adolfson et al., 2007c) attempted to account for this by modifying the UIP condition following the insights in Duarte and Stockman (2005) and allowing for a common unitroot technology shock.

We believe the benchmark model analyzed in this chapter can serve as the starting point to analyze various extensions for topical questions and policy purposes. Specific model extensions combined with broader set of observed data should help us to better identify the various blocks. This applies equally for the financial, fiscal, labor market and the open economy blocks of the models. Bayesian methodology provides the tools to evaluate and combine these model predictions. In this endeavor, a challenge will be to keep the model size manageable by finding the most parsimonious ways to capture the necessary frictions and shocks, and to understand its implications as the models become increasingly complicated. To keep the models tractable, a critical decision point will be which frictions and shocks that are really needed in the core model, and which features that can be abstracted from in the core model and instead meaningfully analyzed in satellite models. Developing and maintaining empirically validated models with strong theoretical foundations is a daunting task ahead for policy making institutions, even the ones with the most resources.

APPENDICES

A. Linearized Model Representation

In this appendix, we summarize the log-linear equations of the basic SW07-model stated in Section 3. The complete model also includes the seven exogenous shocks $\varepsilon_t^a, \varepsilon_t^b, \varepsilon_t^i, \varepsilon_t^p, \varepsilon_t^w, \varepsilon_t^r$ and g_t, but their processes are not stated here as they were already shown in the main text. Consistent with the notation of the log-linearized *endogenous* variables $\widehat{x}_t = dx_t/x$, the exogenous shocks are denoted with a 'hat', ie, $\widehat{\varepsilon}_t = \ln \varepsilon_t$.

First, we have the consumption Euler equation:

$$\begin{aligned}\widehat{c}_t = & \frac{1}{(1+\varkappa/\gamma)} \mathrm{E}_t \widehat{c}_{t+1} + \frac{\varkappa/\gamma}{(1+\varkappa/\gamma)} \widehat{c}_{t-1} - \frac{1-\varkappa/\gamma}{\sigma_c(1+\varkappa/\gamma)} \\ & (\widehat{R}_t - \mathrm{E}_t \widehat{\pi}_{t+1}) - \frac{(\sigma_c - 1)(w_*^h L/c*)}{\sigma_c(1+\varkappa/\gamma)} (\mathrm{E}_t \widehat{L}_{t+1} - \widehat{L}_t) + \widehat{\varepsilon}_t^b,\end{aligned} \tag{A.1}$$

where $\varkappa$ is the external habit parameter, σ_c the reciprocal of the intertemporal substitution elasticity, $w_*^h L/c_*$ the steady state nominal labor earnings to consumption ratio, and the exogenous risk premium shock $\widehat{\varepsilon}_t^b$ is rescaled so that it enters additive with a unit coefficient.

Next, we have the investment Euler equation:

$$\hat{i}_t = \frac{1}{(1+\overline{\beta}\gamma)} \left(\hat{i}_{t-1} + \overline{\beta}\gamma \mathrm{E}_t \hat{i}_{t+1} + \frac{1}{\gamma^2 \varphi} \widehat{Q}_t^k \right) + \widehat{\varepsilon}_t^q, \tag{A.2}$$

where $\overline{\beta} = \beta\gamma^{-\sigma_c}$, φ is the investment adjustment cost, and the investment specific technology shock $\widehat{\varepsilon}_t^q$ has been rescaled so that it enters linearly with a unit coefficient.

Additionally $i_1 = 1/(1+\beta)$ and $i_2 = i_1/\psi$, where β is the discount factor and ψ is the elasticity of the capital adjustment cost function.

The price of capital is determined by:

$$\widehat{Q}^k_t = -(\widehat{R}_t - \mathrm{E}_t\widehat{\pi}_{t+1}) + q_1\mathrm{E}_t r^k_{t+1} + (1-q_1)\mathrm{E}_t Q^k_{t+1} + \frac{\sigma_c(1+\varkappa/\gamma)}{1-\varkappa/\gamma}\widehat{\varepsilon}^b_t, \tag{A.3}$$

where $q_1 \equiv r^k_*/(r^k_* + (1-\delta))$ in which r^k_* is the steady state rental rate to capital, δ the depreciation rate, and $\widehat{\varepsilon}^b_t$ is multiplied by $\frac{\sigma_c(1+\varkappa/\gamma)}{1-\varkappa/\gamma}$ reflecting the rescaling of this shock in the consumption Euler equation (A.1).

Fourth, we have the optimal condition for the capital utilization rate $\hat{u}_t$:

$$\hat{u}_t = (1-\psi)/\psi\widehat{r}^k_t, \tag{A.4}$$

where ψ is the elasticity of the capital utilization cost function and capital services used in production $(\widehat{k}_t)$ is defined as:

$$\widehat{k}_t = \hat{u}_t + \widehat{\bar{k}}_{t-1}, \tag{A.5}$$

where $\widehat{\bar{k}}_{t-1}$ is the physical capital stock which evolves according to the capital accumulation equation:

$$\widehat{\bar{k}}_t = \kappa_1 \; \widehat{\bar{k}}_{t-1} + (1-\kappa_1)\hat{i}_t + \kappa_2\widehat{\varepsilon}^q_t \tag{A.6}$$

with $\kappa_1 = (1-(i_*/\bar{k}_*)$ and $\kappa_2 = (i_*/\bar{k}_*)\gamma^2\varphi$.

The following optimal capital/labor input condition also holds:

$$\widehat{k}_t = \widehat{w}_t - \; \widehat{r}^k_t + \widehat{L}_t, \tag{A.7}$$

where $\widehat{w}_t$ is the real wage.

The log-linearized production function is given by:

$$\widehat{y}_t = \phi_p \;\; (\alpha\widehat{k}_t + (1-\alpha)\widehat{L}_t + \widehat{\varepsilon}^a_t), \tag{A.8}$$

in which ϕ_p is the fixed costs of production corresponding to the gross price markup in the steady state, and $\widehat{\varepsilon}^a_t$ is the exogenous TFP process.

Aggregate demand must equal aggregate supply:

$$\widehat{y}_t = \frac{c_*}{y_*}\widehat{c}_t + \frac{i_*}{y_*}\hat{i}_t + g_t + \frac{r^k_* k_*}{y_*}\hat{u}_t, \tag{A.9}$$

where g_t represents the exogenous demand component.

Next, we have the following log-linearized price-setting equation with dynamic indexation ι_p:

$$\widehat{\pi}_t - \iota_p\widehat{\pi}_{t-1} = \pi_1\left(\mathrm{E}_t\widehat{\pi}_{t+1} - \iota_p\widehat{\pi}_t\right) - \pi_2\widehat{\mu}^p_t + \widehat{\varepsilon}^p_t, \tag{A.10}$$

where $\pi_1 = \beta$, $\pi_2 = (1 - \xi_p\beta)(1 - \xi_p)/[\xi_p(1 + (\phi_p - 1)\epsilon_p)]$, $1 - \xi_p$ is the probability of each firm being able to reoptimize the price each period, ϵ_p is the curvature of the aggregator function Eq. (2), and the markup shock $\widehat{\varepsilon}_t^p$ has been rescaled to enter with a unit coefficient. The price markup $\widehat{\mu}_t^p$ equals the inverse of the real marginal cost, $\widehat{\mu}_t^p = -\widehat{mc}_t$, which in turn is given by:

$$\widehat{mc}_t = (1-\alpha)\ \ \widehat{w}_t^{real} + \alpha\ \ \widehat{r}_t^k - \widehat{\varepsilon}_t^a. \tag{A.11}$$

We also have the following wage-setting equation allowing for dynamic indexation of wages for nonoptimizing households:

$$\begin{aligned}(1+\overline{\beta}\gamma)\widehat{w}_t^{real} - \widehat{w}_{t-1}^{real} - \overline{\beta}\gamma \mathrm{E}_t\widehat{w}_{t+1}^{real} = {} & \frac{(1-\xi_w\overline{\beta}\gamma)(1-\xi_w)}{[\xi_w(1+(\phi_w-1)\epsilon_w)]} \\ & \left(\frac{1}{1-\varkappa/\gamma}\widehat{c}_t - \frac{\varkappa/\gamma}{1-\varkappa/\gamma}\widehat{c}_{t-1} + \sigma_l\widehat{L}_t - \widehat{w}_t\right) \\ & -(1+\overline{\beta}\gamma\iota_w)\widehat{\pi}_t + \iota_w\widehat{\pi}_{t-1} + \overline{\beta}\gamma\mathrm{E}_t\widehat{\pi}_{t+1} + \widehat{\varepsilon}_t^w,\end{aligned} \tag{A.12}$$

where ϕ_w the gross wage markup, $1 - \xi_p$ is the probability of each household being able to reoptimize its wage each period, ϵ_w is the curvature of the aggregator function (eq. 7), and σ_l determines the elasticity of labor supply given σ_c (see Eq. (9)). The exogenous wage markup shock $\widehat{\varepsilon}_t^w$ has been rescaled to enter linearly with a unit coefficient.

Finally, we have the monetary policy rule:

$$\widehat{R}_t = \rho_R\widehat{R}_{t-1} + (1-\rho_R)\left(r_\pi\widehat{\pi}_t + r_y\widehat{y}_t^{gap} + r_{\Delta y}\Delta\widehat{y}_t^{gap}\right) + \widehat{\varepsilon}_t^r, \tag{A.13}$$

where $\widehat{y}_t^{gap} = \widehat{y}_t - \widehat{y}_t^{pot}$, or in words: the difference between actual output and the output prevailing in the flexible price and wage economy in absence of the inefficient price and wage markup shocks. We solve for $\widehat{y}_t^{pot}$ by setting $\xi_p = \xi_w = 0$ (or arbitrary close to nil) and removing $\widehat{\varepsilon}_t^w$ and $\widehat{\varepsilon}_t^p$ from the system of equations given by (A.1)–(A.13). Note that when we impose the ZLB on the model, Eq. (A.13) is replaced by Eq. (17).

B. The ZLB Algorithm and the Likelihood Function

This appendix provides some details on the ZLB algorithm we use and how the likelihood function takes the ZLB into account. For more details on the ZLB algorithm we refer to Hebden et al. (2010), whereas more details on the computation of the likelihood is provided by Jesper et al. (2016).

B.1 The ZLB Algorithm

The DSGE model can be written in the following practical state-space form,

$$\begin{bmatrix} X_{t+1} \\ Hx_{t+1|t} \end{bmatrix} = A \begin{bmatrix} X_t \\ x_t \end{bmatrix} + Bi_t + \begin{bmatrix} C \\ 0 \end{bmatrix} \varepsilon_{t+1}. \tag{B.1}$$

Here, X_t is an n_X–vector of *predetermined* variables in period t (where the period is a quarter) and x_t is a n_x–vector of *forward-looking* variables. The i_t is generally a n_i–vector of (policy) *instruments* but in the cases examined here it is a scalar—the central bank's policy rate—giving $n_i = 1$. The ε_t is an n_ε-vector of independent and identically distributed shocks with mean zero and covariance matrix I_{n_ε}, while A, B, C, and H are matrices of the appropriate dimension. Lastly $x_{t+\tau|t}$ denotes $\mathrm{E}_t x_{t+\tau}$, ie, the rational expectation of $x_{t+\tau}$ conditional on information available in period t. The forward-looking variables and the instruments are the *nonpredetermined* variables.[as]

The variables are measured as differences from steady state values, in which case their unconditional means are zero. In addition, the elements of the matrices A, B, C, and H are considered fixed and known.

We let i_t^* denote the policy rate when we disregard the ZLB. We call it the *unrestricted* policy rate. We let i_t denote the actual or *restricted* policy rate that satisfies the ZLB,

$$i_t + \bar{\imath} \geq 0,$$

where $\bar{\imath} > 0$ denotes the steady state level of the policy rate and we use the convention that i_t and i_t^* are expressed as deviations from the steady state level. The ZLB can therefore be written as

$$i_t + \bar{\imath} = \max\{i_t^* + \bar{\imath}, 0\}. \tag{B.2}$$

We assume the unrestricted policy rate follows the (possibly reduced form) unrestricted linear policy rule,

$$i_t^* = f_X X_t + f_x x_t, \tag{B.3}$$

where f_X and f_x are row vectors of dimension n_X and n_x, respectively. From (B.2) it then follows that the restricted policy rate is given by:

$$i_t + \bar{\imath} = \max\{f_X X_t + f_x x_t + \bar{\imath}, 0\}. \tag{B.4}$$

Consider now a situation in period $t \geq 0$ where the ZLB may be binding in the current or the next finite number T periods but not beyond period $t + T$. That is, the ZLB constraint

$$i_{t+\tau} + \bar{\imath} \geq 0, \qquad \tau = 0, 1, \ldots, T \tag{B.5}$$

may be binding for some $\tau \leq T$, but we assume that it is not binding for $\tau > T$,

[as] A variable is predetermined if its one-period-ahead prediction error is an exogenous stochastic process (Klein, 2000). For (B.1), the one-period-ahead prediction error of the predetermined variables is the stochastic vector $C\varepsilon_{t+1}$.

$$i_{t+\tau} + \bar{\imath} > 0, \qquad \tau > T.$$

We will implement the ZLB with anticipated shocks to the unrestricted policy rule, using the techniques of Laséen and Svensson (2011). Thus, we let the restricted and unrestricted policy rate in each period t satisfy

$$i_{t+\tau,t} = i^{*}_{t+\tau,t} + z_{t+\tau,t}, \tag{B.6}$$

for $\tau \geq 0$. The ZLB policy rule in (B.4)—as we explain in further detail later—implies that all current and future anticipated shocks $z_{t+\tau,t}$ in (B.6) must be nonnegative, and that $z_{t,t}$ is strictly positive in periods when the ZLB is binding.

Disregarding for the moment when z_t are nonnegative, we follow Laséen and Svensson (2011) and call the stochastic variable z_t the deviation and let the $(T+1)$-vector $z^t \equiv (z_{t,t}, z_{t+1,t}, \ldots, z_{t+T,t})'$ denote a projection in period t of future realizations $z_{t+\tau}$, $\tau = 0, 1, \ldots, T$, of the deviation. Furthermore, we assume that the deviation satisfies

$$z_t = \eta_{t,t} + \sum_{s=1}^{T} \eta_{t,t-s}$$

for $T \geq 0$, where $\eta^t \equiv (\eta_{t,t}, \eta_{t+1,t}, \ldots, \eta_{t+T,t})'$ is a $(T+1)$-vector realized in the beginning of period t. For $T = 0$, the deviation is given by $z_t = \eta_t$. For $T > 0$, the deviation is given by the moving-average process

$$z_{t+\tau,t+1} = z_{t+\tau,t} + \eta_{t+\tau,t+1}$$
$$z_{t+\tau+T+1,t+1} = \eta_{t+T+1,t+1},$$

where $\tau = 1, \ldots, T$. It follows that the dynamics of the projection of the deviation can be written more compactly as

$$z^{t+1} = A_z z^t + \eta^{t+1}, \tag{B.7}$$

where the $(T+1) \times (T+1)$ matrix A_z is defined as

$$A_z \equiv \begin{bmatrix} 0_{T\times 1} & I_T \\ 0 & 0_{1\times T} \end{bmatrix}.$$

Hence, z^t is the projection in period t of current and future deviations, and the innovation η^t can be interpreted as the new information received in the beginning of period t about those deviations.

Let us now combine the model, (B.1), the dynamics of the deviation, (B.7), the unrestricted policy rule, (B.3), and the relation (B.6). Taking the starting period to be $t = 0$, we can then write the combined model as

$$\begin{bmatrix} \tilde{X}_{t+1} \\ \tilde{H}\tilde{x}_{t+1|t} \end{bmatrix} = \tilde{A} \begin{bmatrix} \tilde{X}_t \\ \tilde{x}_t \end{bmatrix} + \begin{bmatrix} C & 0_{n_X \times (T+1)} \\ 0_{(T+1)\times n_\varepsilon} & I_{T+1} \\ 0_{(n_x+2)\times n_\varepsilon} & 0_{(n_x+2)\times(T+1)} \end{bmatrix} \begin{bmatrix} \varepsilon_{t+1} \\ \eta^{t+1} \end{bmatrix} \tag{B.8}$$

for $t \geq 0$, where

$$\tilde{X}_t \equiv \begin{bmatrix} X_t \\ z^t \end{bmatrix}, \quad \tilde{x}_t \equiv \begin{bmatrix} x_t \\ i_t^* \\ i_t \end{bmatrix}, \quad \tilde{H} \equiv \begin{bmatrix} H & 0_{n_x \times 1} & 0_{n_x \times 1} \\ 0_{1 \times n_x} & 0 & 0 \\ 0_{1 \times n_x} & 0 & 0 \end{bmatrix}.$$

Under the standard assumption of the saddle-point property (that the number of eigenvalues of $\tilde{A}$ with modulus larger than unity equals the number of nonpredetermined variables, here $n_x + 2$), the system of difference equations (B.8) has a unique solution and there exist unique matrices M and F returned by the Klein (2000) algorithm such that the solution can be written:

$$\tilde{x}_t = F\tilde{X}_t \equiv \begin{bmatrix} F_x \\ F_{i^*} \\ F_i \end{bmatrix} \tilde{X}_t \tilde{X}_{t+1} = M\tilde{X}_t + \begin{bmatrix} C\varepsilon_{t+1} \\ \eta^{t+1} \end{bmatrix} \equiv \begin{bmatrix} M_{XX} & M_{Xz} \\ 0_{(T+1)\times n_X} & A_z \end{bmatrix} \begin{bmatrix} X_t \\ z^t \end{bmatrix} + \begin{bmatrix} C\varepsilon_{t+1} \\ \eta^{t+1} \end{bmatrix},$$

for $t \geq 0$, and where X_0 in $\tilde{X}_0 \equiv (X_0', z^{0\prime})'$ is given but the projections of the deviation z^0 and the innovations η^t for $t \geq 1$ (and thereby z^t for $t \geq 1$) remain to be determined. They will be determined such that the ZLB is satisfied, ie, Eq. (B.4) holds. Thus, the *policy-rate projection* is given by

$$i_{t+\tau,t} = F_i M^\tau \begin{bmatrix} X_t \\ z^t \end{bmatrix} \tag{B.9}$$

for $\tau \geq 0$ and for given X_t and z^t.

We will now show how to determine the $(T+1)$-vector $z^t \equiv (z_t, z_{t+1,t}, \ldots, z_{t+T,t})'$, ie, the projection of the deviation, such that policy-rate projection satisfies the ZLB restriction (B.5) and the policy rule (B.4).

When the ZLB restriction (B.5) is disregarded or not binding, the policy-rate projection in period t is given by

$$i_{t+\tau,t} = F_i M^\tau \begin{bmatrix} X_t \\ 0_{(T+1)\times 1} \end{bmatrix}, \qquad \tau \geq 0. \tag{B.10}$$

The policy-rate projection disregarding the ZLB hence depends on the initial state of the economy in period t, represented by the vector of predetermined variables X_t. If the ZLB is disregarded, or not binding for any $\tau \geq 0$, the projections of the restricted and unrestricted policy rates will be the same,

$$i_{t+\tau,t} = i^*_{t+\tau,t} = f_X X_{t+\tau,t} + f_x x_{t+\tau,t}, \qquad \tau \geq 0.$$

Assume now that the policy-rate projection according to (B.10) violates the ZLB for one or several periods, that is,

$$i_{t+\tau,t} + \bar{\imath} < 0, \quad \text{for some } \tau \text{ in the interval } 0 \leq \tau \leq T. \tag{B.11}$$

In order to satisfy the ZLB, we then want to find a projection of the deviation z^t such that the policy-rate projection satisfies (B.5) and

$$i_{t+\tau,t} + \bar{\imath} = \max\{i^*_{t+\tau,t} + \bar{\imath}, 0\} = \max\{f_X X_{t+\tau,t} + f_x x_{t+\tau,t} + \bar{\imath}, 0\} \tag{B.12}$$

for $\tau \geq 0$. This requires that the projection of the deviation satisfies a nonnegativity constraint

$$z_{t+\tau,t} \geq 0, \qquad \tau \geq 0, \tag{B.13}$$

and that the policy-rate projection and the projection of the deviation satisfies the complementary-slackness condition

$$(i_{t+\tau,t} + \bar{\imath}) z_{t+\tau,t} = 0, \qquad \tau \geq 0. \tag{B.14}$$

Notice that the complementary-slackness condition implies that $z_{t+\tau,t} = 0$ if $i_{t+\tau,t} + \bar{\imath} > 0$.

For given X_t, we now proceed under the presumption that there exists a unique projection of the deviation z^t that satisfies (B.9) and (B.12)–(B.14).[at] We call this projection of the deviation and the corresponding policy-rate projection the *equilibrium*projection. This projection of the deviation either has all elements equal to zero (in which case the ZLB is not binding for any period) or has some elements positive and other elements zero. Let

$$\mathcal{T}_t \equiv \{0 \leq \tau \leq T \mid z_{t+\tau,t} > 0\}$$

denote the set of periods for which the projection of the deviation are positive in equilibrium.

For each $\tau \in \mathcal{T}_t$, the solution will satisfy

$$i_{t+\tau,t} + \bar{\imath} = F_i M^\tau \begin{bmatrix} X_t \\ z^t \end{bmatrix} + \bar{\imath} = 0 \quad \text{for} \quad \tau \in \mathcal{T}_t. \tag{B.15}$$

Let $n_{\mathcal{T}_t}$ denote the number of elements of $\mathcal{T}_t$, that is, the number of periods that the ZLB binds. The equation system (B.15) then has $n_{\mathcal{T}}$ equations to determine the $n_{\mathcal{T}}$ elements of z^t that are positive. From the system (B.15), it is clear that the solution for z^t and the set $\mathcal{T}_t$ will depend on X_t as well as the initial situation, and thereby also on the initial innovation ε_t. For other periods (that is $\tau \notin \mathcal{T}_t$), the ZLB will not be binding and the elements in z^t will be zero. The equation system (B.15) and the periods in the set $\mathcal{T}_t$ hence refer to the periods where the ZLB is *strictly* binding, that is, when $z_{t+\tau,t}$ is positive. Furthermore, it is important to notice that the set of periods τ in (B.11), for which the policy-rate projection (B.10) violates the ZLB, is not necessarily the same as the set of periods $\mathcal{T}_t$ for which the ZLB is strictly binding *in equilibrium*. That is because the projections of

[at] This assumption is discussed in further detail in Hebden et al. (2010).

the predetermined and forward-looking variables $X_{t+\tau,t}$ and $x_{t+\tau,t}$, that determine the unrestricted policy rate differ, depending on whether z^t is zero or not. This means that the whole policy-rate path is affected when the ZLB is imposed.

The difficulty in imposing the ZLB is to find the set $\mathcal{T}_t$ for which the ZLB is strictly binding in equilibrium, that is, to find the periods for which the equation system (B.15) applies. Once this is done, solving the equation system (B.15) is trivial. Hebden et al. (2010) outline a simple shooting algorithm to find the set $\mathcal{T}_t$.

B.2 Computation of the Likelihood Function

To compute the likelihood function, we follow the general idea outlined by Maih (2010). Maih's algorithm allows us to add anticipated policy shocks (using the algorithm outlined earlier) to the state space formulation of the model and filter those shocks with the Kalman filter to impose the zero lower bound on policy rates in the estimation. The appealing feature of Maih's algorithm is that it does not require us to include standard deviations for each of the anticipated policy shocks. Thus, the log-marginal likelihood can be directly compared to the models which does not impose the ZLB. For further details on the computation of the likelihood function in the face of the ZLB constraint, we refer to Lindé et al. (2016).

C. Data

In this appendix, we provide the sources on the data we use in the analysis.

C.1 Benchmark Model

The benchmark model is estimated using seven key macroeconomic time series: real GDP, consumption, investment, hours worked, real wages, prices, and a short-term interest rate. The Bayesian estimation methodology is extensively discussed by Smets and Wouters (2003). GDP, consumption and investment were taken from the US Department of Commerce—Bureau of Economic Analysis data-bank—on September 25, 2014. Real gross domestic product is expressed in billions of chained 2009 dollars. Nominal personal consumption expenditures and fixed private domestic investment are deflated with the GDP-deflator. Inflation is the first difference of the log of the implicit price deflator of GDP. Hours and wages come from the BLS (hours and hourly compensation for the nonfarm business, NFB, sector for all persons). Hourly compensation is divided by the GDP price deflator in order to get the real wage variable. Hours are adjusted to take into account the limited coverage of the NFB sector compared to GDP (the index of average hours for the NFB sector is multiplied with the Civilian Employment (16 years and over). The aggregate real variables are expressed per capita by dividing with the population size aged 16 or older. All series are seasonally adjusted. The interest rate is the Federal Funds Rate. Consumption, investment, GDP, wages, and hours are expressed in 100× log. The interest rate and inflation rate are expressed on a quarterly basis during the estimation (corresponding with their appearance in the model), but in the figures the series are reported on an annualized (400× first log difference) or yearly (100× the four-quarter log difference) basis.

C.2 Model with Financial Frictions

The first seven variables are exactly those used to estimate the benchmark model, which are described in Appendix C.1. In addition to those series, this model features an interest rate spread. Following Bernanke et al. (1999), this spread is measured as the difference between the BAA corporate interest rate and the US 10-year government yield.

ACKNOWLEDGMENTS

We thank the editors John B. Taylor and Harald Uhlig, as well as our discussant Jonas Fisher, for very helpful suggestions that significantly improved the chapter. Comments by participants in the Handbook conference at the Becker-Friedman institute in Chicago, the 18th Central Bank Macro Modeling Workshop (hosted by the Reserve Bank of New Zealand in Wellington), and seminars at the ECB and Sveriges Riksbank were also useful. We are heavily indebted to Junior Maih for developing code in RISE which allowed us to estimate models subject to the zero lower bound. A special thank you is also due to Mattias Villani, who allowed us to use his BVAR-code in which we could set steady state priors. Leonard Voltaire and Jessica Radeschnig provided excellent research assistance, but all remaining errors are ours. Frank Smets and Raf Wouters acknowledge financial support from the EU FP7-Project MACFINROBODS, grant 612796. The views, analysis, and conclusions in this chapter are solely the responsibility of the authors and do not necessarily agree with ECB, National Bank of Belgium, or the Sveriges Riksbank, or those of any other person associated with these institutions.

REFERENCES

Adam, K., Billi, R., 2006. Optimal monetary policy under commitment with a zero bound on nominal interest rates. J. Money Credit Bank. 38 (7), 1877–1906.

Adjemian, S., Bastani, H., Juillard, M., Karamé, F., Mihoubi, F., Perendia, G., Pfeifer, J., Ratto, M., Villemot, S., 2011. Dynare: Reference Manual, Version 4. Dynare Working Papers 1, CEPREMAP.

Adolfson, M., Laséen, S., Lindé, J., Villani, M., 2005. The role of sticky prices in an open economy DSGE model: a bayesian investigation. J. Eur. Econ. Assoc. Pap. Proc. 3 (2-3), 444–457.

Adolfson, M., Andersson, M.K., Lindé, J., Villani, M., Vredin, A., 2007a. Modern forecasting models in action: improving macroeconomic analyses at central banks. Int. J. Cent. Bank. 3 (4), 111–144.

Adolfson, M., Laséen, S., Lindé, J., Villani, M., 2007b. Bayesian estimation of an open economy DSGE model with incomplete pass-through. J. Int. Econ. 72, 481–511.

Adolfson, M., Laséen, S., Lindé, J., Villani, M., 2007c. RAMSES–a new general equilibrium model for monetary policy analysis. Sveriges Riksbank Econ. Rev. 2, 5–40.

Adolfson, M., Lindé, J., Villani, M., 2007d. Forecasting performance of an open economy DSGE model. Econ. Rev. 26, 289–328.

Adolfson, M., Laséen, S., Lindé, J., Villani, M., 2008. Evaluating an estimated new Keynesian small open economy model. J. Econ. Dyn. Control. 32 (8), 2690–2721.

Adolfson, M., Laséen, S., Lindé, J., Svensson, L.E., 2011. Optimal monetary policy in an operational medium-sized model. J. Money Credit Bank. 43 (7), 1287–1330.

Adolfson, M., Laséen, S., Christiano, L.J., Trabandt, M., Walentin, K., 2013. Ramses II–model description. Sveriges Riksbank Occasional Paper Series No. 12.

Altig, D., Christiano, L., Eichenbaum, M., Lindé, J., 2011. Firm-specific capital, nominal rigidities and the business cycle. Rev. Econ. Dyn. 14 (2), 225–247.

Andrade, P., Gaballoy, G., Mengusz, E., Mojon, B., 2015. Forward guidance and heterogeneous beliefs. Banque de France Working Paper Series No. 573.

Andrés, J., López-Salido, J.D., Nelson, E., 2004. Tobin's imperfect asset substitution in optimizing general equilibrium. J. Money Credit Bank. 36 (4), 666–690.

Andrle, M., Kumhof, M., Laxton, D., Muir, D., 2015. Banks in the global integrated monetary and fiscal model. IMF Working Paper No. 15-150.

Anzoategui, D., Comin, D., Gertler, M., Martinez, J., 2015. Endogenous technology adoption and R&D as sources of business cycle persistence. University Working Paper, New York.

Backus, D.K., Kehoe, P.J., Kydland, F.E., 1992. International real business cycles. J. Polit. Econ. 100, 745–773.

Barro, R.J., 1974. Are government bonds net wealth? J. Polit. Econ. 82 (6), 1095–1117.

Benigno, P., Nisticó, S., 2015. Non-neutrality of open-market operations. CEPR Working Paper No. 10594.

Bernanke, B.S., 2013. Communication and monetary policy. In: Herbert Stein Memorial Lecture at the National Economists Club Annual Dinner, November 19, Washington, DC.

Bernanke, B., Gertler, M., Gilchrist, S., 1999. The financial accelerator in a quantitative business cycle framework. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics. North-Holland/Elsevier Science, New York.

Blanchard, O., Kahn, C.M., 1980. The solution of linear difference models under rational expectations. Econometrica 48, 1305–1313.

Bocola, L., 2013. The pass-through of sovereign risk. University of Pennsylvania, manuscript.

Boissay, F., Collard, F., Smets, F., 2015. Booms and banking crises. J. Polit. Econ. 124 (2), 489–538.

Boppart, T., Krusell, P., 2015. Labor supply in the past, present, and future: a balanced-growth perspective. Stockholm University, manuscript.

Brave, S.A., Campbell, J.R., Fisher, J.D., Justiniano, A., 2012. The Chicago fed DSGE model. Federal Reserve Bank of Chicago Working Paper No. 2012-02.

Brubakk, L., Husebø, T., Maih, J., Olsen, K., Østnor, M., 2006. Finding NEMO: documentation of the Norwegian economy model. Staff Memo 2006/6, Norges Bank.

Brunnermeier, M.K., Sannikov, Y., 2014. A macroeconomic model with a financial sector. Am. Econ. Rev. 104 (2), 379–421.

Burgess, S., Fernandez-Corugedo, E., Groth, C., Harrison, R., Monti, F., Theodoridis, K., Waldron, M., 2013. The Bank of England's forecasting platform: COMPASS, MAPS, EASE and the suite of models. Bank of England Working Paper No. 471.

Calvo, G., 1983. Staggered prices in a utility maximizing framework. J. Monet. Econ. 12, 383–398.

Campbell, J.Y., Shiller, R.J., 1991. Yield spreads and interest rate movements: a bird's eye view. Rev. Econ. Stud. 58, 495–514.

Campbell, J.R., Evans, C.L., Fisher, J.D.M., Justiniano, A., 2012. Macroeconomic effects of federal reserve forward guidance. Brook. Pap. Econ. Act. 1–80 (Spring issue).

Carlstrom, C., Fuerst, T., Paustian, M., 2012. Inflation and output in new Keynesian models with a transient interest rate peg. Bank of England Working Paper No. 459.

Chaboud, A.P., Wright, J.H., 2005. Uncovered interest parity: it works, but not for long. J. Int. Econ. 66 (2), 349–362.

Chari, V., Kehoe, P.J., McGrattan, E.R., 2009. New Keynesian models: not yet useful for policy analysis. Am. Econ. J. Macroecon. 1 (1), 242–266.

Chib, S., Ramamurthy, S., 2014. DSGE models with Student-t errors. Econ. Rev. 33 (1-4), 152–171.

Christiano, L., Motto, R., Rostagno, M., 2003a. The great depression and the Friedman-Schwartz hypothesis. J. Money Credit Bank. 35 (6), 1119–1197.

Christiano, L.J., Eichenbaum, M., Vigfusson, R.J., 2003b. What happens after a technology shock? NBER Working Paper Series No. 9819.

Christiano, L.J., Eichenbaum, M., Evans, C., 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. J. Polit. Econ. 113 (1), 1–45.

Christiano, L., Trabandt, M., Walentin, K., 2007. Introducing financial frictions and unemployment into a small open economy model. Sveriges Riksbank Working Paper Series No. 214.

Christiano, L., Motto, R., Rostagno, M., 2008. Shocks, structures or monetary policies? The Euro area and the US after 2001. J. Econ. Dyn. Control. 32 (8), 2476–2506.

Christiano, L., Trabandt, M., Walentin, K., 2010a. Involuntary unemployment and the business cycle. Sveriges Riksbank Working Paper Series No. 238.

Christiano, L.J., Ilut, C., Motto, R., Rostagno, M., 2010b. Monetary policy and stock market booms. In: Proceedings–Economic Policy Symposium–Jackson Hole, Federal Reserve Bank of Kansas City, pp. 85–145.

Christiano, L.J., Motto, R., Rostagno, M., 2014. Risk shocks. Am. Econ. Rev. 104 (1), 27–65.

Christiano, L.J., Eichenbaum, M., Trabandt, M., 2015. Understanding the great recession. Am. Econ. J. Macroecon. 7 (1), 110–167.

Christiano, L.J., Eichenbaum, M., Trabandt, M., 2016. Unemployment and business cycles. Econometrica. (forthcoming in Vol. 84, No.3, 1289).

Chung, H., Laforte, J.P., Reifschneider, D., 2012. Have we underestimated the likelihood and severity of zero lower bound events? J. Money Credit Bank. 44 (2012), 47–82.

Clarida, R., Galí, J., Gertler, M., 1999. The science of monetary policy: a new Keynesian perspective. J. Econ. Lit. 37 (4), 1661–1707.

Clerc, L., Derviz, A., Mendicino, C., Moyen, S., Nikolov, K., Stracca, L., Suarez, J., Vardoulakis, A.P., 2015. Capital regulation in a macroeconomic model with three layers of default. Int. J. Cent. Bank. 15 (3), 9–63.

Coenen, G., Erceg, C., Freedman, C., Furceri, D., Kumhof, M., Lalonde, R., Laxton, D., Lindé, J., Mourougane, A., Muir, D., Mursula, S., de Resende, C., Roberts, J., Roeger, W., Snudden, S., Trabandt, M., in't Veld, J., 2012. Effects of fiscal stimulus in structural models. Am. Econ. J. Macroecon. 4 (1), 22–68.

Curdia, V., Del Negro, M., Greenwald, D.L., 2014. Rare shocks, great recessions. J. Econ. 29 (7), 1031–1052.

De Graeve, F., 2008. The external finance premium and the macroeconomy: US post-WWII evidence. J. Econ. Dyn. Control. 32 (11), 3415–3440.

Del Negro, M., Schorfheide, F., 2013. DSGE model-based forecasting. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting, vol. 2. Elseiver, Amsterdam, pp. 57–140.

Del Negro, M., Sims, C.A., 2014. When does a central bank's balance sheet require fiscal support? FRB of New York Staff Report No. 701.

Del Negro, M., Schorfheide, F., Smets, F., Wouters, R., 2007. On the fit of new Keynesian models. J. Bus. Econ. Stat. 25 (2), 123–162.

Del Negro, M., Eusepi, S., Giannoni, M., Sbordone, A., Tambalotti, A., Cocci, M., Hasegawa, R., Henry Linder, M., 2013. The FRBNY DSGE model. Federal Reserve Bank of New York Staff Report No. 647.

Del Negro, M., Hasegawa, R., Schorfheide, F., 2014. Dynamic prediction pools: an investigation of financial frictions and forecasting performance. NBER Working Paper 20575.

Del Negro, M., Giannoni, M.P., Patterson, C., 2015a. The forward guidance puzzle. Federal Reserve Bank of New York Staff Reports No. 574.

Del Negro, M., Giannoni, M.P., Schorfheide, F., 2015b. Inflation in the great recession and new Keynesian models. Am. Econ. J. Macroecon. 7 (1), 168–196.

Dewachter, H., Wouters, R., 2014. Endogenous risk in a DSGE model with capital-constrained financial intermediaries. J. Econ. Dyn. Control. 43 (C), 241–268.

Doan, T., Litterman, R., Sims, C.A., 1984. Forecasting and conditional projection using realistic prior distributions. Econ. Rev. 3 (1), 1–100.

Dotsey, M., King, R.G., 2005. Implications of state dependent pricing for dynamic macroeconomic models. J. Monet. Econ. 52, 213–242.

Duarte, M., Stockman, A., 2005. Rational speculation and exchange rates. J. Monet. Econ. 52, 3–29.

Edge, R.M., Gürkaynak, R., 2010. How useful are estimated DSGE model forecasts for central bankers? Brook. Pap. Econ. Act. 2, 209–244.

Eggertsson, G., Woodford, M., 2003. The zero bound on interest rates and optimal monetary policy. Brook. Pap. Econ. Act. 1, 139–211.

Eichenbaum, M., Evans, C.L., 1995. Some empirical evidence on the effects of shocks to monetary policy on exchange rates. Q. J. Econ. 110 (4), 975–1009.

Erceg, C.J., Lindé, J., 2010. Is there a fiscal free lunch in a liquidity trap? CEPR Discussion Paper Series No. 7624.

Erceg, C.J., Henderson, D.W., Levin, A.T., 2000. Optimal monetary policy with staggered wage and price contracts. J. Monet. Econ. 46, 281–313.

Evans, G.E., Honkapohja, S., 2001. Learning and Expectations in Macroeconomics. Princeton University Press, Princeton.

Fair, R.C., Taylor, J.B., 1983. Solution and maximum likelihood estimation of dynamic nonlinear a rational expecations models. Econometrica 51 (4), 1169–1185.

Fernald, J., 2012. A quarterly, utilization-adjusted series on total factor productivity. Federal Reserve Bank of San Francisco Working Paper 2012-19.

Fernández-Villaverde, J., Rubio-Ramírez, J., 2007. Estimating macroeconomic models: a likelihood approach. Rev. Econ. Stud. 74, 1059–1087.

Fernández-Villaverde, J., Guerrón-Quintana, P.A., Kuester, K., Rubio-Ramírez, J., 2011. Fiscal volatility shocks and economic activity. NBER Working Paper 17317.

Fisher, J.D., 2006. The dynamic effects of neutral and investment-specific technology shocks. J. Polit. Econ. 114 (3), 413–451.

Fisher, J.D., 2015. On the structural interpretation of the Smets–Wouters "risk premium" shock. J. Money Credit Bank. 47 (2-3), 511–516.

Fratto, C., Uhlig, H., 2014. Accounting for post-crisis inflation and employment: a retro analysis. NBER Working Paper No. 20707.

Galí, J., 2008. Monetary Policy, Inflation and the Business Cycle: An Introduction to the New Keynesian Framework. Princeton University Press, Princeton.

Galí, J., Gertler, M., 1999. Inflation dynamics: a structural econometric analysis. J. Monet. Econ. 44, 195–220.

Galí, J., Pau, R., 2004. Technology shocks and aggregate fluctuations: how well does the RBC model fit postwar U.S. data? NBER Macroeconomics Annual.

Galí, J., Gertler, M., López-Salido, D., 2001. European inflation dynamics. Eur. Econ. Rev. 45, 1237–1270.

Galí, J., Smets, F., Wouters, R., 2011. Unemployment in an estimated new Keynesian model. NBER Macroeconomics Annual.

Gerali, A., Neri, S., Sessa, L., Signoretti, F.M., 2010. Credit and banking in a DSGE model of the Euro area. J. Money Credit Bank. 42, 107–141.

Gertler, M., Kiyotaki, N., 2010. Financial intermediation and credit policy in business cycle analysis. In: Friedman, B.M., Woodford, M. (Eds.), Handbook of Monetary Economics, vol. III. North-Holland Elsevier Science, New York (Chapter 11).

Gertler, M., Sala, L., Trigari, A., 2008. An estimated monetary DSGE model with unemployment and staggered nominal wage bargaining. J. Money Credit Bank. 40 (8), 1713–1764.

Geweke, J., 1999. Using simulation methods for bayesian econometrics models: inference, development and communication. Econ. Rev. 18 (1), 1–73.

Gilchrist, S., Zakrajsek, E., 2012. Credit spreads and business cycle fluctuations. Am. Econ. Rev. 102 (4), 1692–1720.

Gilchrist, S., Ortiz, A., Zakrasej, E., 2009. Credit risk and the macroeconomy: evidence from an estimated DSGE model. Manuscript.

Gilchrist, S., Sim, J.W., Schoenle, R., Zakrajsek, E., 2015. Inflation dynamics during the financial crisis. Finance and Economics Discussion Series 2015-012, Board of Governors of the Federal Reserve System.

Guerrieri, L., Iacoviello, M., 2013. Collateral constraints and macroeconomic asymmetries. International Finance Discussion Papers 1082, Board of Governors of the Federal Reserve System.

Gust, C., López-Salido, D., Smith, M.E., 2012. The empirical implications of the interest-rate lower bound. Finance and Economics Discussion Series 2012-83, Board of Governors of the Federal Reserve System.

He, Z., Krishnamurthy, A., 2012. A model of capital and crises. Rev. Econ. Stud. 79 (2), 735–777.

Hebden, J.S., Lindé, J., Svensson, L.E., 2010. Optimal monetary policy in the hybrid new-Keynesian model under the zero lower bound. Federal Reserve Board, manuscript.

Howard, G., Martin, R., Wilson, B.A., 2011. Are recoveries from banking and financial crises really so different? International Finance Discussion Papers No. 1037, Board of Governors of the Federal Reserve System.

Iacoviello, M., 2005. House prices, borrowing constraints, and monetary policy in the business cycle. Am. Econ. Rev. 95 (3), 739–764.

Iacoviello, M., Guerrieri, L., 2015. OccBin: a toolkit for solving dynamic models with occasionally binding constraints easily. J. Monet. Econ. 70, 22–38.

Iversen, J., Laséen, S., Lundvall, H., Söderström, U., 2016. Real-time forecasting for monetary policy analysis: the case of Sveriges riksbank. Sveriges Riksbank, manuscript.

Jordà, O., Moritz, H.P.S., Alan, M.T., 2012. When credit bites back: leverage, business cycles, and crises. Federal Reserve Bank of San Francisco Working Paper 2011-27.

Justiniano, A., Preston, B., 2010. Can structural small open-economy models account for the influence of foreign disturbances? J. Int. Econ. 81 (1), 61–74.

Justiniano, A., Primiceri, G.E., 2008. The time varying volatility of macroeconomic fluctuations. Am. Econ. Rev. 98 (3), 604–641.

Justiniano, A., Primiceri, G.E., Tambalotti, A., 2013a. Investment shocks and the relative price of investment. Rev. Econ. Dyn. 14 (1), 101–121.

Justiniano, A., Primiceri, G.E., Tambalotti, A., 2013b. Is there a trade-off between inflation and output stabilization. Am. Econ. J. Macroecon. 5 (2), 1–31.

Kimball, M.S., 1995. The quantitative analytics of the basic neomonetarist model. J. Money Credit Bank. 27 (4), 1241–1277.

Klein, P., 2000. Using the generalized schur form to solve a multivariate linear rational expectations model. J. Econ. Dyn. Control. 24, 1405–1423.

Kocherlakota, N., 2009. Modern macroeconomic models as tools for economic policy. 2009 Annual Report Essay, Federal Reserve Bank of Minneapolis.

Kydland, F., Prescott, E., 1982. Time to build and aggregate fluctuations. Econometrica 50, 1345–1371.

Laséen, S., Svensson, L.E., 2011. Anticipated alternative instrument-rate paths in policy simulations. Int. J. Cent. Bank. 7 (3), 1–36.

Leeper, E.M., Leith, C., 2016. Understanding inflation as a joint monetary—fiscal phenomenon. In: Taylor, J., Uhlig, H. (Eds.), Handbook of Macroeconomics, vol. 2B. Elsevier, Amsterdam, Netherlands, pp. 2305–2415.

Leeper, E.M., Traum, N., Walker, T.B., 2015. Clearing up the fiscal multiplier morass. NBER Working Paper No. 21433.

Levin, A.T., Natalucci, F.M., Zakrajsek, E., 2004. The magnitude and cyclical behavior of financial market frictions. Finance and Economics Discussion Series 2004-70, Board of Governors of the Federal Reserve System.

Lindé, J., 2005. Estimating new Keynesian Phillips curves: a full information maximum likelihood approach. J. Monet. Econ. 52 (6), 1135–1149.

Lindé, J., Maih, J., Wouters, R., 2016. Alternative approaches to incorporate the ZLB in the estimation of DSGE models. National Bank of Belgium, manuscript.

Liu, Z., Wang, P., Zha, T., 2013. Land-price dynamics and macroeconomic fluctuations. Econometrica 81 (3), 1147–1184.

Lucas, R.E., 1976. Econometric policy evaluation: a critique. Carn. Roch. Conf. Ser. Public Policy 1, 19–46.

Maih, J., 2010. Conditional forecasts in DSGE models. Norges Bank Working Paper No. 2010/7.

Maih, J., 2015. Efficient perturbation methods for solving regime-switching DSGE models. Norges Bank Working Paper No. 2015/1.

Mendoza, E.G., 2010. Sudden stops, financial crises, and leverage. Am. Econ. Rev. 100, 1941–1966.

Queijo von Heideken, V., 2009. How important are financial frictions in the United States and the Euro area? Scand. J. Econ. 111 (3), 567–596.

Queralto, A., 2013. A model of slow recoveries from financial crises. International Finance Discussion Papers No. 1097, Board of Governors of the Federal Reserve System.

Reifschneider, D., Williams, J.C., 2000. Three lessons for monetary policy in a low inflation era. J. Money Credit Bank. 32 (4), 936–966.

Reinhart, C.M., Rogoff, K.S., 2009. The aftermath of financial crises. Am. Econ. Rev. 99 (2), 466–472.

Rudebusch, G.D., Svensson, L.E., 1999. Policy rules for inflation targeting. In: Taylor, J.B. (Ed.), Monetary Policy Rules. University of Chicago Press, Chicago, pp. 203–246.

Schorfheide, F., 2000. Loss function-based evaluation of DSGE models. J. Appl. Econ. 15 (6), 645–670.

Sims, C.A., 1980. Macroeconomics and reality. Econometrica 48 (1), 1–48.

Sims, C.A., 2003. Implications of rational inattention. J. Monet. Econ. 50 (3), 665–690.

Sims, C.A., 2010. Rational inattention and monetary economics. In: Friedman, B.M., Woodford, M. (Eds.), Handbook of Monetary Economics, vol. 3A. Elsevier, Amsterdam, pp. 155–181.

Smets, F., Wouters, R., 2003. An estimated stochastic dynamic general equilibrium model of the Euro area. J. Eur. Econ. Assoc. 1 (5), 1123–1175.

Smets, F., Wouters, R., 2007. Shocks and frictions in US business cycles: a bayesian DSGE approach. Am. Econ. Rev. 97 (3), 586–606.

Suh, H., Walker, T.B., 2016. Taking financial frictions to the data. J. Econ. Dyn. Control. 64, 39–65.

Taylor, J.B., 2007. Housing and monetary policy, in housing, housing finance, and monetary policy. In: Proceedings–Economic Policy Symposium–Jackson Hole, Federal Reserve Bank of Kansas City, vols. 463-476.

Taylor, J.B., Wieland, V., 2012. Surprising comparative properties of monetary models: results from a new monetary model base. Rev. Econ. Stat. 94 (3), 800–816.

Vavra, J., 2013. Time-varying phillips curves. University of Chicago, manuscript.

Vayanos, D., Vila, J.L., 2009. A preferred-habitat model of the term structure of interest rates. London School of Economics, manuscript.

Villani, M., 2009. Steady state priors for vector autoregressions. J. Appl. Econ. 24, 630–650.

Warne, A., Coenen, G., Christoffel, K., 2015. Marginalized predictive likelihood comparisons of linear Gaussian state-space models with applications to DSGE, DSGE-VAR, and VAR models. J. Appl. Econ. (forthcoming). http://dx.doi.org/10.1002/jae.2514.

Wieland, V., Wolters, M., 2013. Forecasting and policy making. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting, vol. 2, Elseiver, Amsterdam, pp. 239–325 (Chapter 5).

Wieland, V., Cwik, T., Müller, G.J., Schmidt, S., Wolters, M., 2012. A new comparative approach to macroeconomic modeling and policy analysis. J. Econ. Behav. Organ. 83, 523–541.

Williams, J.C., 2014. Monetary policy at the zero lower bound: putting theory into practice. Hutchins Center Working Paper No. 2, Brookings Institution, Washington, DC.

Woodford, M., 2003. Interest Rates and Prices. Princeton University Press, Princeton.

Woodford, M., 2014. Stochastic choice: an optimizing neuroeconomic model. Am. Econ. Rev. 104 (5), 495–500.

Yun, T., 1996. Nominal price rigidity, money supply endogeneity, and business cycles. J. Monet. Econ. 37, 345–370.

CHAPTER 29

Liquidity Requirements, Liquidity Choice, and Financial Stability

D.W. Diamond[*,†], **A.K. Kashyap**[*,†]
[*]University of Chicago Booth School of Business, Chicago, IL, United States
[†]National Bureau of Economic Research, Cambridge, MA, United States

Contents

Abstract

We study a modification of the Diamond and Dybvig (1983) model in which the bank may hold a liquid asset, some depositors see sunspots that could lead them to run, and all depositors have incomplete information about the bank's ability to survive a run. The incomplete information means that the bank is not automatically incentivized to always hold enough liquid assets to survive runs. Regulation similar to the liquidity coverage ratio and the net stable funding ratio (that are soon be implemented) can change the bank's incentives so that runs are less likely. Optimal regulation would not mimic these rules.

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.011

Keywords

Bank runs, Bank regulation, Liquidity regulation, Net stable funding ratio, Liquidity coverage ratio

1. INTRODUCTION

In September 2009, the leaders of 20 major economies created the Financial Stability Board (FSB) whose purpose is to "coordinate at the international level the work of national financial authorities and international standard setting bodies (SSBs) in order to develop and promote the implementation of effective regulatory, supervisory, and other financial sector policies." Since that time the financial system has undergone a regulatory overhaul.

The term "macroprudential" regulation has become synonymous with much of this effort. As we explain in the next section, what that means in practice remains somewhat elusive. But, there are two tangible changes that are on track to occur over the remainder of this decade. One widely studied set of reforms pertain to the rules regarding capital requirements for banks. Less well-understood is that, through their cooperation via the Basel Committee on Bank Supervision, the major economies have also agreed also to implement by 2019 new rules governing banks' debt structures and requirements to hold certain types of liquid assets.

To date there is a remarkable asymmetry in the economic analysis of the capital and liquidity regulations. The pioneering work of Modigliani and Miller (1958) provides a solid theoretical framework for analyzing capital regulation. Any student taking a first course in corporate finance will encounter this theory, and there is a massive empirical literature that explores the theory's predictions. International regulations governing bank capital were introduced in 1988 and there many empirical examinations of the impact of these regulations. A recent book directed toward the general public, Admati and Hellwig (2013), makes a case for substantially increased capital requirements for commercial banks.

The discussion about regulating liquidity is much less advanced. For example, there is no benchmark theory regarding liquidity provision by intermediaries. Indeed, financial economists even have competing concepts that they have in mind when discussing liquidity, so that there is no generally accepted empirical measure of liquidity economists study. Allen (2014), in his survey of the nascent literature on liquidity regulation, concludes by writing "much more research is required in this area. With capital regulation there is a huge literature but little agreement on the optimal level of requirements. With liquidity regulation, we do not even know what to argue about."

Nonetheless, the global regulatory community has agreed on certain liquidity requirements (Basel Committee on Bank Supervision, 2013a, 2014). Two new concepts, the liquidity coverage ratio (LCR) and the net stable funding ratio (NSFR), have been proposed and banks by 2019 will be compelled to meet requirements for these ratios. Thus, it seems fair to say we are in a situation where practice is ahead of both theory and measurement.

In this chapter, we survey the existing work on liquidity regulation and develop a framework for discussing the regulation. The theory that we propose suggests, in certain parameterizations, regulations bearing some resemblance to the LCR and NSFR can emerge as ones which will improve outcomes relative to an unregulated benchmark. However, the regulations that arise in our model would naturally differ across banks, depending on certain bank characteristics, so they do not mimic exactly the ones that are on track to be implemented.

The critical ingredients in our model are the following. First, we consider banks which are spatially separated and hence do not compete aggressively for deposits. Treating the bank as monopolist simplifies the analysis by allowing us to side-step some complications that arise from having to model the deposit market equilibrium. The model can also be interpreted as a description of the aggregate banking system, which for many financial stability and regulatory discussions, is the object of primary concern, and under this interpretation ignoring the deposit competition is perhaps more natural.

Second, we assume that intermediaries provide liquidity insurance for customers who have uncertain withdrawal needs (or consumption desires). We build on the Diamond and Dybvig (1983), henceforth DD, model of banking in which banks provide this insurance by relying on the law of large numbers to eliminate idiosyncratic customer liquidity needs.

For those familiar with DD, we make two modifications. The first is allowing the bank to invest in a liquid asset that has a rate of return exceeding the return from liquidating illiquid assets and thus is the efficient way to arrange to pay customers that need liquidity. This introduces a trade-off between lending and holding liquidity as in Bhattacharya and Gale (1987), several papers of Allen and Gale (1997), and others.

The other modification to DD is the form of run risk that the banks face. Banks are assumed to have a good assessment of the aggregate needs of their customers for fundamental reasons. But, they also know that some customers will receive a signal about the bank which could lead to a run. The sunspots that we consider are a metaphor for people being concerned with the health of the bank, but not having a fully formed set of beliefs about the bank's solvency status. In making their decisions, we assume that customers are unable to fully evaluate the ability of the bank to honor deposits. Given the complexity of modern banks it seems realistic to presume that most customers cannot precisely determine their bank's maturity mismatch and hence its vulnerability to a run. The imperfect information creates a challenge for the banks because their customers will not necessarily know if the bank is prudently holding liquidity or not, which reduces the incentive to hold liquidity.

In the event that a run does occur, we depart from DD and Ennis and Keister (2006) to allow for the possibility that not all customers seek to withdraw their funds. We believe it is useful to analyze partial runs for two separate reasons. One is that in practice there do seem to be some sticky deposits that do not flee even in times of considerable banking stress. In addition, even before troubles occur it is usually clear which types of deposits are prone to running. So this allows us to talk about policies for different types of withdrawal risk.

Within this environment we can assess the vulnerability of the financial system to runs under different regulatory arrangements. In the baseline case, we assume that banks simply maximize their profits and see which types of equilibria arise. As usual in DD style models, the outcomes depend critically on how depositors form beliefs. It is possible, under certain parameter configurations, that the pure self-interest motives of the banks will sufficient to insure that the system will be run proof even if depositors had no detailed information about a bank's liquidity holdings. In these situations, added liquidity could not influence whether a given depositor would choose to join a run if one was feared.

We describe several reasons why depositors may not be able to use some types of disclosure of a bank's liquidity holdings to determine if the holding is sufficient to allow it to survive a run. To fix ideas, one can consider whether a bank would choose to hold this sufficient amount of liquidity even if its choice between liquid assets and illiquid loans was completely unobservable. In circumstances where depositors cannot be sure about how changes in liquidity holdings impact the robustness of banks to runs, the banks will typically face a tension in deciding how much to fortify themselves against the risk of a run. They can always choose to be sufficiently conservative to be able to withstand a worst case of fundamental withdrawals as well as a panic. But in order to do that, they will engage in very little lending, and the forgone profits from deterring the run will be high. The additional liquidity to survive a run will turn out to be excessive whenever a run is avoided. Hence, it is possible they will make more profits from added lending which would leave them unable to always be able to sustain a run.

We next allow regulatory interventions that place restrictions on present and possibly on future bank portfolio choices. In the baseline setup, the banks have perfectly aligned incentives to prepare to service fundamental aggregate withdrawal needs. So the regulatory challenge is to determine whether a requirement that distorts their private incentives toward being more robust to a run will improve outcomes. We allow for regulation that is inspired by the two impending Basel rules.

One variant requires an initial liquidity position that must be established before depositors make their intentions clear. This can function like the "NSFR" that is proposed as part of the Basel reforms. A second option is a mandate to always hold additional liquid assets beyond those needed for the fundamental withdrawals. This imposes both present and future minimum holdings of liquid assets. This regulation looks like a traditional reserve requirement for the bank but can also be interpreted as a kind of "liquidity coverage" ratio that is part of the Basel reforms.

One point of contention regarding the LCR that has emerged is whether required liquidity can be deployed in the case of a crisis. Goodhart (2008) framed the issue nicely with a now famous analogy of "the weary traveller who arrives at the railway station late at night, and, to his delight, sees a taxi there who could take him to his distant destination. He hails the taxi, but the taxi driver replies that he cannot take him, since local bylaws require that there must always be one taxi standing ready at the station."

One way to interpret the Goodhart conundrum is to recognize that, broadly speaking, there are two ways to think about the purpose behind liquidity regulations. One motivation can be to make sure that banks can better withstand a surge in withdrawals should one occur. From this perspective, mandating that the last cab cannot depart the station seems foolish. Another possible motivation is to design regulations aimed at reducing the likelihood of a withdrawal surge in the first place. Our model helps highlight the potential incentive properties of regulation and can potentially explain why mandating the presence of some unused liquidity could be beneficial.

In studying how private and social incentives for liquidity choices diverge, our main conclusion from analyzing the two Basel-style regulations is that they may improve outcomes relative to the ones that arise from pure self-interest, but each brings potential inefficiencies. Hence, we briefly also describe the solution of the mechanism design problem for a social planner who has less information about withdrawal risk than the bank does and seeks to optimally regulate banks to avoid runs. That solution provides a natural benchmark against which to judge the Basel-style regulations.

The remainder of the chapter is divided into five parts. Section 2 contains our selective overview of previous work. We organize this into three subsections. We begin with an overview of the emerging policy proposals and research regarding macroprudential regulation. We then hone in on the enormous and rapidly growing literature on capital regulation. We provide our perspective on how to group these papers and highlight several recent excellent surveys on the pure effects of capital regulation. We close with a review of the most relevant papers for our questions that motivate us about liquidity regulation.

Section 3 introduces the benchmark model. We explain how it works under complete information. We also derive a generic proposition that holds with incomplete information that describes when the bank's preferred liquidity choice will be sufficient to deter a run. Generically, however, privately chosen levels of liquidity need not be sufficient to deter runs. So this opens the door for regulations that might do so.

In Section 4, we analyze the two types of liquidity regulation that are akin to the ones contemplated under the Basel process. We first demonstrate that a particular type of regulation that requires the bank to hold liquid assets equal to a fixed percentage of deposits at all times can potentially deter runs. This works because the liquidity mandate, combined the bank's self-interest to prepare to service predictable deposit outflows, leads the bank to hold more overall liquidity than it would otherwise. Because depositors understand this, it removes the incentive to run in some cases. We also consider alternative assumptions about depositors' knowledge and the information available to regulators and assess the vulnerability of the bank to runs in these scenarios.

In Section 5, we describe a couple of extensions of the baseline model. The first sketches a mechanism design problem where the regulator does not have all of the bank's information and seeks to implement run-free banking. We fully characterize the solution to this problem in Diamond and Kashyap (2016), here we describe the main findings from this exercise. It turns out that a regulator with sufficient tools can induce the bank to hold the proper amount of liquidity despite the private information advantage possessed by the bank.

We also briefly discuss capital regulation. We explain why, as a tool for managing liquidity problems, capital requirements can be relatively inefficient compared to the other regulations that we have reviewed. Obviously in a richer model where both credit risk and liquidity risk are present, capital, and liquidity regulations can serve different purposes. We describe some of these differences.

Section 6 presents our conclusions. Besides summarizing our findings, we also pose a few open questions that are natural next steps to consider in addressing the issues analyzed in this chapter.

2. LITERATURE REVIEW

Research on financial regulation has exploded since the global financial crisis (GFC), and the number of regulatory interventions and tools has also expanded massively. To review all of this work would require a book. To keep our review manageable, we limit our discussion to focusing on the theoretical underpinnings and rationale behind these changes.[a]

2.1 Macroprudential Regulation

Clement (2010) provides the interesting history of the origins and evolution in the meaning of the phrase "macroprudential." His best estimate is that the term appeared first in 1979 in the documents of the committee that was the fore-runner to the Basel Committee on Bank Supervision. The first public document using the term which he can identify was a report by the committee now known as the Committee on the Global Financial System. It defined macroprudential policy as promoting "the safety and soundness of the broad financial system and payments mechanism."

The phrase took on added prominence when it was the focus of a Sept. 2000 speech by Andrew Crockett (who was then the General Manager of the Bank for International Settlements (Crockett, 2000)). He defined the objective of macroprudential policy to be "limiting the costs to the economy from financial distress, including those that arise from any moral hazard induced by the policies pursued." Crockett's rational for calling for macroprudential policies was his belief that optimal choices for a single institution could create problems for the financial system as a whole. He was explicitly focused on the distinction between the supervisory challenges for monitoring an individual institution and those for protecting the aggregate financial system.

[a] For a diverse set of perspectives on the changing postcrisis regulatory landscape see Čihák et al. (2013), Financial Stability Board (2015), Claessens and Kodres (2014), Basel Committee on Bank Supervision (2013a,b), and Fisher (2015).

Crockett did not offer precise microeconomic foundations for why the private actions of individual actors would not be aligned with social welfare, but he did give a few examples where he saw the potential for divergence. One possibility he cited is that one bank seeking to limit its credit exposures could choose to cut lending to its clients, but if all banks did this a credit crunch could ensue that would trigger a recession. A second example was the possibility of what we would now dub to be a fire-sale where all agents simultaneously cut back on asset exposures due to falling prices and in the course of doing so exacerbate the price decline. A third problem arises if many lenders shorten the maturity of their funding to a particular borrower, then the risk of a run can increase so that they are all more vulnerable.

Our view is that Crockett's spotlight on the divergence between the narrow private interests of individual institutions (or supervisors monitoring a single institution) and the interests of overall society is exactly the right focus for considering macroprudential policies. Indeed, this literature would be well-served to move in the direction where all macroprudential papers start by clarifying why (and when) social and private interests diverge. The challenge for both for researchers and policymakers is the difficulty in formalizing and prioritizing the exact reasons for the divergence. To clearly see the problem, compare three prominent perspectives on macroprudential regulation that have followed Crockett.

First, various BIS documents (eg, Clement, 2010) now interpret Crockett as having identified two types of problems that are to be addressed. One relates to the buildup of risks over time that are often now referred to as the procyclicality of the financial system or the "time dimension" of the macroprudential policy problem. The other relates to the distribution of risks within the financial system, the so-called cross-sectional dimension of the problem. Many official sector documents adopt the convention of separating time-series and cross-sectional macroprudential problems. As Clement (2010) notes, while the BIS work in this area has been relatively precise in the way these issues are discussed, "the usage of the term in the public sphere has on occasion been loose. It is not uncommon for it to be employed almost interchangeably with policies designed to address systemic risk or concerns that lie at the intersection between the macroeconomy and financial stability, regardless of the specific tools used."

In contrast, Hanson et al. (2011) start with a particular view of "how modern financial crises unfold, and why both an unregulated financial system, as well as one based on capital rules that only apply to traditional banks, is likely to be fragile." Their perspective, appealing to the model in Stein (2012), presumes that banks will find it cheaper to fund themselves with short-term debt than equity, so that banks have limited incentives to build strong equity buffers in normal times. If, in a crisis, such banks suffer substantial losses, then the market value of debt claims can fall below the face value, which will deter them from raising new equity (Myers, 1977). Consequently, in this case the banks are likely to comply with capital regulations by shrinking their asset base. Hence, Hanson et al. argue that the goal of macroprudential regulation should be to "control the social costs associated with excessive balance sheet shrinkage on the part of multiple financial institutions hit with a common shock."

A recent survey by the Norges bank staff, Borchgrevink et al. (2014), argues that in fact there are six market failures that can give rise to macroprudential concerns. These are pecuniary externalities, interconnectedness externalities, strategic complementarities, aggregate demand externalities, market for lemons, and deviations from full rationality. Not surprisingly they conclude "Because of the diversity of these categories, policy lessons diverge. There is yet no 'workhorse' model for policy analysis." Though they do argue that capital and liquidity regulation should tuned to aggregate conditions, not just those of individual banks, and that borrowers should be subjected to time-varying policies that aim to force them to internalize the costs of excessive borrowing.[b]

We share the Borchgrevink et al (2014) conclusion that the macroprudential literature at this point remains in sufficient flux that it is too soon to reach firm conclusions about where it will lead. Hence, for the remainder of our analysis we focus on capital and liquidity regulation where the range of issues to be considered can be narrowed and where specific global policies are being implemented.

2.2 Capital Regulation

For an overview of the literature on capital regulation, it is useful to sort papers along two dimensions. The first regards what is assumed regarding the Modigliani–Miller (1958) (henceforth MM) capital structure propositions. As in all models of corporate finance, absent failures of one of the MM propositions any choices regarding capital structure will be inconsequential. There have been four primary MM violations that have drawn attention in the literature.

One concerns that existence of deposit insurance. If certain parts of a bank's capital structure is protected from losses by the government, that can create risk-shifting incentives for equity holders. In many models, bank managers working on behalf of the equity owners face an incentive to gamble after adverse shocks that goes unchecked because depositors are immune from losses that they would suffer if the gamble fails.

A second distortion is concerns over guarantees to protect equity holders of banks from losses. Usually this is couched as a problem of having some banks that are assumed to be "too big" or "too-interconnected" to fail. But, in the recent GFC, there were also

[b] Others have also chosen to organize their analyses around distinctions between the kinds of tools that can be deployed. For example, Aikman et al. (2013) classify tools into three groups: those that operate on financial institutions' balance sheets; those that affect the terms and conditions on financial transactions; and those that influence market structures. While Cerutti et al. (2015) present empirical analyses comparing 12 types of different regulations.

cases in some countries where equity owners of smaller, nonsystemic banks were insulated from losses due to political connections.

A third violation regards the MM assumption of complete financial markets. With incomplete markets, an institution that creates new securities could be valuable. In the banking context, deposits are a leading example of special security that banks might create.

Finally, there are many models where either asymmetric information or moral hazard problems are considered. Some of the prominent examples include the possibility that borrowers know more about their investment opportunities than lenders, or that borrowers can shift the riskiness of their investments after receiving funding.

So unlike much of the research on nonfinancial corporations, the trade-off theory of capital structure, whereby firms prefer debt for its tax advantages and balance those benefits against costs of financial distress, has not figured prominently in the banking research on capital regulation. Rather, regulation is usually justified on the grounds of addressing one of these other four problems. The type of regulation that can be welfare improving will differ depending on which of these other frictions is assumed to be present.

The second important dimension one which the literature can be organized concerns the economic services that banks are assumed to provide.[c] Broadly, there are three types of services that have been modeled. The first presumes that certain financial institutions can expand the amount of credit that borrowers can obtain (say, relative to direct lending by individual savers). The micro-founded theories typically assume that borrowers can potentially default on loans and so any lender has to be diligent in monitoring borrowers (Diamond, 1984). By concentrating the lending with specialized agents, these monitoring costs can be conserved and the amount of credit extended can be expanded.

A second widely posited role for intermediaries is helping people and businesses share risks (Allen and Gale, 1997; Benston and Smith, 1976). There are many ways to formalize how this takes place, but perhaps the simplest is to recognize that because banks offer both deposits and equity to savers, they can create two different types of claims that would be backed by bank assets. These two choices allow savers to hedge some risks associated with lending, and this hedging improves the consumption opportunities for savers. More broadly, these theories suppose that banks help pool and tranche risks.[d]

A third class of models, which complements the second, supposes that the financial system creates liquid claims that facilitate transactions. There are various motivations behind how this can be modeled. In DD style models, an intermediary can cross-insure consumers' needs for liquidity by exploiting the law of large numbers among customers. But doing so exposes banks to the possibility of a run, which can be disastrous for the bank and its borrowers and depositors. Calomiris and Kahn (1991) and Diamond and Rajan (2001)

[c] The next few paragraphs are taken from Kashyap et al. (2014).

[d] For instance, if there are transactions costs associated with buying securities, a bank that makes no loans but holds traded securities could still be valuable.

explain that the very destructive nature of a run is perhaps helpful in disciplining the bank to work hard to honor its claims. So the fragility of runs is potentially important in allowing both high amounts of lending and large amounts of liquidity creation.

Gorton and Winton (2003) give a much more complete review of these three classes of theories and one clear conclusion that emerges is that depending on which of these three services is presumed to be operative, and which of the MM failures are present, one can reach very different conclusions about the efficacy of capital regulation in improving welfare. For instance, in models where liquidity creation is not one of the services provided by banks, the costs of mandating higher amounts of equity financing are often modest. Likewise, the benefits of protecting taxpayers from having to bail out banks or depositors by forcing more equity issuance are potentially substantial.

Rather than reviewing the results from many papers on capital regulation we refer interested readers to several recent surveys including Brooke et al (2015), Martynova (2015), Rochet (2014), and the references therein. Both Brooke et al. and Rochet attempt to compare the macroeconomic costs and benefits of higher levels of required capital and use a variety of calculations to assess them. In both cases, the benefits are presumed to be a reduction in likelihood and potential severity of financial crises (and the associated reductions in output). While the costs of higher capital requirements are the possible potential reductions in lending and losses of output. One humbling observation from both of these papers is that despite drawing on many different types of evidence, empirically estimating the net effects is difficult and there is substantial uncertainty about the overall net effects.

One other important observation is that most of the papers in these reviews are not very informative regarding liquidity regulation or the potential interactions of liquidity and capital regulation because in the environment being analyzed there is no value to liquidity creation (and hence no cost to limiting it). Indeed, Bouwman (2015), in a review article, emphasizes the dearth of research on potential interactions between capital and liquidity regulation and argues that it "is critically important to develop a good understanding of how capital and liquidity requirements interact."

2.3 Liquidity Regulation

As mentioned in Section 1, there are far fewer papers that seek to investigate the purpose and effect of liquidity regulation. Allen (2014) offers a survey of this nascent literature and we share the sentiment of the concluding paragraph of his survey. He writes, "much more research is required in this area. With capital regulation there is a huge literature but little agreement on the optimal level of requirements. With liquidity regulation, we do not even know what to argue about."

It is possible to again use a similar kind of two-way to classification regarding capital regulation to describe much of the thinking on liquidity. Trivially, if the economic services offered by a bank do not include the provision of liquidity, then regulation that focuses on liquidity will not be particularly interesting to consider. It is possible that in such environments regulating liquidity could make sense to achieve other aims, such as supplementing or substituting for capital requirements. However, if maturity transformation is not one of the outputs of the financial system, assessments of the efficacy of liquidity regulation in such models will be incomplete. Put bluntly, if there are no costs to limiting liquidity provision per se, then obviously the cost of regulations that have this effect cannot be fully assessed.

It is worth noting that will most of the literature on liquidity and liquidity regulation label the institutions that undertake this activity as "banks." However, as became evident in the GFC this activity is hardly limited to banks. Fig. 1, reproduced from

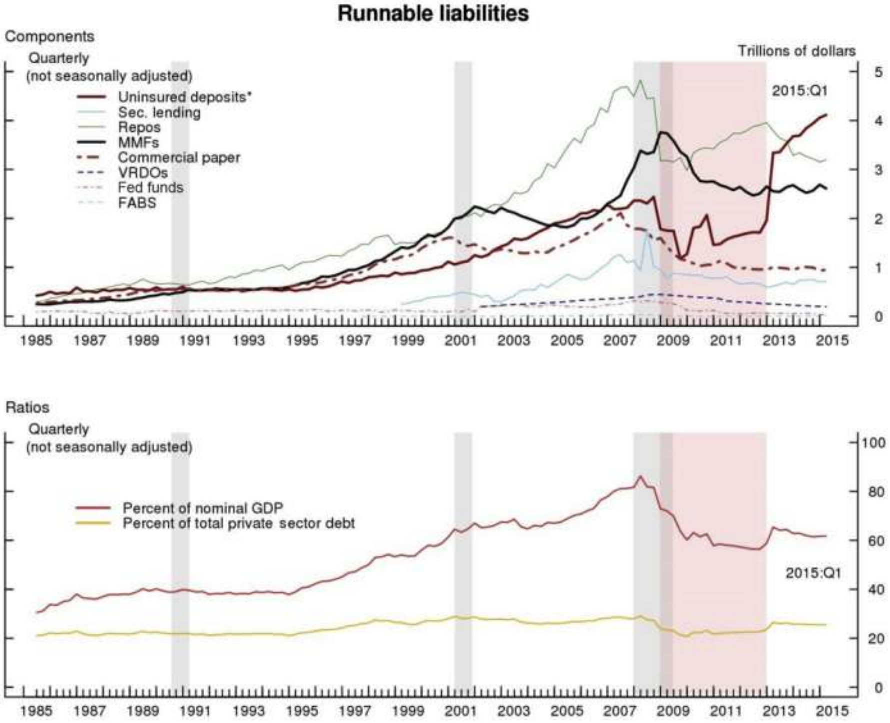


Fig. 1 Bao et al. (2015) estimates of runnable funding in the United States. *Uninsured deposits equal the difference between total deposits and insured deposits. The quarterly insured deposits series between 1985 and 1990 are obtained by interpolating the available annual data. For 2008:Q4–2012:Q4 (*red* (*light gray* in print version)) *shades*, insured deposits increased due to the Transaction Account Guarantee (TAG) program. For 2008:Q4–2009:Q2, some insured deposits were not accounted for because the FDIC did not collect data on insured amounts for those TAG accounts with balances between $100,000 and $250,000. *Note*: The *gray shades*, which overlap the *red* (*light gray* in print version) *shades*, indicate NBER recession dates. Source: *Staff calculations using data from RMA, DTCC, SIFMA, Call Reports, Financial Accounts, M3 monetary aggregates, and Bloomberg Finance LP.*

Bao et al. (2015), shows the total amount of runnable funding inside the US financial system over the past 30 years.

We draw three conclusions from their estimates that are worth bearing in mind throughout the rest of the discussion. First, there has been a sizable increase in the amount maturity transformation over the last 20 years. From 1995 until 2015, the scale of such activity rose by 50% as measured relative to gross domestic product (GDP). Second, as far back as 1985 as much of this activity has occurred outside the banking system as inside it. Third, the decline immediately after the GFC was sizable. The drop in repurchase agreements and money market funds were especially pronounced, but even as a percent of GDP, the level in 2015 is very similar to the level in 2005 (just before the frenzied period ahead of the GFC). Hence, maturity transformation is still happening on a substantial scale even after the GFC and all of the various regulatory reforms that have been introduced.

Given this evidence, we focus only on papers where one of the services of the financial system is to provide liquidity. Among these it is helpful to separate them into papers that model liquidity provision in the same way or similarly to DD, and those that introduce other mechanisms.

Among the DD style models, we focus on three that are closely related to our analysis. Ennis and Keister (2006) have a DD style model (related to Cooper and Ross, 1998) which determines how much liquidity banks need to hold to deter runs. They compute the amount of excess liquidity the bank must hold to buffer it against a run by all depositors, and also determine the optimal amounts to promise depositors. In their model with full information, when depositors desire safe banks, there will be private incentives to hold enough excess liquidity to deter a sunspot-based run. They do not study regulation because there is no need for any under their assumptions, but we will see that some of the same forces that are present in their model arise in ours.

Vives (2014) analyzes a question similar to that in Ennis and Keister (2006): what are the efficient combinations of equity capital and liquidity holdings to make a bank safe when it subject to runs based on private information about its solvency? He studies a global game where a bank can be insolvent or illiquid. The need for regulation is not considered explicitly, but he does examine what capital and liquidity levels would make the bank safer. He finds that capital and liquidity are differentially successful in attending to insolvency and illiquidity. In particular, if depositors are very conservative (and which makes them more inclined to run in the model), increased liquidity holdings which reduce profits by investing more in liquid assets can enhance stability.

Farhi et al. (2009) investigate a DD model where consumers need banks to invest and where the consumers can trade bank deposits. Absent a minimum liquidity regulation, it is profitable to free ride on the liquidity held by other banks, because banks offer rates which subsidize those who need to withdraw their deposit early (which is the spirit of Jacklin, 1987). A floor on liquidity holdings removes the incentive for this free riding.

Among the non-DD models, one that is related is Calomiris et al. (2014). They have a six period model where banks can potentially engage in risk-shifting so that when banks suffer loan losses they may not be able to honor their deposit contracts. Cash is observable and mandating that banks must have minimum levels of cash reserves can limit the risk-shifting.

Santos and Suarez (2015) examine another role for liquidity when runs occur slowly; it allows time to decide if the bank's assets are sufficient to imply solvency absent a run. This channel is foreclosed in our setup with assets which are free of risk.

More generally, our approach is closely related to the mechanism design approach to regulation of monopolists in Baron and Myerson (1982). They also were interested in investigating how regulation could be structured to induce the party being regulated to efficiently use information that is private.

3. BASELINE MODEL

We begin by describing a baseline setup in which the timing and preferences are as in DD. We then modify certain informational assumptions to bound the possible outcomes. Throughout we maintain that there are three dates: $T=0$, 1, and 2. The interest rates that bank must offer are taken as given, motivated by a monopoly bank which must meet the outside option of depositors to attract deposits. Equivalently, the single bank can be thought of as representing the overall banking system.

For a unit investment at date 0, the bank offers a demand deposit which pays either r_1 at date 1 or r_2 at date 2. This effectively offers a gross rate of return r_2/r_1 between dates 1 and 2 which is equal to the exogenous outside option (such as government bonds) for depositors between these dates. Essentially, the bank offers one period deposits which equal the interest rate on the outside option. We will assume that depositors are sufficiently risk averse that they would like the banking system to supply one period deposits that are riskless. Hence, when we consider interventions they will be designed to deliver as this as the only possible equilibrium.

The residual claim after deposits are paid is limited liability equity retained by the banker. All equity payments are made at date 2.[e]

The bank can invest in two assets with constant returns to scale. One is a liquid asset (which we will interchangeably refer to as the safe asset) that returns $R_1>0$ per unit invested in the previous period. The other is an illiquid asset for which a unit investment at date 0 returns at date 2 an amount that exceeds the return from rolling over liquid assets ($R_2>R_1{}^*R_1$). The illiquid asset (which we will interchangeably refer to a loan)

[e] We could introduce another incentive problem for the banker to motive a minimum value of equity at all dates and states, but for now the bank will operate efficiently as long as equity remains positive in equilibrium.

can be liquidated for θR_2 date 1, where $\theta R_2 < R_1$ and $\theta \geq 0$. These restrictions imply that when the bank knows it must make a payment at date 1, it is always more efficient to do that by investing in the safe asset rather than planning to liquidate the loan.

We also assume that banking is profitable even if the bank invests exclusively in the liquid asset, so that $r_1 \leq R_1$ and $r_2 \leq R_1^2$. This is a sufficient condition to guarantee that requiring excess liquidity will not make the bank insolvent (though it still will reduce the efficiency of investment). In addition, we assume that bank profits from investing in illiquid assets when depositors hold their deposits for two periods (borrowing short-term repeatedly to fund long-term illiquid investment) is greater than from investing in liquid assets when depositors hold their deposits for only one period (or $\frac{r_2}{R_2} < \frac{r_1}{R_1}$). This implies that a bank is most profitable when in can finance loans returning R_2 with deposits for two periods at cost r_2 (as compared with financing liquid assets for one period). This second assumption is used only to obtain some results on optimal liquidity holdings.

There are many possible reasons to presume that the illiquid asset can be liquidated for only θR_2. For instance, in DD liquidation can be thought of as a nontradable production technology. Alternatively it could reflect the bank's lending skills, implying that it would be worth less to a buyer than to the bank because (compared to the bank) the buyer would be able to collect less from a borrower, as in Diamond and Rajan (2001). Nothing in our analysis hinges on why this discount exists, though we do insist that it is operative for everyone in the economy including a potential lender of last resort (LOLR). Also, our assumption that θ is a constant implies that we are not modeling a situation where the sale price depends only on the amount of remaining liquidity held by potential buyers (as in Bhattacharya and Gale, 1987; Allen and Gale, 1997; and Diamond, 1997).

For fundamental reasons, a fraction t_s of depositors want to withdraw at date 1 and $1 - t_s$ want to withdraw at date 2 in state s. The realizations of t_s are bounded below by $\underline{t} \geq 0$ and above by $\bar{t} \leq 1$. The banker will know the realization of t_s when the asset composition choice is made. This assumption is meant to capture the fact that banks have superior information about their customers. Indeed, some early theories of banking supposed that the advantage of tying lending and deposit making was that by watching a customer's checking account activities a bank could gauge that customer's creditworthiness (Black, 1975).

Mester et al. (2007) provide direct evidence supporting the assumption that banks can learn about customer credit needs by monitoring transactions accounts. Drawing on a unique data set from a Canadian bank, they demonstrate the bank is able to infer changes in the value of borrowers' collateral that is posted against commercial loans by tracking flows into and out of the borrowers' transaction accounts. At this bank, they document that the number of prior borrowings in excess of collateral is an important predictor of credit downgrades and loan write-downs. Most importantly, the bank uses this information in making credit decisions. Loan reviews become longer and more frequent for borrowers with deteriorating collateral.[f] In what follows, we make the simplifying assumption t_s is always known exactly by the bank, but the analysis also goes through so long as the bank is simply better informed than the depositors and the regulator.

To understand agents' incentives, note that if the ex-post state is s and there is not a run, a fraction $f_1 = t_s$ will withdraw r_1 each, requiring $r_1 t_s$ in date 1 resources, and this will leave a fraction $1 - t_s$ depositors at date 2 who are collectively owed $r_2(1 - t_s)$ (in date 2 resources). If we let α_s be the fraction of the bank's portfolio that is invested in the liquid asset and $(1 - \alpha_s)$ be the portion invested in the illiquid one, then the bank's profits, and hence its value of equity in general will be

$$\text{Value of equity} = \left\{ \begin{array}{ll} (1-\alpha_s)R_2 + (\alpha_s R_1 - f_1 r_1)R_1 - (1-f_1)r_2 & \text{if } f_1 r_1 \leq \alpha_s R_1 \\ \text{Max}\left\{0, \left(1 - \alpha_s - \dfrac{(f_1 r_1 - \alpha_s R_1)}{\theta R_2}\right) R_2 - (1-f_1)r_2\right\} & \text{if } f_1 r > \alpha_s R \end{array} \right\} \tag{1}$$

Because we are assuming that the bank knows t_s, its own self-interest will lead it to make sure to always have enough invested in the liquid asset to cover these withdrawals. So absent a run, the profits are very intuitive and easy to understand. The first term in Eq. (1) when $f_1 r_1 \leq \alpha_s R_1$ represents the returns from the illiquid investment, the second reflects the spread on the safe asset relative to deposits (recognizing that any leftover funds are rolled over), and the third term reflects the funding costs of the remaining two period deposits. When $f_1 r_1 > \alpha_s R_1$, the bank needs to pay out more than its liquid assets are worth at date 1. To honor its promises, the bank must liquidate illiquid assets worth θR_2 each, implying that each unit of withdrawn in excess of $\alpha_s R_1$ removes $(1/\theta R_2)$ loans from the bank's balance sheet. These loans would each be worth R_2 at date 2. For a bank in this situation that can honor all early and late withdrawals the residual profits go to the banker (otherwise the bank is insolvent). Given our assumptions about interest rates and liquidation discounts, if actual withdraws, f_1, were known, the bank would choose to hold enough liquid assets to avoid needing to liquidate any loans. We know that at all times, even absent a run in state s, $f_1 \geq t_s$. As a result, the bank will always have an incentive to choose $\alpha_s \geq \dfrac{t_s r_1}{R_1} \equiv \alpha_s^{\text{AIC}}$. As a result, we refer to α_s^{AIC} as the *automatically incentive compatible* liquidity holding of the bank.

It is interesting to consider what happens when a run is possible. We suppose that a fixed number Δ of the patient depositors are highly likely to see a sunspot. All depositors (and the bank) know Δ and upon seeing the sunspot they must decide whether they believe that the others who see it will decide withdraw their funds early. As mentioned

[f] Norden and Weber (2010) also find that credit line usage, credit limit violations, and cash inflows into checking accounts are unusual in the periods preceding defaults by small businesses and individuals in Germany.

earlier, the sunspot is intended to stand in for general fears about the solvency of the bank, so the inference problem relates to their conjecture about whether others investors might panic. In that case, they have to decide whether to join the run.[g] So in general $f_1 > t_s$ is possible.

If the bank will be insolvent with a fraction of withdrawals of any amount less than $t_s + \Delta$, then we assume each depositor who sees then sunspot will withdraw and $f_1 = t_s + \Delta$. This will give zero to all who do not withdraw, and the goal of bank or its regulator is to prevent this outcome from ever being a Nash equilibrium. We will refer to a bank as unstable if its asset holdings admit the possibility of a run. Alternatively, we refer to a bank as stable if its asset holding eliminate the possibility of a run.

In addition, we will assume that if the bank is exactly solvent at $f_1 = t_s + \Delta$, no depositor who does not need to withdraw (and only sees the sunspot) will withdraw. This condition establishes exactly how much liquidity is needed to deter a run (as opposed to providing a floor which must be exceeded). We define the *minimum stable amount of liquidity holdings*, α_s^{Stable} as the minimum fraction of liquid assets in state s which eliminate the possibility of a run. This implies that a bank with $\alpha_s \geq \alpha_s^{\text{Stable}}$ will be run-free.

3.1 Complete Information

We presume that depositors desire run-free bank deposits. As a first benchmark, suppose that depositors know all of the choices and information which banks know, and thus observe α_s, Δ, and t_s. In this case, the need to attract deposits will force the bank to make itself run-free. If, given depositor knowledge of α_s, Δ, and t_s, the bank would remain solvent in a run, then it never is individually rational to react to the sunspot, and there will be no runs. Proposition 1 shows that it is possible that the bank will not need to distort its holding of liquidity to implement run-free banking.

Proposition 1

If the bank chooses $\alpha_s^{\text{AIC}} = \dfrac{t_s r_1}{R_1}$, and if

$$t_s + \Delta_s < \frac{t_s r_1 + \left(1 - \dfrac{t_s r_1}{R_1}\right)\theta R_2 - r_2\theta}{r_1 - r_2\theta} \left(\text{equivalently } \theta \geq \frac{\Delta r_1 R_1}{R_2(R_1 - r_1 t_s) - r_2 R_1(1 - t_s - \Delta)}\right),$$

investors will not run and the bank is stable with $\alpha_s^{\text{AIC}} = \dfrac{t_s r_1}{R_1}$.

[g] Uhlig (2010) shows that partial bank runs in a DD style model can arise if there other types of dispersion in agents' beliefs. For instance, if depositors are highly uncertainty averse and differ in their estimates of θ that heterogeneity can lead to a partial bank run in his setup.

Proof

If $f_1 r > \alpha R_1$, the bank's equity is positive when $\left(1-\alpha_s - \frac{(f_1 r_1 - \alpha_s R_1)}{\theta R_2}\right) R_2 - (1-f_1) r_2 \geq 0.$

So, when the automatically incentive compatible level of initial liquidity α^{AIC} is chosen $\left(\alpha_s^{\mathrm{AIC}} = \frac{t_s r_1}{R_1}\right)$, the value of equity is decreasing in f_1 and equal zero when $f_1^* = \frac{\frac{t_s r_1}{R_1} R_1 + \left(1 - \frac{t_s r_1}{R_1}\right) \theta R_2 - r_2 \theta}{r_1 - r_2 \theta}$. Therefore, if $t_s + \Delta$ is less than f_1^*, then the depositors always know the bank will be solvent and there is no Nash equilibrium with a run. □

The proposition simply states the condition when the bank is sufficiently profitable and liquid, so that by holding only enough of the liquid asset to service fundamental withdrawals, the bank will nonetheless be solvent in the event of a run. Under these conditions, $\alpha_s^{\mathrm{AIC}} \geq \alpha_s^{\mathrm{Stable}}$.

When the conditions for Proposition 1 fail, because loans are quite illiquid or the bank is not very profitable, then to deter runs, a bank must hold more liquidity than is needed to meet normal withdrawals. One useful case to contemplate is when loans are totally illiquid ($\theta = 0$). In this case, the bank must always hold enough liquidity to fully finance the run because there is no other way to get access to liquidity or $\alpha_s^{\mathrm{Stable}} = (t_s + \Delta)\frac{r_1}{R_1} > \alpha_s^{\mathrm{AIC}} = \frac{t_s r_1}{R_1}$. Therefore, the bank must always hold more liquidity than is needed for normal withdrawals in order to deter a run. More generally, whenever

$$t_s + \Delta > \frac{t_s r_1 + \left(1 - \frac{t_s r_1}{R_1}\right) \theta R_2 - r_2 \theta}{r_1 - r_2 \theta} \quad \text{or } \theta < \frac{\Delta r_1 R_1}{R_2 (R_1 - r_1 t_s) - r_2 R_1 (1 - t_s - \Delta)}$$

then the bank must increase α_s to $\alpha_s^{\mathrm{stable}}$ to definitely deter the run, where $\alpha_s^{\mathrm{stable}}$ is such that $t_s + \Delta = \frac{\alpha_s^{\mathrm{stable}} R_1 + \left(1 - \alpha_s^{\mathrm{stable}}\right) \theta R_2 - r_2 \theta}{r_1 - r_2 \theta}$. This yields

$$\alpha_s^{\mathrm{stable}} = \frac{(t_s + \Delta) r_1 + \theta((1 - t_s - \Delta) r_2 - R_2)}{R_1 - \theta R_2}.$$

So when it is sufficiently illiquid, merely preparing to service fundamental withdrawals will not always be enough to deter a run.

This threshold tells us how much liquidity is needed when there is full information such that all variables including t_s are known and all parties understand the bank's incentives. Under the conditions of Proposition 1, the bank will choose $\alpha_s = \frac{t_s r_1}{R_1}$ and no unused liquidity is held from dates 1 to 2. Because depositors might choose to run and the incentive for this must be removed, this will not be enough liquidity when the conditions for Proposition 1 do not hold. To always deter a possible run, the bank will have to hold $\alpha_s = \alpha_s^{\text{Stable}} > \alpha_s^{\text{AIC}}$. This will require that some unused liquidity, $\left(\alpha_s^{\text{Stable}} - \alpha_s^{\text{AIC}}\right) R_1 \equiv U(t_s) > 0$, to be held from date 1 to 2, after the normal withdraws are met at date 1. If the bank is free to use all of this unused liquidity if a run should occur, then depositors can see that the liquidity is present and will never choose to run. Once the run is deterred, the liquidity will be in excess of what is needed. This is the simplest example of the benefits of holding unused liquidity or leaving extra taxicabs at the train station.

With full information available to all parties, market forces will produce run-free banking. Alternatively, suppose the depositors do not observe t_s or α_s, but the bank and a regulator do. Then the following arrangement is possible, but only by regulation.

Proposition 2

With full information available to the bank or regulators, a bank (or a regulator) seeking to deter runs will choose $\alpha_s^* = \max\left\{\alpha_s^{\text{AIC}}, \alpha_s^{\text{Stable}}\right\}$.

Proof

The bank is automatically stable when $\frac{r_1 t_s}{R_1} \geq \alpha_s^{\text{Stable}}$ so the regulator would always want to maximize lending and allow the bank to follow its self-interest and select that level $\left(\frac{r_1 t_s}{R_1}\right)$ of liquidity. Otherwise, the minimum amount of liquidity that is needed is α_s^{Stable}. □

More generally, for arbitrary anticipated withdrawals of t_s, α_s^{AIC} and α_s^{Stable} will differ and if liquidity, θ, is not too high or too low, and their relationship will be similar to what is shown in Fig. 2. For very low levels of anticipated withdrawals, where the condition in

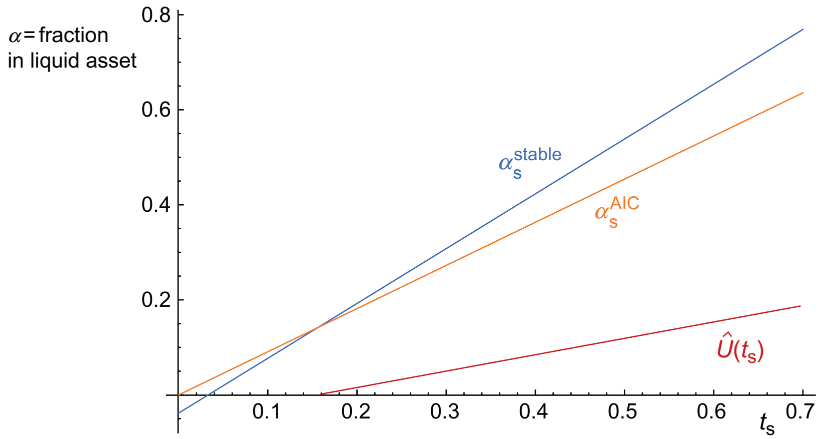


Fig. 2 Comparison of automatically incentive compatible and stable liquidity choices and the implied amount of unused liquidity held from date 1 to 2. *Note*: parameter values are $\Delta = 0.3$, $\theta = 0.5$, $R_1 = 1.1$, $R_2 = 1.33$, $r_1 = r_2 = 1$.

Proposition 1 holds, the bank is sufficiently solvent that chooses to hold more ex-ante liquidity than is need to be stable, so that runs are impossible. At some point, however, this ceases to be true and the amount needed to just be solvent in a run is higher than the bank would hold out of pure self-interest. So in this case run deterrence would require a higher level of initial liquidity. This observation will be helpful in understanding some of the regulatory trade-offs that we subsequently explore.

Note that because some liquidity must be unused, it will appear that there is an unneeded amount of liquidity. With full information, this amount will serve to deter runs and will be chosen at date 0 with all knowing the amount of normal withdrawals at date 1 t_s. If more than a fraction t_s were to withdraw at date 1, the unused liquidity could be used because all would know that a run was occurring. The bank, or regulator acting for depositors, could use liquidity holdings to deter runs in the efficient way which maximizes lending. Because depositors always desire run-free deposits, with full information, banks would be forced to hold the extra unused liquidity because otherwise no deposits would be attracted.

To summarize, with complete information, a bank will be forced to hold enough liquidity to deter runs, and its desire to maximize profits will assure that it holds no more than this amount. The next section explains why the complete information benchmark may not be very informative. Once the possibility of incomplete information is considered, we can see that arriving at run-free banking can be challenging.

3.2 Incomplete Information: Is It a Problem?

While the full-information benchmark is helpful, we think it is too extreme to be realistic. Banks disclosures may be very difficult to interpret. We describe a few compelling reasons to doubt that simply disclosing some information about liquidity holdings will make depositors (or regulators) well informed about all of these quantities. This suggests that disclosure of such information may not, by itself, force a bank to make the decisions, which they would make under complete information.

There is one important situation where incomplete information is not necessarily a problem. Even if there is no disclosure of asset holdings, depositors, who know Δ and observe a such t_s the conditions of Proposition 1 are satisfied, will know that the bank's choice will eliminate run risk in state s, because $\alpha_s^{\text{AIC}} = \dfrac{t_s r_1}{R_1} \geq \alpha_s^{\text{Stable}}$ (the bank is automatically stable in this case). If this was satisfied for all states, s, a bank would always choose a level of liquid asset holdings which always results in stability even if no one could verify those holdings and if no depositor knew the state s. Whenever this condition is not universally satisfied, the bank's incentives to hold liquidity will depend on the information available to depositors (or regulators) and on the incentives provided to the bank.

We believe that in most cases a bank's liquidity choice is not always automatically stable. This suggests that some forms of disclosure or regulation will influence its choice of liquidity. We describe two types of reasons that simple disclosure of liquidity is difficult to interpret. First, if disclosure (or a regulatory requirement) regarding liquidity only applies on some dates (such as the end of an accounting period), the bank can distort the disclosure. Second, even if a liquidity disclosure (or requirement) is on all dates, it is plausible that the bank knows much more about its customers liquidity needs than anyone else, which makes it very difficult to determine if a given level of liquidity is sufficient to make the bank stable and run-free.

3.2.1 Problems with the Periodic Disclosure of Liquidity

One important problem facing depositors is the difficulty in interpreting the kind of accounting data that must be parsed in order to decide whether to join a run. Disclosures that are made on liquidity positions typically occur with a delay and are periodic (such as at the end of a quarter or a fiscal year). The inference problem for depositors can be compounded by the temptation for banks to engage in window dressing of their accounting information.

One eye-opening example of the problem, analyzed in Munyan (2015), is the tendency of (mostly) European banks to disguise borrowing around quarter-end dates. As Munyan (2015) explains, many non-US banks are required to report their accounting information that forms the basis various regulatory ratios only on the last day of the quarter. In the United States, banks also have to show average daily ratios for critical balance sheet variables which caps the gains from manipulating end-of-quarter data. The non-US banks apparently sell some safe assets just before the end of the quarter and then buy them back shortly afterwards. This transaction allows them to report lower leverage across the quarter-end date.

The ingenious aspect of Munyan's analysis is using detailed data on the tri-party repo market to infer this behavior. He explains how the banks' would normally be borrowing in this market to fund these assets. Because they step back only briefly, their window dressing shows up in reduced repo volumes. Fig. 3 (reproduced from fig. 1 of Munyan) shows the raw data on repo volumes with the quarter-end dates indicated with dashed vertical lines. The pattern is so strong that it is clearly evident from inspection. Munyan's econometric estimates suggest that the non-US banks trim their end of quarter borrowing by about $170 billion, with the vast majority of the decline coming from European banks.

This problem of the potential window dressing of periodic disclosures is relevant to measuring bank liquidity, due to the complicated nature of the kind of liquidity information that needs to be inferred. Cetina and Gleason (2015) provide a series of examples about how the LCR is vulnerable to this type of manipulation. Some of the problems come because of the ability to use repurchase agreements (and reverse repurchase

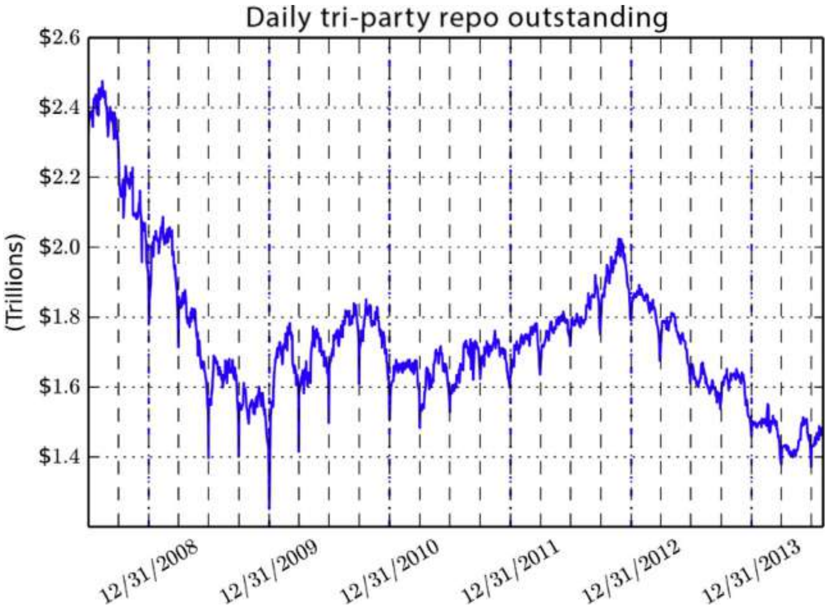


Fig. 3 Tri-party repo volumes outstanding from Munyan (2015).

agreements) to move the timing of cash flows. But the rules also distinguish between the assumed levels of liquidity of different asset class and some types of transactions can alter both the numerator and denominator of a ratio in different ways. Moreover, the computations in different jurisdictions vary which further complicates comparisons.

In summary, this possibility for window dressing implies that liquidity disclosures and regulations should hold on all dates rather than being applied periodically. In our model, this will mean that it may be difficult to credibly disclose α_s, the initial holding of liquidity, because this could be invested in illiquid loans after the disclosure. Requiring liquidity to be held on all dates (after date 1 in our model) will of course limit its use to meet withdraws of deposits. This again brings back the problem of not allowing the last taxicab to leave the station. In addition, disclosure or regulation of complicated liquidity holdings may require careful auditing (for disclosure) or supervision (for regulation).

3.2.2 Liquidity Disclosures Are Difficult to Interpret

A second challenge facing depositors and regulators in interpreting disclosed information is placing it in appropriate context. Suppose all parties are truthfully told the level of liquid asset holdings in the banking system at a given date (or even on every date). Judging whether these are adequate to service impending withdrawals requires knowledge of how far along a potential run might be on that date and how many normal withdrawals are anticipated. If a bank has a small amount of liquidity after its normal withdraws (of $t_s r_1$ in state s), this is very different than if normal withdrawals have not yet occurred. It is possible that very little additional liquidity would be needed if most potential withdrawals have already occurred. How could banks credibly communicate such information? The next section provides a model of this, based on the bank's private information about the normal level of withdrawals, t_s.

3.2.3 A Bank Has Private Information About Needed Liquidity

Before turning to the details of the model, it is helpful to provide some intuition about how private information possessed by the bank interacts with the incentives of depositors to run. Similar problems arise at both date 0 and date 1 in the model, but we will describe them in turn. One reason for separating out the discussion is because in our framework the most natural analogs to the Basel-style regulation can be thought of in terms what they imply as of different dates.

If there is no way to communicate what the bank knows, and it is not automatically stable, then disclosing a level of liquidity at date 0, α_s, which would make the bank stable only in some states of nature, t_s, will not be adequate completely eliminate runs. In these cases, depositors will have two reasons to be worried. First, in the states of nature where it would not be stable, a run would cause the bank to fail and thus would be self-fulfilling, leading to losses by depositors who did not run.

Second, because depositors do not know t_s, a depositor (whom we assume to be very risk averse) who sees a sunspot and worries about a run will always withdraw rather than face losses if the unknown state turns out to be one that makes the bank fail. As a result, a level of liquidity disclosure, which is not sufficient to makes a bank run-free for all levels of t_s, will lead to runs whenever they are feared, even for the levels of t_s where this does not cause bank failure. In the next section, we will explain why an NSFR approach to liquidity regulation (which can be mapped into restrictions on date 0 liquidity choices) can be susceptible to such concerns.

Suppose that a positive level of liquidity held at date 1, after withdrawals from a fraction f_1 of deposits, is regulated and required. It can also be very difficult to interpret this level when the normal level of withdrawals, t_s, is unknown. Any liquidity which must be held from date 1 to date 2 is not available to service withdrawals at date 1. From Proposition 2, we would like to require a level of unused liquidity $U(t_s)$ that coincides with the amount specified under full information. This amount would deter runs in state s by being available to be completely used to meet the withdrawals in a run from a fraction $t_s + \Delta$ of depositors.

When depositors must guess about the level of normal withdrawals, merely observing the actual outflows in period 1 is not necessarily enough to assure them about the safety of their deposits. To see the problem consider two levels of normal withdrawals, High and Low such that $t_{s=\text{High}} = t_{s=\text{Low}} + \Delta$. A positive level of liquidity which must be held if $f_1 = t_{s=\text{high}}$ cannot be released to meet with the same number of withdrawals during a

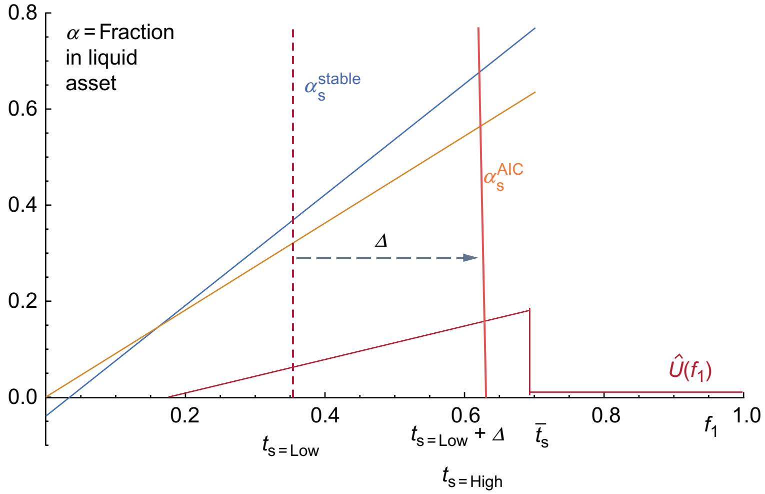


Fig. 4 Inability to distinguish between runs and large fundamental withdrawals.

run with $f_1 = t_{s=\text{low}} + \Delta$. This is shown in Fig. 4. Therefore, the full-information level of liquidity required at date 1 cannot be implemented without a way to learn the bank's information about the normal level of withdrawals, t_s. We will show how this is related to the implementation of the LCR approach to regulating liquidity in the next section.

For the balance of this chapter, we assume that liquidity on date 0 and date 1 can be measured (for example by a regulator) but that the bank has private information about the normal levels of withdrawals, t_s. This information friction alone is sufficient to study many interesting issues in the regulation of liquidity needed to make banks run-free. If the regulator cannot learn this information, it will constrain the efficiency of regulation.

4. BASEL-STYLE REGULATORY OPTIONS

Based on these observations about the efficacy of disclosure, for the remainder of our analysis we assume that liquidity on date 0 and date 1 can be measured (for example by a regulator), but we want understand the limitations that arise if the bank has private information about the normal levels of withdrawals, t_s. This information friction alone is sufficient to study many interesting issues in the regulation of liquidity needed to make banks run-free. To see how things unfold, we will begin again with a case where the regulator can also observe the state and then contrast that with what happens if the regulator cannot learn this information. Throughout we continue to assume that depositors are sufficiently risk averse so that run-free banking is social optimum which we seek to implement.

We consider two potential approaches that a regulator could pursue. These are inspired by the kinds of regulations that are proposed as part of Basel III. We suppose that she can credibly certify that the bank has some level of the liquid asset present (as a percentage of deposits). One option is to report on this ratio at the time when the liquid assets are acquired at time zero. This would amount to regulating α, and this is similar in spirit the NSFR. The NSFR requires "banks to maintain a stable funding profile in relation to the composition of their assets and off-balance sheet activities" (Basel Committee on Bank Supervision 2014). Loosely speaking, the NSFR can be thought of as forcing banks to match long-term assets with long-term funding. Our interpretation of this requirement is that the bank is free to violate the requirement temporarily in the future, so it is not always a binding restriction. As a result, it is very much like a requirement that the bank chooses a level of liquid holdings at date 0, α_s. From Proposition 2, we know that with complete information a regulation that is bank and state specific can be effective in delivering run-free banking, the question we ask now is what happens in other situations.

Alternatively, a regulator could insist that the bank will always have a certain amount of liquid assets relative to deposits at all times, including after any withdrawals. This kind of regulation is more like the LCR. The LCR requires "that banks have an adequate stock of unencumbered high-quality liquid assets that can be converted easily and immediately in private markets into cash to meet their liquidity needs for a 30 calendar day liquidity stress scenario" (Basel Committee on Bank Supervision, 2013a).

4.1 An LCR Regulation

Ultimately we are interested in understanding how the LCR works when the regulator cannot learn the bank's information about t_s. As a first step, we consider an LCR regulation where the state is known by the regulator and where the regulation says the bank must always (on both dates) hold a fraction ρ_s of deposits in liquid assets in state s. At date, 1 the bank has promised depositors $r_1(1-f_1)$. The important consequence of this is that regulation would even apply after first period withdrawals (f_1), when the bank would have to have a minimum level of safe assets equal to $\rho_s r_1(1-f_1)$.

If the bank is subject to this requirement, and it conjectures that f_1 depositors will withdraw in state s, then its optimal initial level of safe assets (α_s) will satisfy $\alpha_s R_1 = f_1 r_1 + \rho_s r_1(1-f_1)$. This choice follows trivially because it is never efficient to make loans with intention of liquidating them, and this is the minimum amount of liquid assets that will satisfy the regulation. Accordingly, the bank knows that the depositors will know this (and also understand that the bank is trying to maximize its profits). The residual value of the bank's equity will be:

$$E_2(f_1;\rho)=\begin{cases}(\alpha_s R_1-f_1r_1)R_1+(1-\alpha_s)R_2-(1-f_1)r_2 & \text{if } f_1<\dfrac{\alpha_s R_1-r_1\rho_s}{r_1(1-\rho_s)},\\ \left((1-\alpha_s)-\dfrac{f_1r_1-\alpha_s R_1+\rho_s(1-f_1)}{\theta R_2}\right)R_2 & \text{if } f_1\geq\dfrac{\alpha_s R_1-r_1\rho_s}{r_1(1-\rho_s)} \text{ and}\\ +(\rho_s R_1-r_2)(1-f_1) & \text{if } f_1\leq\dfrac{\alpha_s R_1+(1-\alpha_s)\theta R_2-\rho_s r_1(1-\theta R_1)-r_2\theta}{r_1-\rho_s r_1(1-\theta R_1)-r_2\theta},\\ 0 & \text{if } f_1>\dfrac{\alpha_s R_1+(1-\alpha_s)\theta R_2-\rho_s r_1(1-\theta R_1)-r_2\theta}{r_1-\rho_s r_1(1-\theta R_1)-r_2\theta}.\end{cases}$$

Each branch of the expression is intuitive. The top branch shows the profits that accrue when withdrawals are small enough that the bank can pay all depositors without liquidating any loans and still satisfy the LCR; this will be the case whenever $f_1r_1<\alpha_s R_1+\rho_s r_1(1-f_1)$, which when rearranged is the threshold condition that is listed. In this case, the bank has two sources of revenue, one coming from rolling over the residual safe assets after paying early depositors and the other coming from the return on the loans. The date 2 depositors must be paid and the banker keeps everything that is left.

The second branch represents a case where the bank must liquidate some loans to service the early withdrawals. In this case, the bank liquidates just enough loans so that after the deposits are paid, it exactly satisfies the LCR. The same two sources of revenues and deposit cost are present, but the formula adjusts for the liquidations. Recall that each loan that is liquidated yields θR_2 at date 1. Hence rather than having the revenue from the full set of loans $(1-\alpha_s)$ that were initially granted, the bank only receives returns on the portion that remains after some loans that were liquidated in order to pay the depositors and comply with the LCR. Because the LCR is binding from date 1 until date 2, the bank has exactly $\rho_s r_1(1-f_1)$ of the safe asset that is rolled over and that money can also be used to pay the remaining patient depositors. Notice that if the loans are totally illiquid and $\theta=0$, then there is no possibility of this second branch (where free liquidity in excess of the coverage ratio is fully used but the bank remains solvent).

The third branch obtains when the level of withdrawals is sufficiently large that the bank becomes insolvent. Insolvency occurs when $f_1>\dfrac{\alpha_s R_1+(1-\alpha_s)\theta R_2-\rho_s r_1(1-\theta R_1)-r_2\theta}{r_1-\rho_s r_1(1-\theta R_1)-r_2\theta}$ because at that point the depositors can see that the liquidations do not generate enough to fully cover the promised repayments.

The bank knows that depositors consider all these possibilities in trying to infer what the bank will do. If the coverage ratio can be set such that the bank chooses to hold sufficient liquidity to remain solvent during a run, then runs will be deterred. Proposition 3 examines the outcomes if the bank faces a state-contingent LCR in state s, $\rho_s\in[0,1]$.

Proposition 3

There is an LCR in state s, $\rho_s \in [0, 1]$ which will deter runs. When ρ_s is not zero or one it satisfies:

$$t_s + \Delta = \frac{\frac{t_s r_1 + \rho_s r_1(1-t_s)}{R_1} R_1 + \left(1 - \frac{t_s r_1 + \rho_s r_1(1-t_s)}{R_1}\right)\theta R_2 - \rho_s r_1(1-\theta R_1) - r_2\theta}{r_1 - \rho_s r_1(1-\theta R_1) - r_2\theta},$$

implying that

$$\rho_s = \frac{\theta R_1((1-t_s-\Delta)r_2 - R_2) + r_1(\Delta R_1 + t_s\theta R_2)}{r_1(\Delta R_1 + (1-t_s-\Delta)\theta R_1^2 - (1-t_s)\theta R_2)}.$$

A regulator who knows t_s can choose ρ_s so as to deter runs.

Proof

If the bank is run-free in state s and $f_1 = t_s$, then it will pick α_s to satisfy:

$$\alpha_s R_1 = t_s r_1 + \rho_s r_1(1-t_s)$$

Because the bank will be solvent for all $f_1 \leq \bar{f}_1(\rho_s) = \dfrac{\alpha_s R_1 + (1-\alpha_s)\theta R_2 - \rho_s r_1(1-\theta R_1) - r_2\theta}{r_1 - \rho_s r_1(1-\theta R_1) - r_2\theta}$, the regulator can pick ρ_s such that it delivers $t_s + \Delta \leq \bar{f}_1(\rho_s)$ and $\alpha_s = \dfrac{t_s r_1 + \rho_s r_1(1-t_s)}{R_1}$. If $t_s + \Delta \leq \bar{f}(\rho_s)$ at $\rho_s = 0$, then $\rho_s = 0$ suffices. If instead $t_s + \Delta > \bar{f}(\rho_s)$ at $\rho_s = 0$, then either a $\rho_s < 1$ solves (2) as an equality or if not then we will see that $\rho_s = 1$ satisfies (2):

$$t_s + \Delta \leq \frac{\frac{t_s r_1 + \rho_s(1-t_s)}{R_1} R_1 + \left(1 - \frac{t_s r_1 + \rho_s(r_1 1-t_s)}{R_1}\right)\theta R_2 - \rho_s r_1(1-\theta R_1) - r_2\theta}{r_1 - \rho_s r_1(1-\theta R_1) - r_2\theta} \tag{2}$$

From our assumptions that $r_1 \leq R_1$ and $r_2 \leq R_1^2$, the bank is solvent with $\rho_s = 1$ and there will always be a value of ρ between 0 and 1 which satisfies (2). If the bank is not solvent given a run with $\rho_s = 0$, then either a $\rho_s \in (0, 1)$ exists where (2) holds with equality or no $\rho_s < 1$ keeps the bank solvent and then lowest ρ_s is given by $\rho_s = \min\left[1, \dfrac{\theta R_1((1-t_s-\Delta)r_2 - R_2) + r_1(\Delta R_1 + t_s\theta R_2)}{r_1(\Delta R_1 + (1-t_s-\Delta)\theta R_1^2 - (1-t_s)\theta R_2)}\right]$. □

If the regulator chooses an appropriate level of $\rho_s \leq 1$ knowing t_s, then depositors can be sure that the bank is stable and will never want to join a run, even though they cannot observe or interpret the level of liquidity at any instant. The intuition for why the regulation (which is a combination of a rule which can be enforced and credibly auditing) is sufficient to foreclose a run, even when the bank's liquidity choice is unobservable to depositors, is straightforward. The LCR forces the bank to invest in more liquid assets than it would voluntarily prefer to hold and the depositors know that the regulator is doing this to try to prevent runs. The bank's own self-interest continues to insure that it plans to always hold enough liquid assets to cover its anticipated fundamental withdrawals, and we are assuming that it can do that perfectly. Consequently, knowing that the extra liquidity cannot be avoided removes the incentive to run.

Importantly, once the run has been prevented the liquidity still will have to remain on the bank's balance sheet. So, under these assumptions it is beneficial to force the last taxi cab to always remain at the train station.

There are several special cases where there is an interesting corner solution. If $\theta = 0$, then for all values of t_s, $\rho_s = \dfrac{r_1(\Delta R_1)}{r_1 \Delta R_1} = 1$ and as a result $\alpha_s = \dfrac{t_s r_1 + r_1(1-t_s)}{R_1} = \dfrac{r_1}{R_1}$. This implies that the bank must invest in a sufficient fraction of liquid assets to finance a withdrawal of 100% of deposits. This is not surprising because in this case the loans are so illiquid as to have no value during a run.

Alternatively, if the asset is not totally illiquid, $\theta > 0$, but $r_1 = R_1 = 1$, then the bank earns no spread between its deposits and its liquid asset holdings and there exists a possible value of t_s such that a complete run is possible, $t_s + \Delta = 1$, then for that value of t_s

$$\rho_s = \frac{\theta R_1((1-t_s-\Delta)r_2 - R_2) + r_1(\Delta R_1 + t_s \theta R_2)}{r_1(\Delta R_1 + (1-t_s-\Delta)\theta R_1^2 - (1-t_s)\theta R_2)} = \frac{\Delta - (1-t_s)\theta R_2}{\Delta - (1-t_s)\theta R_2} = 1.$$

In this case, if the bank experiences a complete run and holds just enough liquid assets to meet withdrawals in a complete run ($\alpha_s = \dfrac{r_1}{R_1} = 1$, ie, 100% liquid assets), it is just solvent (net worth is just zero), and any reduction in holdings of liquidity would make it insolvent given a run by all depositors.

To better understand how the model works, consider the following example (which is not calibrated in any particular way). Suppose the value of t_s is $t_s = ½$, and $\theta = ½$, $R_1 = 1.1$, $R_2 = 1.5$, $r_1 = r_2 = 1$, then it is possible to solve for the ρ_s needed to deter the run as a function of Δ. Fig. 5 shows this correspondence.

For these parameters, there are two interesting regions. First, up until the point when Δ reaches about 0.32, the optimal value of ρ_s is zero. In this region runs that are small enough so that the condition in Proposition 1 holds and the bank selfishly will always hold enough liquid assets so as to deter a run.

At certain point, however, the condition in Proposition 1 no longer applies and profits are no longer sufficient to prevent the run. For potential runs that are this size (or larger), ρ_s must be positive and it increases as the size of the potential run does, up until the point where a full run is a possibility.

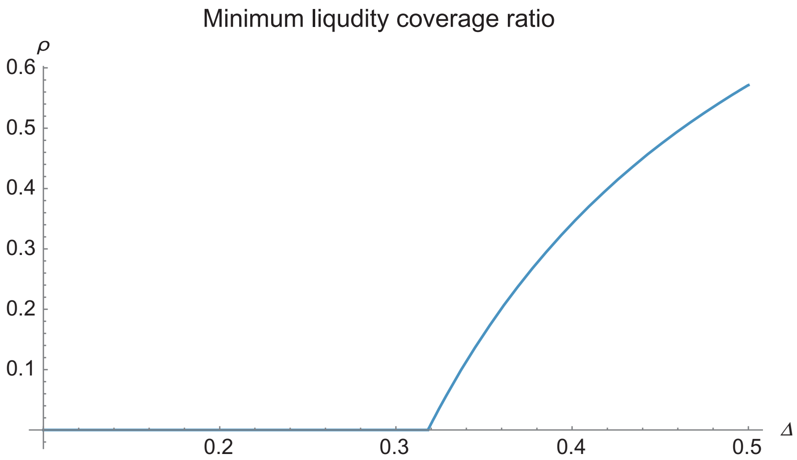


Fig. 5 Liquidity coverage ratio as a function of the potential run risk.

While a highly contingent LCR type regulation is a useful benchmark, we believe a more accurate description of this type of regulation is one where ρ is identical in all states. The constancy could arise because of limited information available to regulators and/or the desire for simplicity. Proposition 4 characterizes the optimal LCR when t_s is private information to the bank and required ratio is constant, ρ.

Proposition 4

If the regulator must specify an LCR with a constant ρ knowing only the distribution of outcomes, then a value which leads the bank to be stable for all t_s must be specified. The worst case for solvency given a run is the bank with anticipated withdrawals of $\bar{t}$ (the highest possible value of t_s). An LCR ratio which makes the bank with $\bar{t}$ anticipated withdrawals just solvent in a complete run will make all types of banks safe.

Proof

A bank of type t_s, subject to an LCR of ρ will choose $\alpha_s R_1 = t_s r_1 + \rho r_1(1 - t_s)$ and given a run, the value of its equity when withdrawals exceed t_s and $f_1 = t_s + \Delta$ is

$$E(\rho, t = t_s, f_1 = t_s + \Delta)$$
$$= \left(1 - \left(\frac{t_s r_1 + r_1 \rho(1 - t_s)}{R_1} \right) - \frac{t_s r_1 - \dfrac{t_s r_1 + \rho r_1(1 - t_s)}{R_1} R_1 + \rho r_1(1 - t_s - \Delta)}{\theta R_2} \right) R_2$$
$$+ (\rho r_1 R_1 - r_2)(1 - t_s - \Delta)$$

Define $\hat{\rho}$ to be the lowest ρ for a type $\hat{t}_s$, such that the value of equity given a run for that type will be exactly zero (so it will just be solvent). To determine the solvency of types

$t_s < \hat{t}_s$ subject to this sort of regulation, note each will choose $\alpha_s = \dfrac{t_s r_1 + \hat{\rho} r_1 (1 - t_s)}{R_1}$. Differentiating $E(\hat{\rho}, t = t_s, f_1 = t_s + \Delta)$ with respect to t_s yields:

$$\frac{\partial E(\hat{\rho}, t = t_s, f_1 = t_s + \Delta)}{\partial t_s} = r_2 + \frac{(r_1 \hat{\rho} - r_1) R_2}{R_1} - \hat{\rho} r_1 R_1.$$

From the assumption that it is more profitable to finance illiquid assets with deposits absent a withdrawal than to finance liquid asset with one period deposits, $\dfrac{r_1}{R_1} > \dfrac{r_2}{R_2}$, we know $r_2 < \dfrac{R_2 r_1}{R_1}$, which implies that:

$$r_2 + \frac{(\hat{\rho} r_1 - r_1) R_2}{R_1} - \hat{\rho} r_1 R_1 < \frac{R_2 r_1}{R_1} + \frac{(\hat{\rho} r_1 - r_1) R_2}{R_1} - \hat{\rho} r_1 R_1 = \hat{\rho} r_1 \left(\frac{R_2}{R_1} - R_1 \right) < 0.$$

The final inequality follows from the profitability of the illiquid asset (ie, $R_2 > R_1^2$). This implies that for all $t_s \leq \hat{t}$, banks are stable and no one would join an anticipated run. An LCR ratio $\rho^* = \hat{\rho}$ which makes the bank with anticipated withdrawals of $t_s = \bar{t}$ just solvent in a run of $\bar{t}_s + \Delta$ will therefore make all types of banks stable. No lower value of ρ will suffice. □

Finally, recall that we already have seen a couple of special cases where stability requires that the LCR must be set at $\rho = 1$: when either the assets are totally illiquid, $\theta = 0$, or when there is no spread earned from investing in liquidity, $r_1 = R_1 = 1$, and the worst case is a complete run, $\bar{t}_s + \Delta = 1$.

4.2 NSFR Regulation

In our interpretation of an NSFR, a bank is subject to a long-term limit on how many illiquid assets it can fund, but this is not imposed as a real time constraint at all times in the future. As a result, when a bank is subject only to an NSFR, it gets to release all of its liquidity in the event of a run. That is the initial level of liquidity is regulated but not future liquidity after withdrawals have occurred. If the regulator knows all the information as in Proposition 2, then the best NSFR is the full-information amount, $\alpha_s^* = \max\{\alpha^{\text{AIC}}, \alpha_s^{\text{stable}}\}$. This will always be better than the LCR which does not release all liquidity after a run, except in the case of a complete run where $t_s = 1 - \Delta$.

More realistically, suppose depositors or regulators can perfectly observe α_s, but do not know how many people need to withdraw for fundamental reasons (t_s) and only know its probability distribution (where we again denote the maximum value by $\bar{t}$). The bank can continue to see t_s, and all parties know Δ.

While these assumptions allow for regulations akin to the NSFR, the regulation still must be very crude. The only certain way to assure the depositors that adequate ex-ante liquidity is being held is to insist that the bank invests in enough safe assets to cover the worst case withdrawals, $\bar{t}+\Delta$. Otherwise there will be an equilibrium where there is a run under the belief that other depositors conjecture that $t_s=\bar{t}$.[h] Only covering this worst case will definitely remove the incentive to run, but whenever fewer fundamental withdrawals are required, the bank is left with many liquid assets that must be rolled over.

4.3 Comparing the LCR and NSFR

Having characterized the two types of regulation, we can now compare them. First, we contrast an NSFR which is sufficient to make stable a bank with $t_s=\bar{t}$ to an LCR which will make that same type of bank stable. Either will make stable banks of all values of t_s (and no lower values will achieve this). To illustrate the possible disadvantages of a constant NSFR, we show what happens when the worst case is $\bar{t}+\Delta=1$, and where the best possible LCR is implemented

Proposition 5

An LCR regulation can potentially support more lending than an NSFR regulation when depositors and regulators cannot condition on t_s.

Proof

The simplest way to see that this might occur is to suppose that in the worst case the run is complete, $\bar{t}+\Delta=1$. In this case, we know that $\alpha=\alpha^*=\dfrac{r_1-\theta R_2}{R_1-\theta R_2}$ is the optimal NSFR, because this is the full-information level of liquidity given by α_s^{Stable} when $t_s=1-\Delta$. But in this case, the regulator can choose $\rho=\rho^*$, where $\rho^*=\dfrac{\theta R_1((1-t_s-\Delta)r_2-R_2)+r_1(\Delta R_1+t_s\theta R_2)}{r_1(\Delta R_1+(1-t_s-\Delta)\theta R_1^2-(1-t_s)\theta R_2)}=\dfrac{\theta R_1(-R_2)+r_1(\Delta R_1+(1-\Delta)\theta R_2)}{r_1(\Delta R_1+(\Delta)\theta R_2)}$ and implement the same outcome with the same amount of liquidity when $t_s=\bar{t}=1-\Delta$ such that $\alpha^*=\dfrac{\bar{t}r_1+\rho^* r_1(1-\bar{t})}{R_1}$. Because a run on a bank with $\bar{t}+\Delta=1$ will be complete, all its liquidity can be released in a run (the LCR becomes $\rho^*(1-\bar{t}-\Delta)=0$). From Proposition 4, this LCR will make stable the other types of banks with lower $t_s<\bar{t}$, and they will be able to invest a smaller amount in liquid assets $\alpha_s=\dfrac{t_s r_1+\rho^* r_1(1-t_s)}{R_1}$. Because they are stable, there will not be runs and they will never need to liquidate illiquid assets. Each bank will choose $\alpha_s=\dfrac{t_s r_1+\rho^*(1-t_s)}{R_1}<\alpha^*$ while a bank subject to the NSFR would still have to hold α^*. □

The complete run case is some sense the most favorable environment for the LCR-style regulation because in the event of a full run, the requirement to maintain extra

[h] Because depositors are very risk averse and there is a positive probability of receiving zero if there is a run, then a signal (observed by a fraction Δ of depositors) which indicates a positive probability of a run will always lead to a run if the other depositors who see the signal believe that it will.

liquidity after the first date is irrelevant. In this case, the last taxicab is allowed to depart (because $\rho^*(1-t_s-\Delta)=0$). If the worst possible case involves only a partial run, there would then be a trade-off because the incentive effects of the LCR require that some liquid assets remain on the balance sheet and the NSFR ratio does not. Further, if there is no private information (uncertainty about t_s), then the NSFR achieves the full-information outcome, with $\alpha_s=\alpha_s^{\text{Stable}}$ (and for a partial run, the LCR cannot).

These polar cases provide some general guidance about the relative efficacy of the two types of regulations. The LCR will work well when monitoring the bank's liquidity is difficult because the regulation forces the bank to carry more safe assets than it would prefer to. Depositors understand this and in some cases this will be enough to quell any concerns about the bank having insufficient funds to withstand a run.

The main cost of the LCR is that deterring the run requires the bank to continue to have some funds invested in liquid assets, even if a run has occurred. Ex-post this liquidity is inefficient and everyone would be better off if more loans had been made instead. But, the incentive effects vanish if the depositors are not convinced that the liquidity will always be present. The only situation when this is not true in the case of a full run.

Conversely, the NSFR is an attractive run deterrent when the regulator is well informed about the fundamental deposit outflows, so that initial liquidity requirement can be varied. In this case, the bank can be forced to hold just enough to survive a run, but never have to hold more than is needed. Importantly, during a run a bank subject to an NSFR can always use all of its liquid assets to serve depositors. So this kind of regulation does not require the bank to liquidate any more loans than is necessary, and hence in the best-case it avoids the inefficiency associated with the LCR.

Once the regulator does not have good knowledge about the fundamental needs of the depositors, using the NSFR becomes less efficient. In this case, depositors cannot generally be confident that the bank will have a portfolio that guarantee solvency in all cases. The best the regulator can, therefore, accomplish is to protect against a worst case set of withdrawals. This can remove the incentive to run, but doing so will mean that all but the worst case the bank over-invests in safe assets. The LCR potentially is less distorting in this case.

This intuition suggests that the relative advantages of the two approaches to regulation will hinge on two considerations. One is the variability of potential fundamental withdrawal requirements. When t_s fluctuates considerably, then regulation that relies on a fixed value of α will only deter runs if the liquidity requirement is set high enough to cover the worst case outcome. When the worst case does not materialize, this will result in the banking holding surplus liquidity. Because the LCR regulation exploits the bank's knowledge about impending withdrawals and relies on its incentives to plan for these withdrawals, variability of t_s is not as severe a problem for this kind of regulation.

The other consideration is the size of the runs that are possible. The Achilles' heel of the LCR is that even after a run has taken place, the bank must continue to hold liquid assets. The NSFR avoids this (ex-post) inefficiency because all the liquid assets that the bank has can be used in the event of a run. So if runs are never complete, the inefficiency associated with the LCR will be at a disadvantage.

It strikes us that the information requirements that would favor the NSFR are relatively onerous. One of the most difficult challenges in a real-time crisis is gauging the extent of a run. In that case, even if it possible to verify and certify that some liquid assets are present at any given point in time, it make be difficult to forecast whether they will be adequate to meet potential subsequent withdrawals. Hence, releasing all liquidity on hand can be risky.

One can see a further disadvantage of the NSFR by introducing the possibility that the bank can secretly alter its liquidity holdings after meeting the NSFR, and this can happen at date 0. This is similar to window dressing when liquidity must be reported only at the end of a calendar year. In this case, liquidity must be disclosed at every date and there must be a future commitment to hold liquidity. The LCR is just such a commitment. It does not do as well as the full-information commitment, but it does succeed in forcing the bank to remain free of runs (while an NSFR single disclosure will not).

5. EXTENSIONS

Having characterized the properties of Basel-style regulations in this model, we now discuss the implications of extending the model in two directions. First, there is no reason to restrict regulations to only look like the NSFR and the LCR, so it makes sense to expand the range of regulatory tools considered. Diamond and Kashyap (2016) provide a complete analysis of how to optimally regulate liquidity in this kind of a model, and we begin with a review of those results.

Second, we discuss several of the issues that arise if the bank faces capital regulation. Allowing savers to have a choice between investing in deposits and equity greatly complicates the model. Part of the complication comes because our model abstracts from asset risk, and many of the benefits of capital regulation arise from creating a buffer against loan losses so that any discussion of capital without asset risk is necessarily incomplete. Nonetheless, there are a couple of interesting possible comparisons between capital and liquidity regulation that can be made even without developing a full-blown model.

5.1 Optimal Regulation of Liquidity

Stepping away from the Basel approach, how should liquidity optimally be regulated in this kind of environment? To find the most efficient set of choices which can be implemented, we describe the results from undertaking a mechanism design analysis.[i] This will achieve the best outcome by providing incentives for the bank to reveal to a regulator the

[i] The analysis here is a special case of the more general treatment in Diamond and Kashyap (2016).

information needed to implement run-free banking most efficiently. Proposition 2 already describes the full-information choices, and it turns out that these can be implemented with the optimal regulation.

To understand what happens where t_s is known only to the bank, we describe a mechanism to which accounts for this information asymmetry and still induces the bank to make efficient choices. The challenge in this situation is that a bank with private information about t_s could have an incentive to misreport t_s. The condition for efficient investment from the bank's point of view without a run remains $\alpha_s = \frac{r_1 t_s}{R_1}$. When the conditions of Proposition 1 regarding the range of possible withdrawals are satisfied, this level of liquidity automatically leads to a stable, run-free bank. This is what both the bank and its depositors desire and runs will be avoided without any regulation or even any disclosure.

When the bank is not automatically stable, to make it incentive compatible to honestly report t_s, the bank must be provided an incentive for reporting high levels of anticipated withdrawals that offsets any increased profits that could arise from underreporting. The potential gains from underreporting come from making more loans and hence having less unused liquidity which is held after normal withdrawals occur. Diamond and Kashyap (2016) prove that under our assumptions there is a way to implement the full-information choice of $\alpha_s{}^*$ (from Proposition 2) and, which is similar to, but not exactly the same as an LCR requirement.

This is possible whenever the regulator has sufficient tools to penalize the banker when actual withdrawals deviate from those that the banker reports are anticipated. These tools share the feature that they eliminate the profits that accrue from underreporting. There are various tools that can achieve this outcome. For instance, one approach is to place limits on compensation whenever reports turn out to be inaccurate (to reduce spoils from underreporting). Another strategy is to deploy fines that would be tied to the use of the supposedly required liquidity given the report. If such tools exist, then the regulator can require the bank to hold $\alpha_s{}^*$ and can punish any cases where the unused liquidity after the withdrawals departs from what would be needed when the bank is run-free (ie, when actual withdrawals, f_1, deviate from what the bank reports as anticipated withdrawals, t_s), but also allow the bank to use the extra liquidity if a run were to occur.

In other words, it is possible to implement the full-information outcomes because if the bank can be induced to be run-free, the actual withdrawals, f_1, will be exactly equal to the (state-contingent) fraction of normal withdrawals, t_s. In essence, an honestly reported value of t_s allows the regulator to determine whether the realized withdrawal f_1 is or is not due to a run and release liquidity only in a run. The critical decision by the regulator is to carefully choose how much mandated excess liquidity must be held in all circumstances to create the right incentives for the banker to truthfully report anticipated fundamental withdrawals. Diamond and Kashyap (2016) characterize these choices under various assumptions about the nature of run risk.

The formal mechanism design problem in Diamond and Kashyap (2016) solves for the optimal mechanism by looking for the one where the bank is given incentives to honestly report its private information to the regulator and the regulator uses the honestly reported information to choose a run-free level of liquidity to make the bank stable. Once the bank is run-free, any misreporting of t_s by the bank will be measured by a level of withdrawals, f_1, which differ from t_s. Assessing a sufficiently large penalty (such as driving the banker's compensation to zero without imposing losses on depositors) for such a misreport will provide incentives for accurate reporting without the need to distort liquidity holding away from the full-information level.

As in the analysis in this chapter, the full-information level of liquidity results in excess liquidity held from date 1 to date 2, but with the optimal mechanism all of this liquidity can always be used if a run should occur. While the excess liquidity is available for use, because it deters runs, it is in fact never needed. Returning to our metaphor, the last taxicab is allowed to leave the station, but in equilibrium there are enough cabs such that some always remain at the station.

5.2 Integrating Liquidity Regulation with LOLR Policy

If the regulator can also serve as an LOLR, then the efficient mechanism, that we just described earlier, can be implemented by requiring a level of liquidity holdings which depends on the quantity of deposits. This is essentially a generalized LCR. In this case, no actual report of anticipated withdrawals, t_s, is required. This is implemented by requiring the amount of unused liquidity at date 1 (to be held until date 2) equal to the full-information level from Proposition 2, α_s^{Stable}, which is given by $U(f_1 = t_s) = \text{Max}\left[0, \alpha_s^{\text{Stable}} - \alpha_s^{\text{AIC}}\right] R_1$, and allowing the bank to use this liquidity in a run, but with a penalty which drives banker compensation to zero in that case.

The goal from this policy is to induce the bank to always use its private information to choose to hold just enough liquidity to make sure that after normal withdrawals, $t_{s,}$ it will meet the requirement, $U(f_1 = t_s)$. This is the equal to the investment in liquidity from the full-information level presented in Proposition 2.

To accomplish this outcome, liquidity requirements and LOLR policy should be integrated in the following way. Banks are forced to hold the specified amount of liquidity but are allowed to borrow against it for use during a run. If there is a sufficient penalty to the bank for violating its liquidity requirement, the bank will hold the specified amount of liquidity and will never use borrowing to meet normal withdrawals. As a result, the run will be deterred and the extra liquidity need not be borrowed against.

Remarkably, there is historical precedent for this sort of policy: the original United States Federal Reserve Act prohibited dividend payments for banks which were in violation of the reserve (liquidity) requirement. In that period, most banks were closely held, implying that a dividend was a significant part of management compensation. This policy is not necessarily akin to charging a high interest rate for such borrowing, because a penalty rate could be so severe that it might make the bank fail due to a run (making the bank unstable and defeating the purpose of holding extra liquidity).

Note that this type of LOLR lends against liquid assets, allowing them to be used during a crisis while providing incentives to get the bank to hold the higher level of liquidity needed to make it stable. This lending does not have moral hazard of inducing the bank to hold excessive amounts of illiquid assets (as described in Bagehot (1873), Goodfriend and King (1988), and Diamond and Rajan (2012)). It may appear pointless to lend against liquid assets, but the ability to penalize the bank for such the borrowing induces the bank to make the proper ex-ante liquidity choice. Once the liquidity is in place, its existence can deter bank runs. Finally, notice that if the LOLR acquired (or lent against) illiquid assets and could then only recover their illiquid value, θR_2, if the bank were to fail, then lending an amount in excess of this value could distort bank incentives and lead to losses by the LOLR. Lending against liquid assets has no such problem.

To summarize, the optimal mechanism induces a bank to hold excess liquidity but allows access to it during a run. The robust conclusion from this analysis is that the optimal regulation requires less unused liquidity than the simple Basel-style regulations because the excess liquidity can be released if a run should occur. If there are additional constraints on what the regulator can do, which limit the ability to release this liquidity, then a regulation like the LCR could be nearly optimal. If all liquidity cannot be released in a run, then the best regulations will have the property that as anticipated withdrawals rise, the amount of required surplus liquidity falls.

5.3 Interactions Between Capital and Liquidity Regulations

Finally, it is worth noting several observations about interactions between capital and liquidity regulation. In our baseline model, there is no credit risk associated with loans, so the usual arguments for capital requirements do not hold. Generically, however, the incentive to run is still related to depositors' assessments about the solvency of the bank so the presence of equity could still matter.

The role that capital would play in deterring a run is subtle. On the one hand, if the bank issued capital (nondemandable liabilities) and invested the proceeds in loans, this can leave the bank more solvent when a fixed number of deposits are withdrawn, moving the bank to a situation where it is solvent during a (potential) run of fixed size. This is due to the liquidation value of the additional loans made. On the other hand, added equity would be irrelevant if a (potential) run of given size given is still going to make the bank insolvent. In our framework, this is easiest to see if the liquidation value of the loans (θ) is zero. In that case, the future value of the assets that would otherwise be the basis of the equity value would be of no value in a run. So the liquidity requirements needed to deliver stability would be unchanged, and capital requirements would be completely ineffective.

Once assets become risky, the analysis becomes much more complicated. In this kind of environment, depositors will make withdrawals based both on their fundamental liquidity needs and based on beliefs about the future value a bank's assets. In addition, if the bank can fail simply because loans turn out to default, the bank's choice between loans and liquid assets can also be distorted if there is limited liability; banks in this kind of an environment can in some situations have an incentive to shift risk on depositors.

A full analysis of this kind of model is beyond the scope of our survey, especially because there are so many additional assumptions that are needed to maintain tractability. However, Kashyap et al. (2015) have solved one particular version of this kind of model, and their analysis does deliver one apparently general result about the interactions between capital and liquidity requirements in deterring runs that is worth mentioning.

They show that there is a fundamental asymmetry in the way that liquidity and capital regulations work in preventing runs. Capital requirements essentially work on the liability-side of a bank's balance sheet without directly constraining the bank's asset choices. Hence, when a bank is forced to have higher equity, it can on the margin reduce its reliance on deposit financing. The need for fewer deposits means that the bank can marginally reduce liquid asset holdings too. This frees up the bank to make marginally more loans. While this marginal adjustment is not enough to raise the overall risk of a run, it does suggest that the bank's assets will become less liquid.[j]

Conversely, liquidity regulation, either in the form of an LCR or NSFR, work very differently. The LCR, as we have seen, directly forces the bank to substitute from illiquid assets toward liquid assets. So the run deterrence automatically is accompanied by having less liquidity risk. The NSFR forces the bank to finance illiquid assets with long-term liabilities. Therefore, if the bank wants to take on additional illiquid assets, it cannot fund them with runnable deposits. Instead, short-term deposits will shrink along with liquid assets.

Kashyap et al. (2015) describe many other ways in which capital regulations and liquidity regulations can complement or substitute for each other. The asymmetry in how they marginally influence asset illiquidity is robust.

6. CONCLUSION

Our analysis provides some novel insights that can inform subsequent discussions of how to design liquidity regulation. Our starting point is the recognition that for a forward looking intermediary, anticipated withdrawals, and access to other funding influences the desired ex-ante, profit-maximizing choice of how much liquidity to hold. Absent

[j] This is not arising because of a Modigliani–Miller type fallacy whereby depositors fail to recognize that the bank's deposits are safer. Instead, this happens because the liquid assets are held only to deter runs, and when capital requirements make them less likely the bank cut back on liquid assets.

any regulation, the bank will voluntarily opt to hold more liquidity when higher exogenous deposit reductions are anticipated. Hence, it is helpful to understand whether, and when, this incentive alone will lead to banking stability even when it is not directly a goal of the bank.

In this kind of model that we have explored, stability is not guaranteed when bank assets are sufficiently illiquid and profitability is below a certain level because depositors may have doubts about whether the bank will make choices which lead it to able to withstand a panic. The lack of confidence that creates this problem can arise for various reasons. Banks are opaque and even for sophisticated counterparties assessing their balance sheet can be challenging. Information about the balance sheet is rarely available contemporaneously, so some forecasting (about the bank's condition and the decisions of other depositors) is inevitable. This will cause problems when the bank's incentives are not automatically aligned with enhancing stability.

Imperfect information also creates a problem for the bank. Cutting back on lending and holding additional liquidity is not fully rewarded when depositors cannot determine if the given amount of liquidity is sufficient to make the bank stable, so the bank's private incentive to become super-safe is limited (unless it can show depositors that it is sufficiently stable to cover all possible circumstances). Regulation that mandates some additional liquidity can potentially circumvent this problem.

Analogs to both of the two regulations contemplated as part of the Basel process, the NSFR and LCR, are among the various types of regulations that we explore. These can arise as approximations of a general type of regulation that is optimally designed to resolve the information friction. All of the ones we consider are designed to eliminate runs.

The generic form of the optimal regulation specifies that the bank must hold a level of liquid assets that is tied to anticipated withdrawals, but which often will exceed the level that it would choose on its own. If the regulator is well-informed about these withdrawals (and the risk of a run), then there are many equivalent ways to guarantee that the bank makes adequate liquidity choices. In particular, stability can be achieved either by having the bank hold the correct amount of liquid assets up-front as with an NSFR, or by imposing restrictions that require liquidity be available even after withdrawals are underway (as with an LCR). Using combinations of these kinds of policies will work too.

To achieve the efficient outcome (which in our model is the same as that which would prevail with full-information available to all), the regulator must be able to induce the bank to disclose everything it knows about the deposit risk that it faces (or have access to that information from some other way). With the ability to impose taxes on bank compensation, the regulator could elicit this from the bank. This need not even involve any direct communication of information to regulators by the bank. A liquidity regulation combined with an LOLR policy which penalizes liquidity regulation violations by limiting compensation, but allows the bank to borrow can implement this optimal arrangement.

One generic property of all of the optimally designed regulations when banks are not automatically stable is that they involve requiring the bank to hold some liquidity that goes unused. So even in the best possible case, the last taxi cab often remains at the station. Fundamentally, this occurs because the unused liquidity is needed to deter the run.

There are two separate forces that lead to this result. First, a prudent provision that forecloses a run necessarily requires that the bank has enough liquidity to be able to service depositors if they did run. This might be possible through liquidating loans. But liquidations are highly inefficient so this typically this will not be sufficient and the bank needs to have some liquid assets which could be deployed if needed. By mandating the "dry powder," the regulator preserves solvency in a run and thus removes the depositors' incentive to run.

The second consideration is that a regulator cannot count on being able to distinguish a run from a situation where fundamental withdrawal needs are simply high. The goal in preventing runs is to do so without mandating more dry powder than is needed. Unfortunately, even when exceptionally high levels of withdrawals are anticipated, some dry powder is needed.

These observations suggest are a number of other directions that would be interesting to explore. In Diamond and Kashyap (2016), we generalize the environment to allow for different types of run dynamics and investigate the implications for regulation. Let us close with three much broader issues that merit further consideration.

First, our analysis suggests a novel type of interaction between LOLR policy and liquidity regulation. Most discussions of the LOLR start with the Bagehot dictum of lending freely against good collateral but at a penalty rate. However, the reason which many loans are illiquid is because they are difficult to quickly value and their value may depend on actions or relationships of the bank. A system where all assets were illiquid and all liquidity (even for normal withdrawals) is provided by the LOLR could be highly problematic. If a private bank is to provide much of its own liquidity, our analysis shows that there is a role for integrating liquidity regulation with an LOLR which lends against required liquid holdings of a bank. This allows banks to access to liquidity without distorting their incentives to minimize the risk of a run.

Carlson et al. (2015) make one attempt to investigate the degree to which liquidity requirements and LOLR policies complement each other. More work in this vein that could probe other interactions between these tools seems promising.

Second, it would also be interesting in future research to examine other mechanisms to provide incentives for banks to hold sufficient liquidity to make them stable and run-free. We focus on liquid asset quantity requirements, but there may be interesting price-based mechanisms. One example is adjusting the interest rates paid on central bank reserves. This is especially relevant in times (like today) when the aggregate quantity of central bank reserves is large in many countries. On a related note, the large central bank balance sheets and the low interest rates in the many counties today make it difficult to use historical data to calibrate the incentive effect of liquidity requirements or the effects of changes in the interest on reserves on endogenous liquidity holdings.

Finally, there are interesting issues involving the need for and effect of liquidity regulation on interbank competition for funding and liquidity sharing between banks. When banks can raise liquidity from the customers of other banks (or from others outside the banking system) then some interactions that we have ignored come into play. As noted by Bhattacharya and Gale (1987) and Farhi et al. (2009), in these circumstances, liquidity regulation can be needed to prevent banks from free riding on others' liquidity. This becomes even more difficult if some of the participants in the market are unregulated "shadow banks." It would be interesting to examine how this interacts with our notion of providing incentives for banks to choose an efficient level of liquidity based on their private information about their own future needs for liquidity.

ACKNOWLEDGMENTS

We thank Franklin Allen, Gary Gorton, Guido Lorenzoni, Annette Vissing-Jorgenson, Nancy Stokey, Nao Sudo, John Taylor, Harald Uhlig, and seminar participants at the Asian Development Bank Institute, East Asian Economic Seminar, Imperial College, Melbourne Institute Macroeconomic Policy Meetings, National Bureau of Economic Research Monetary Economics meeting, the Bank of England, European Central Bank, Centre for Economic Policy Research and Centre for Macroeconomics Conference on Credit Dynamics and the Macroeconomy, Riksbank, and the University of Chicago for helpful comments and Adam Jorring for expert research assistance. We thank the Initiative on Global Markets at Chicago Booth, the Fama Miller Center at Chicago Booth and the National Science Foundation for grants administered through the NBER for research support. All errors are solely our responsibility.

REFERENCES

Admati, A., Hellwig, M., 2013. The Bankers' New Clothes: What's Wrong with Banking and What to Do about It. Princeton University Press.

Aikman, D., Haldane, A., Kapadia, S., 2013. Operationalising a macroprudential regime: goals, tools and open issues. Banco Espana Financ. Stability J. 24, 9–30.

Allen, F., 2014. How Should Bank Liquidity Be Regulated? Mimeo, Imperial College London.

Allen, F., Gale, D., 1997. Financial markets, intermediaries, and intertemporal smoothing. J. Polit. Econ. 105 (3), 523–546.

Bagehot, W., 1873. Lombard Street: A Description of the Money Market. H. S. King, London.

Bao, J., David, J., Han, S., 2015. The Runnables, FED Notes, Board of Governors of the Federal Reserve. https://www.federalreserve.gov/econresdata/notes/feds-notes/2015/the-runnables-20150903.html.

Baron, D.P., Myerson, R.B., 1982. Regulating a monopolist with unknown costs. Econometrica 50, 911–930.

Basel Committee on Bank Supervision, 2013a. Basel III: The Liquidity Coverage Ratio and Liquidity Risk Monitoring Tools. Bank for International Settlements, Basel, Switzerland.

Basel Committee on Bank Supervision, 2013b. Liquidity Stress Testing: A Survey of Theory, Empirics and Current Industry and Supervisory Practices. Bank for International Settlements. Basel Committee on Bank Supervision Working Paper 24.

Basel Committee on Bank Supervision, 2014. Basel III: The Net Stable Funding Ratio. Bank for International Settlements.

Benston, G.J., Smith, C.W., 1976. A transactions cost approach to the theory of financial intermediation. J. Financ. 31 (2), 215–231.

Bhattacharya, S., Gale, D., 1987. Preference shocks, liquidity and central bank policy. In: Barnett, W.A., Singleton, K.J. (Eds.), New Approaches to Monetary Economics. Cambridge University Press, Cambridge.

Black, F., 1975. Bank funds management in an efficient market. J. Financ. Econ. 2 (4), 323–339.

Borchgrevink, H., Ellingsrud, S., Hansen, F., 2014. Macroprudential Regulation: What, Why and How. Norges Bank Staff Memo Number 13, 2014.

Bouwman, C.H.S., 2015. Liquidity: how banks create it and how it should be regulated. In: Berger, Al., Molyneux, P., Wilson, J. (Eds.), The Oxford Handbook of Banking, second ed. Oxford University Press, Oxford, UK, pp. 184–218.

Brooke, M., Bush, O., Edwards, R., Ellis, J., Francis, B., Harimohan, R., Neiss, K., Siegert, C., 2015. Measuring the Macroeconomic Costs and Benefits of Higher UK Bank Capital Requirements. Bank of England Financial Stability Paper No. 35.

Calomiris, C.W., Kahn, C.M., 1991. The role of demandable debt in structuring optimal banking arrangements. Am. Econ. Rev. 81 (3), 497–513.

Calomiris, C.W., Heider, F., Hoerova, M., 2014. A Theory of Bank Liquidity Requirements. Columbia Business School Research Paper No. 14-39.

Carlson, M., Duygan-Bump, B., Nelson, W., 2015. Why Do We Need Both Liquidity Regulations and a Lender of Last Resort? A Perspective from Federal Reserve Lending during the 2007–09 U.S. Financial Crisis. Board of Governors of the Federal Reserve System, Washington. Finance and Economics Discussion Series 2015-011, http://dx.doi.org/10.17016/FEDS.2015.011.

Cerutti, E., Claessens, S., Laeven, L., 2015. The Use and Effectiveness of Macroprudential Policies: New Evidence. International Monetary Fund WP/15/61.

Cetina, J., Gleason, K., 2015. The Difficult Business of Measuring Banks' Liquidity: Understanding the Liquidity Coverage Ratio. Office of Financial Research Working Paper 15-20.

Čihák, M., Demirgüç-Kunt, A., Martínez Pería, M.S., Mohseni-Cheraghlou, A., 2013. Bank Regulation and Supervision Around the World: A Crisis Update. World Bank Policy Research Working Paper 6286.

Claessens, S., Kodres, L., 2014. The Regulatory Responses to the Global Financial Crisis: Some Uncomfortable Questions. International Monetary Fund Working Paper 14/46.

Clement, P., 2010. The term "macroprudential": origins and evolution. BIS Q. Rev. 2010, 59–67.

Cooper, R., Ross, T.W., 1998. Bank runs: liquidity costs and investment distortions. J. Monet. Econ. 41 (1), 27–38.

Crockett, A., 2000. Marrying the micro- and macro-prudential dimensions of financial stability. Remarks by Mr. Andrew Crockett, General Manager of the Bank for International Settlements and Chairman of the Financial Stability Forum, before the Eleventh International Conference of Banking Supervisors, held in Basel, 20–21 September.

Diamond, D.W., 1984. Financial intermediation and delegated monitoring. Rev. Econ. Stud. 51 (3), 393–414.

Diamond, D.W., 1997. Liquidity, banks and markets. J. Polit. Econ. 105, 928–956.

Diamond, D.W., Dybvig, P.H., 1983. Bank runs, deposit insurance and liquidity. J. Polit. Econ. 91 (3), 401–419.

Diamond, D.W., Kashyap, A.K., 2016. Optimal Regulation of Bank Liquidity. (still in preparation).

Diamond, D.W., Rajan, R.G., 2001. Liquidity risk, liquidity creation and financial fragility: a theory of banking. J. Polit. Econ. 109 (2), 287–327.

Diamond, D.W., Rajan, R.G., 2012. Illiquid banks, financial stability, and interest rate policy. J. Polit. Econ. 120 (3), 552–591.

Ennis, H., Keister, T., 2006. Bank runs and investment decisions revisited. J. Monet. Econ. 53 (2), 217–232.

Farhi, E., Golosov, M., Tsyvinski, A., 2009. A theory of liquidity and regulation of financial intermediation. Rev. Econ. Stud. 76 (3), 973–992.

Financial Stability Board, 2015. Transforming Shadow Banking into Resilient Market-based Finance: An Overview of Progress. Financial Stability Board Working Paper.

Fisher, P., 2015. The Financial Regulation Reform Agenda: What Has Been Achieved and How Much Is Left to Do? Speech at Richmond, the American International University, London 30 September 2015.

Goodfriend, M., King, R.G., 1988. Financial deregulation, monetary policy, and central banking. Fed. Reserve Bank Richmond Econ. Rev. 74 (3), 3–22.

Goodhart, C.A.E., 2008. Liquidity risk management. Banque France Financ. Stability Rev. 12, 39–44.

Gorton, G., Winton, A., 2003. Financial intermediation. In: Constantinides, G.M., Harris, M., Stulz, R. (Eds.), The Handbook of the Economics of Finance: Corporate Finance, North Holland, pp. 431–552.

Hanson, S.G., Kashyap, A.K., Stein, J.C., Winter 2011. A macroprudential approach to financial regulation. J. Econ. Perspect. 25 (1), 3–28.

Jacklin, C.J., 1987. Demand deposits, trading restrictions, and risk sharing. In: Prescott, E.C., Wallace, N. (Eds.), Contractual Arrangements for Intertemporal Trade. University of Minnesota Press, Minneapolis, MN, pp. 26–47.

Kashyap, A.K., Tsomocos, D.P., Vardoulakis, A.P., 2014. Principles for macroprudential regulation. Banque France Financ. Stability Rev. 18, 173–181.

Kashyap, A.K., Tsomocos, D.P., Vardoulakis, A.P., 2015. How Does Macroprudential Regulation Change Bank Credit Supply? Revision of National Bureau of Economic Research Working Paper 20165.

Martynova, N., 2015. Effect of bank capital requirements on economic growth: a survey. De Nederlandsche Bank Working Paper, DNB Working Paper No. 467.

Mester, L., Nakamura, L., Renualt, M., 2007. Transactions accounts and loan monitoring. Rev. Financ. Stud. 20 (3), 529–556.

Modigliani, F., Miller, M.H., 1958. The cost of capital, corporate finance and the theory of investment. Am. Econ. Rev. 48 (3), 261–297.

Munyan, B., 2015. Regulatory Arbitrage in Repo Markets. Office of Financial Research Working Paper 15-22.

Myers, S.C., 1977. Determinants of corporate borrowing. J. Financ. Econ. 5, 147–175.

Norden, L., Weber, M., 2010. Credit line usage, checking account activity, and default risk of bank borrowers. Rev. Financ. Stud. 23 (10), 3665–3699.

Rochet, J.C., 2014. The Extra Cost of Swiss Banking Regulation. Swiss Finance Institute White Paper.

Santos, J.C., Suarez, J., 2015. Liquidity Standards and the Value of an Informed Lender of Last Resort. Working paper, Federal Reserve Bank of New York, May.

Stein, J.C., 2012. Monetary policy as financial-stability regulation. Q. J. Econ. 127 (1), 57–95.

Uhlig, H., 2010. A model of a systemic bank run. J. Monet. Econ. 57, 78–96.

Vives, X., 2014. Strategic complementarity, fragility, and regulation. Rev. Financ. Stud. 27 (12), 3547–3592.

Understanding Inflation as a Joint Monetary–Fiscal Phenomenon

E.M. Leeper*, **C. Leith**†
*Indiana University and NBER, IN, United States
†University of Glasgow, Glasgow, United Kingdom

Contents

Handbook of Macroeconomics
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.012

Abstract

We develop the theory of price-level determination in a range of models using both ad hoc policy rules and jointly optimal monetary and fiscal policies and discuss empirical issues that arise when trying to identify monetary–fiscal regime. The chapter concludes with directions in which theoretical and empirical developments may go.

Keywords

Monetary policy, Fiscal policy, Price level determination, Optimal policy, Tax smoothing, Government debt

JEL Classification Codes

E4, E5, E6, H3, H6

1. INTRODUCTION

There is a long tradition in macroeconomics of modeling inflation in stable economies by focusing on monetary policy and abstracting from fiscal policy.[a] As the global financial crisis and its aftermath rocked the world economy, the tenability of that modeling approach has been strained.

This chapter introduces readers to the interactions between monetary and fiscal policies and their role in determining macroeconomic outcomes, particularly the aggregate price level. By incrementally widening the scope of those interactions and considering both simple ad hoc rules and optimal policy, we aim to make accessible the intricacies that policy interactions entail. We hope the material will entice young macroeconomists to engage a set of issues that we regard as both not fully resolved and fundamental to macroeconomic policy analysis.

1.1 Some Observations

Let us start with a few observations of economic developments since 2008:

1. Many countries reacted to the financial crisis and recession that began in 2008 with joint policy actions that sharply reduced monetary policy interest rates and implemented large fiscal stimulus packages.
2. Central banks reacted to the financial crisis by purchasing large quantities of private assets and government bonds in actions that bear a striking resemblance to fiscal policy (Brunnermeier and Sannikov, 2013; Leeper and Nason, 2014).
3. Sovereign debt crises in the Euro zone culminated in the European Central Bank's 2012 policy of "outright monetary transactions," a promise to purchase sovereign debt in secondary markets in unlimited quantities for countries that satisfied conditionality restrictions.
4. Rapid adoption of fiscal austerity measures beginning in 2010 and 2011 created challenges for central banks that were already operating at or near the lower limits for nominal interest rates.
5. Exploding central bank balance sheets also grew riskier, increasing concerns about whether the requisite fiscal backing or support for monetary policy is guaranteed (Del Negro and Sims, 2015).
6. In 2013, Japan's newly elected prime minister Shinzō Abe adopted "Abenomics," a mix of fiscal stimulus, monetary easing, and structural reforms designed to reinflate a Japanese economy that has languished since the early 1990s.

[a] Focusing on stable economies rules out hyperinflations, which are widely believed to have fiscal origins.

Table 1 Net general government debt as percentage of GDP

	2008	2015
Euro area	54.0	74.0
Japan	95.3	140.0
United Kingdom	47.5	85.0
United States	50.4	80.9

Projections for 2015.
Source: International Monetary Fund, 2014. Fiscal Monitor-Back To Work: How Fiscal Policy Can Help. IMF, Washington, DC.

7. Table 1 reports that government debt expansions during the recession were significant: net debt as a share of GDP rose between 37% and 79% across four advanced-economy country groups. As central banks begin to raise interest rates toward more normal levels, these debt expansions will carry with them dramatically higher debt service to create fresh fiscal pressures. The Congressional Budget Office (2014) projects that U.S. federal government net interest payments will rise dramatically as a share of GDP from 2014 to 2024. Evidently, there are substantial fiscal consequences from central bank exits from very low policy interest rates.
8. With an increasing number of central banks now paying interest on reserves at rates close to those on short-term government bonds, one important distinction between high-powered money and nominal government bonds has disappeared, removing a principal distinction between monetary and fiscal policy (Cochrane, 2014).
9. Sovereign debt troubles in the Euro area and political polarization in many countries remind us that every country faces a fiscal limit, which is the point at which the adjustments in primary surpluses needed to stabilize debt are not assured. Uncertainty about future fiscal adjustments can untether fiscal expectations, making it difficult or impossible for monetary policy to achieve its objectives (Davig et al., 2010, 2011).
10. Exacerbating the fiscal fallout from the crisis, aging populations worldwide create long-run fiscal stress whose resolution in most countries is uncertain. This kind of uncertainty operates at low frequencies and may conflict with the long-run objectives of monetary policy (Carvalho and Ferrero, 2014).

It is hard to think about these developments without bringing monetary and fiscal policy *jointly* into the analysis. Several of these examples also run counter to critical maintained assumptions in monetarist/Wicksellian perspectives, including:

- fiscal policies will adjust government revenues and expenditures as needed to finance and stabilize government debt; this ensures that fiscal actions are "self-correcting" and need not concern monetary policymakers;
- sufficiently creative monetary policies—which include interest rate settings, quantitative easing, credit easing, government debt management, forward guidance—can always achieve desired inflation and macroeconomic objectives;
- impacts of monetary policy on fiscal choices are small enough to be of negligible importance to monetary policy decisions, freeing central banks to focus on a narrow set of goals.

As even this handful of examples makes clear, it is unlikely to be fruitful to interpret recent macroeconomic policy issues by studying monetary or fiscal policy in isolation. This chapter takes that premise as given to explore how macro policies interact to determine aggregate prices and quantities.

1.2 Our Remit

We were invited to write a chapter on the "fiscal theory of the price level," an assignment that we gladly accepted, but chose to broaden to the theory of price-level determination. A broader perspective, like the observations earlier, brings monetary and fiscal policy jointly into the picture to produce a more general understanding of the inflation process than either the monetarist/Wicksellian or the fiscal theory alone provide. We show that only in very special circumstances can the two perspectives be treated as distinct theories. Despite this broader perspective, both to fulfill our remit and to draw attention to aspects of monetary and fiscal policy interaction that are often overlooked, the chapter will often (but not solely) focus on the mechanisms that the fiscal theory emphasizes.

1.3 What Is the Fiscal Theory?

We consider a class of dynamically efficient models with monetary policy, a maturity structure for nominal government debt, taxes—distorting or lump-sum—government expenditures—purchases or transfers—and a government budget identity. In models of this kind, four key features of equilibrium may emerge:

1. There is a prominent role for nominal government debt revaluations that stabilize debt through surprise changes in inflation and bond prices.
2. It is possible for monetary–fiscal policy mixes to permit nominal government debt expansions or increases in the monetary policy interest rate instrument to increase nominal private wealth, nominal aggregate demand, and the price level.
3. Expectations of fiscal policy are equally important to those of monetary policy in determining prices and, sometimes, quantities, as in Brunner and Meltzer (1972), Tobin (1980), and Wallace (1981).[b]
4. Debt management policies matter for equilibrium dynamics, contributing an additional instrument to the standard macroeconomic policy toolkit, as Tobin (1963) argued.

Analyses of the implications of these features in this class of models constitute what we call the "fiscal theory of the price level."[c]

[b] Brunner and Meltzer anticipate the fiscal theory by showing that a government debt expansion unaccompanied by higher base money is inflationary when the fiscal deficit is held constant. But they dismiss this result on the grounds that "Price-level changes of this kind have not been important [foonote 13]."

[c] Early contributors to the theory include Begg and Haque (1984), Auernheimer and Contreras (1990), Leeper (1991), Sims (1994), Woodford (1995), and Cochrane (1999).

The fiscal theory is a complement to, rather than a substitute for, conventional views of price-level determination. It emerges by filling in the fiscal sides of models and broadening the rules that monetary and fiscal authorities can obey. By doing so, the fiscal theory extracts what assumptions about fiscal behavior are required to deliver conventional views. More importantly, being explicit about both monetary and fiscal behavior reveals that a far richer set of equilibria can arise from the previously suppressed, but undeniable, fact that monetary and fiscal policies are intrinsically intertwined.

The chapter aims to be constructive and instructive, so it does not refight the battles that surround the fiscal theory. Accusations against the fiscal theory include: it confuses equilibrium conditions with budget constraints; it violates Walras' law; it treats private agents and the government differently; it is merely an equilibrium selection device; it is little more than a retread of Sargent and Wallace's (1981) unpleasant monetarist arithmetic.[d] Each of these arguments has been discussed at length in Sims (1999a), Cochrane (2005), and Leeper and Walker (2013). Rehashing those debates detracts from the chapter's aims.

Cochrane (2011b, 2014) and Sims (1999b, 2013), two leading proponents of the fiscal theory, explore a wide range of issues through the lens of the fiscal theory to reach conclusions that contrast sharply with conventional perspective. This chapter also reexamines some practical issues in the light of the fiscal theory.

Most of the chapter focuses on the nature of equilibrium, including price-level determination, in models with nontrivial specifications of monetary and fiscal policy behavior. In this sense, the chapter, like the fiscal theory itself, echoes Wallace's (1981) insight that the effects of central bank open-market operations hinge on the precise sense in which fiscal policy is held constant. Under some assumptions on fiscal behavior, open-market operations are neutral, but different fiscal behavior permits monetary policy actions to have different impacts. Wallace did not explore the nature of price-level determination in the presence of nominal government bonds, which the fiscal theory emphasizes, but his results nonetheless foreshadow the newer literature. We also examine interactions in the opposite direction: how monetary policy behavior can influence the impacts of fiscal actions.

1.3.1 Real vs Nominal Government Debt

Central to the fiscal theory is the distinction between real and nominal government debt. This distinction matters little in conventional views that maintain that future revenues and expenditures always adjust to stabilize government debt. But the presence—in fact, the prevalence, of nominal government debt in many countries—lies at the core of the fiscal theory.[e]

[d] These accusations appear in Kocherlakota and Phelan (1999), McCallum (2001), Bassetto (2002), Buiter (2002), and Ljungqvist and Sargent (2004).

[e] See Cochrane (2011b) and Sims (2013).

Real debt can take the form of inflation-indexed bonds or bonds denominated in units whose supply the country does not control. Real debt is a claim to real goods, which the government must acquire through taxation. This imposes a budget constraint that the government's choices must satisfy. If the government does not have the taxing capacity to acquire the goods necessary to finance outstanding debt, it has no option other than outright default. Under the gold standard with fixed parities, countries effectively issued real debt because the real value of government bonds was determined by factors outside their control—worldwide supply and demand for gold.

Nominal debt is much like government-issued money: it is merely a claim to fresh currency in the future. The government may choose to raise taxes to acquire the requisite currency or it may opt to print up new currency, if currency creation is within its purview. Because the value of nominal debt depends on the price level and bond prices, the government really does not face a budget constraint when all its debt is nominal. Some readers may object to the idea that a government does not face a budget constraint, but the logic here is exactly the logic that underlies fiat currency. By conventional quantity theory reasoning, the central bank is free to double or half the money supply without fear of violating a budget constraint because the price level will double or half to maintain the real value of money. The direct analog to this reasoning is that the government is free to issue any quantity of nominal bonds, whose real value adjusts with the price level, without reference to a budget constraint. Of course, as with a money rain, by doing so the government is giving up control of the price level.

Member nations of the European Monetary Union issue debt denominated in euros, their home currency, but because monetary policy is under the control of the ECB rather than individual nations, the debt is effectively real from the perspective of member nations. The United States issues indexed debt, but it comprises only 10% of the debt outstanding. Even in the United Kingdom, which is known for having a thick market in indexed bonds, the percentage is only about 20. Five percent or less of total debt issued is indexed in the Euro Area, Japan, Australia, and Sweden.

1.3.2 Themes of the Chapter

Several themes run through this paper. First, it is always the *joint* behavior of monetary and fiscal policies that determine inflation and stabilize debt. While this point might seem obvious—echoing, as it does, a viewpoint that dates back at least to Friedman (1948)—it is easily missed in the classes of models and descriptions of policy typically employed in modern macroeconomic policy analyses. In those models, inflation appears to be determined entirely by monetary policy behavior—specifically, by the responsiveness of monetary policy to inflation—while debt dynamics seem to be driven only by fiscal behavior—the strength of primary surplus responses to debt. Of course, *in equilibrium* the two policies must interact in particular ways to deliver a determinate equilibrium with bounded debt, but this point is often swept under the carpet in order to focus the analysis solely on monetary policy.[f]

In dynamic models, macroeconomic policies have two fundamental tasks to achieve: determine the price level and stabilize debt. Two distinct monetary–fiscal policy mixes can accomplish those tasks. A second theme is that it is useful for some purposes to categorize those policy mixes in terms of "active" or "passive" policy behavior.[g] An active authority pursues its objectives unconstrained by the state of government debt and is free to set its control variables as it sees fit. But then the other authority must behave passively to stabilize debt, constrained by the active authority's actions and private-sector behavior. A determinate bounded equilibrium requires the mix of one active and one passive policy; that mix achieves the two macroeconomic objectives of delivering unique inflation and stable debt processes.[h] The combination of active monetary and passive fiscal policies delivers the usual monetarist/new Keynesian setup in which monetary policy can target inflation and fiscal policy exhibits Ricardian equivalence. We call this policy mix regime M, but it also goes by the label "monetary dominance." An alternative combination of passive monetary and active fiscal policies gives fiscal policy important effects on inflation, while monetary policy ensures that debt is stable. The latter policy regime has been given the unfortunate label "the fiscal theory of the price level." The fiscal theory mix is called regime F or "fiscal dominance."

Third, regime F policies produce equilibria in which the maturity structure of government debt affects equilibrium dynamics, as Cochrane (2001) and Sims (2011) emphasize. In contrast, without frictions that make short and long debt imperfect substitutes and in the special case of flexible prices and lump-sum taxes, maturity structure is irrelevant in regime M. Under the fiscal theory, long debt permits both current and future inflation (bond prices) to adjust to shocks that perturb the market value of debt, which serves to make inflation and, if prices are sticky, real activity less volatile than they would be if all debt were one period.

Fourth, only in the special cases of flexible prices and lump-sum fiscal shocks/surplus adjustments can simple active monetary policy rules hit their inflation target in regime M. More generally, with sticky prices and distortionary taxation, we observe revaluation effects and pervasive interactions between monetary and fiscal policy across both the M and F regimes.

Fifth, the "active/passive" rubrics also lose their usefulness once one considers optimal policies. Jointly optimal monetary and fiscal policies generally combine elements of

[f] See, for example, Woodford (2003) and Galí (2008).

[g] Leeper (1991) develops this categorization to study bounded equilibria.

[h] There are unbounded equilibria also. Sims (2013) and Cochrane (2011a) emphasize the possibility of solutions with unbounded inflation; McCallum (1984) and Canzoneri et al. (2001b) display solutions with unbounded debt that hinge on the presence of nondistorting taxes.

both regimes M and F: when long-maturity government debt is outstanding, it is always optimal to stabilize debt partly through distorting taxes and partly through surprise changes in inflation and bond prices (Cochrane, 2001; Leeper and Zhou, 2013; Sims, 2013). How important inflation is as a debt stabilizer—or in Sims' (2013) terminology, a "fiscal cushion"—depends on model specifics: the maturity structure of debt, the costliness of inflation variability, the level of outstanding government debt, whether optimal policy is with commitment or discretion, proximity of the economy to its fiscal limit, and so forth.

The fact that key features of the fiscal theory emerge as jointly optimal monetary and fiscal policy elevates the theory from a theoretical oddity to an integral part of macroeconomic policies that deliver desirable outcomes.

1.4 Overview of the Chapter

As we progress through the chapter we gradually widen the extent of monetary and fiscal policy interactions. We start with a simple flexible-price endowment economy subject to shocks to lump-sum transfers. This environment limits the extent of monetary and fiscal interactions to the revaluation effects emphasized by the fiscal theory and supports the strong dichotomy between the M and F regimes. Even in this simple environment, though, there are important spillovers between monetary and fiscal policy under either regime when we allow for either government spending or monetary policy shocks.

We then turn to consider the same rules in a production economy subject to nominal rigidities, but where we retain the assumption that taxes are lump sum. This adds a new channel for monetary and fiscal interactions because monetary policy can affect real interest rates when prices are sticky which, in turn, influence debt dynamics through real debt service costs. We then generalize this further by adding distortionary taxation to a new Keynesian economy. Then tax policy affects inflation through its impact on marginal costs, government spending feeds into aggregate demand, and monetary policy affects real interest rates to influence the size of the tax base. In this richer specification, equilibrium outcomes are always the result of interactions between monetary and fiscal policy and a key issue is the balance between monetary and fiscal policy in the control of inflation and stabilization of debt. We show that the conventional policy assignment of delegating monetary policy to achieve an inflation target and fiscal policy to stabilize debt is not always optimal.

Most expositions of the fiscal theory posit simple ad hoc rules for monetary and fiscal behavior and characterize the nature of equilibria under alternative settings of those rules. This chapter follows that path in the next two sections to derive clean analytical results that explain how the fiscal theory operates and how it differs from alternative policy mixes. Then the paper turns to study jointly optimal monetary and fiscal policies as an alternative vehicle for describing the economic mechanisms that underlie the fiscal theory. Optimal policies make clear that the distinguishing features of the fiscal theory are generally part of a policy mix that produces desirable economic outcome. But the incentive to use surprise inflation to stabilize debt, especially when debt levels are high, can also create significant time-consistency issues when policymakers cannot credibly commit. When private agents know that policymakers may be tempted to induce inflation surprises to reduce the debt burden, economic agents raise their inflation expectations as debt levels rise until that temptation has been offset. This produces a sizeable debt stabilization bias that drives policymakers to reduce debt levels rapidly, at large cost in terms of social welfare, to avoid the high equilibrium rates of inflation associated with the temptation to inflate that debt away. We explore the sharp contrast between time-consistent and time-inconsistent optimal policy in this context in detail.

After those purely theoretical explorations, the paper turns to consider the empirical relevance of those mechanisms. We describe some subtle issues that arise in efforts to identify monetary–fiscal regime and review existing evidence both for and against fiscal interpretations of time series. The chapter then discusses three practical applications of the theory: fiscal prerequisites for successful inflation targeting, consequences of alternative fiscal reactions to a return to more normal levels of interest rates, and why the central bank needs understand the prevailing monetary–fiscal regime in order to conduct monetary policy. To wrap up, we describe outstanding issues in both theoretical and empirical analyses of monetary and fiscal policy interactions to point out directions for future research.

2. ENDOWMENT ECONOMIES WITH AD HOC POLICY RULES

This section aims to present the distinguishing features of the fiscal theory listed in Section 1.3 in the simplest possible model. A representative consumer lives forever and receives a constant endowment of goods, y, each period. The economy is cashless and financial markets are complete.

2.1 A Simple Model

The consumer optimally chooses consumption, c_t, may buy or sell nominal assets, D_t, at price $Q_{t,t+1}$, receives lump-sum transfers from the government, z_t, and pays lump-sum taxes, τ_t.[i] The representative household maximizes

$$E_0\left\{\sum_{t=0}^{\infty}\beta^t U(c_t)\right\}$$

with $0<\beta<1$, subject to the sequence of flow budget constraints

[i] D_t consists of privately issued, B_t^p, and government issued, B_t, assets. Government bonds cost $\$1/R_t$ per unit and are perfectly safe pure discount bonds.

$$P_t c_t + P_t \tau_t + E_t[Q_{t,t+1} D_t] = P_t y + P_t z_t + D_{t-1} \tag{1}$$

given D_{-1}. $Q_{t,t+1}$ is the nominal price at t of an asset that pays \$1 in period $t+1$ and P_t is the general price level in units of mature government bonds required to purchase one unit of goods. Government bonds sold at t, which are included in D_t, pay gross nominal interest R_t in period $t+1$. Letting $m_{t,t+1}$ denote the real contingent claims price, a no-arbitrage condition implies that

$$Q_{t,t+1} = m_{t,t+1} \frac{P_t}{P_{t+1}} \tag{2}$$

The short-term nominal interest rate, R_t, which is also the central bank's policy instrument, is linked to the nominal bond price: $1/R_t = E_t[Q_{t,t+1}]$.

Setting government purchases of goods to zero,[j] the primary surplus is simply $s_t \equiv \tau_t - z_t$. The household's intertemporal budget identity comes from iterating on (1) and imposing the no-arbitrage condition, (2), and the transversality condition

$$\lim_{T\to\infty} E_t\left[m_{t,T} \frac{D_{T-1}}{P_T}\right] = 0 \tag{3}$$

to yield

$$E_t \sum_{j=0}^{\infty} m_{t,t+j} c_{t+j} = \frac{D_{t-1}}{P_t} + E_t \sum_{j=0}^{\infty} m_{t,t+j}(y - s_{t+j}) \tag{4}$$

where $m_{t,t+j} \equiv \prod_{k=0}^{j} m_{t+k,t+k+1}$ is the real discount factor, with $m_{t,t} = 1$.

After imposing equilibrium in the goods market, $c_t = y$, the real discount factor is constant, $m_{t,t+1} = \beta$, and the nominal interest rate obeys a Fisher relation

$$\frac{1}{R_t} = \beta E_t \frac{P_t}{P_{t+1}} = \beta E_t \frac{1}{\pi_{t+1}} \tag{5}$$

where $\pi_t \equiv P_t/P_{t-1}$ is the gross inflation rate. In equilibrium there will be no borrowing or lending among private agents, so the household's bond portfolio consists entirely of government bonds. Imposing both bond and goods market clearing and the constant real discount factor the household's intertemporal constraint produces the ubiquitous equilibrium condition

$$\frac{B_{t-1}}{P_t} = E_t \sum_{j=0}^{\infty} \beta^j s_{t+j} \tag{6}$$

Cochrane (2001) refers to (6) as an "equilibrium valuation equation" because it links the market value of debt outstanding at the beginning of period t, B_{t-1}/P_t, to the expected

[j] We shall relax this assumption below.

present value of the cash flows that back debt, primary surpluses. Notice that we derived this valuation equation entirely from private optimizing behavior and market clearing, without reference to government behavior or to the government's budget identity. The valuation equation imposes no restrictions on the government's choices of future surpluses, in the same way that the Fisher relation does not limit the central bank's choices of the nominal interest rate.

For each date t, equations (5) and (6) constitute two equilibrium conditions in four unknowns: $R_t, P_t, E_t(1/P_{t+1}), E_t\sum_{j=0}^{\infty}\beta^j s_{t+j}$. Private-sector behavior alone cannot uniquely determine the equilibrium. We turn now to a class of monetary and fiscal policy rules that may deliver determinate equilibria.

2.1.1 Policy Rules

The central bank obeys a simple interest rate rule, come to be called a Taylor (1993) rule, that makes deviations of the nominal interest rate from steady state proportional to deviations of inflation from steady state

$$\frac{1}{R_t} = \frac{1}{R^*} + \alpha_\pi\left(\frac{1}{\pi_t} - \frac{1}{\pi^*}\right) + \varepsilon_t^M \tag{7}$$

where ε_t^M is an exogenous shock to monetary policy. The government sets deviations of the primary surplus from steady state proportional to steady-state deviations of debt

$$s_t = s^* + \gamma\left(\frac{1}{R_{t-1}}\frac{B_{t-1}}{P_{t-1}} - \frac{b^*}{R^*}\right) + \varepsilon_t^F \tag{8}$$

where ε_t^F is an exogenous fiscal shock to the primary surplus. The inverse of the nominal interest rate is the price of nominal debt, so $\frac{1}{R_{t-1}}\frac{B_{t-1}}{P_{t-1}}$ is the real market value of debt issued at $t-1$. Policy choices must be consistent with the government's flow budget identity

$$\frac{1}{R_t}\frac{B_t}{P_t} + s_t = \frac{B_{t-1}}{P_t}$$

where the steady state of the model is

$$\frac{B}{P} = b^*, \quad s^* = (\beta^{-1} - 1)\frac{b^*}{R^*}, \quad R^* = \frac{\pi^*}{\beta}, \quad m^* = \beta$$

It is convenient to express things in terms of the inverse of inflation (ie, deflation) and real debt, so let $\nu_t \equiv \pi_t^{-1}$ and $b_t \equiv B_t/P_t$. Combining the monetary policy rule with the Fisher equation yields the difference equation in deflation

$$E_t(\nu_{t+1} - \nu^*) = \frac{\alpha_\pi}{\beta}(\nu_t - \nu^*) + \frac{1}{\beta}\varepsilon_t^M \tag{9}$$

Combining the fiscal rule and the government's flow budget identity, taking expectations, and employing the Fisher relation yield real debt dynamics

$$E_t\left(\frac{b_{t+1}}{R_{t+1}}-\frac{b^*}{R^*}\right)=(\beta^{-1}-\gamma)\left(\frac{b_t}{R_t}-\frac{b^*}{R^*}\right)-E_t\varepsilon^F_{t+1} \tag{10}$$

Equations (9) and (10) constitute a system of expectational difference equations in inflation and real debt, which is driven by the exogenous policy disturbances ε^M and ε^F. Given the consumer's discount factor, β, this system appears as though inflation dynamics depend only on the monetary policy choice of α_π, while debt dynamics hinge only on the fiscal policy choice of γ: it is not obvious that monetary and fiscal behavior *jointly* determine inflation and real debt. This apparent separation of the system is deceptive. Because the government issues *nominal* bonds, B_t, the price level appears in both equations and $1/P_t$ is the value of bonds maturing at t.

2.1.2 Solving the Model

We focus on bounded solutions.[k] Stability of inflation depends on α_π/β and stability of debt depends on $\beta^{-1}-\gamma$.[l]

2.1.2.1 Regime M

If $\alpha_\pi/\beta > 1$, then the bounded solution for inflation is

$$\nu_t=\nu^*-\frac{1}{\alpha_\pi}\sum_{j=0}^{\infty}\left(\frac{\beta}{\alpha_\pi}\right)^j E_t\varepsilon^M_{t+j} \tag{11}$$

which delivers a solution for $\{P_{t-1}/P_t\}$ for $t\geq 0$ and the equilibrium nominal interest rate is

$$\frac{1}{R_t}=\frac{1}{R^*}-\sum_{j=1}^{\infty}\left(\frac{\beta}{\alpha_\pi}\right)^j E_t\varepsilon^M_{t+j}$$

In this simple model, both actual and expected inflation depend on the monetary policy parameter and shock, but they appear not to depend in any way on fiscal behavior.

[k] Unbounded solutions for inflation also exist, as Benhabib et al. (2001) show. Sims (1999b), Cochrane (2011a), and Del Negro and Sims (2015) thoroughly explore those equilibria to argue that a determinate price level requires appropriate fiscal backing. As Del Negro and Sims (2015, p. 3) define it: "Fiscal backing requires that explosive inflationary or deflationary behavior of the price level is seen as impossible because the fiscal authority will respond to very high inflation with higher primary surpluses and to near-zero interest rates with lower, or negative, primary surpluses." Solutions with unbounded debt inevitably rely on nondistorting taxes, which permit revenues to grow forever at the same rate as interest receipts on government bond holdings. Although such paths for revenues are equilibria in the present model, because they are infeasible in economies where taxes distort, we find them to be uninteresting.

[l] We consider the implications of temporarily being in active–active or passive–passive regimes in Section 7.3.

This appearance is deceptive because (11) does not constitute a complete solution to the model; we also need to ensure that there is a bounded solution for real debt. If fiscal policy chooses $\gamma > \beta^{-1} - 1$, then when real debt rises, future surpluses rise by more than the net real interest rate with the change in debt in order to cover both debt service and a little of the principal. In this case, the debt dynamics in (10) imply that for arbitrary deviations of real debt from steady state, $\lim_{T \to \infty} E_t b_{T+1} = b^*$, so debt eventually returns to steady state.

Digging into exactly what fiscal policy does to stabilize debt reveals the underlying policy interactions. Suppose that at time t news arrives of a higher path for $\{\varepsilon^M_{t+j}\}$. This news reduces ν_t, raising the price level P_t. With fiscal rule (8), in the first instance the monetary news leaves s_t unaffected, but household holdings of outstanding bonds, B_{t-1}/P_t, decline. From the government budget identity, this implies that the market value of debt issued at t also falls, even if there is no change in the price of bonds, $1/R_t$

$$\frac{B_t}{P_t R_t} = -s_t + \frac{B_{t-1}}{P_t}$$

In the absence of future fiscal adjustments—such as those in which $\gamma > \beta^{-1} - 1$—household wealth would decline, reducing aggregate demand and counteracting the inflationary effect of the monetary expansion. But when fiscal policy reduces surpluses with debt by more than the real interest rate, surpluses are expected to fall by an amount equal in present value to the initial drop in the value of household bond holdings. This eliminates the negative wealth effect to render monetary policy expansionary.

When the news of higher $\{\varepsilon^M_{t+j}\}$ extends to affect the equilibrium beyond the current period, the nominal interest rate rises, reducing the price of new bonds at t. Lower bond prices implicitly raise interest yields on these bonds that mature in period $t+1$ to create a second channel by which monetary policy affects household wealth. As with the first channel, though, these wealth effects evaporate with the expected adjustments in surpluses.

These fiscal adjustments connect to Wallace's (1981) point that the impacts of open-market operations hinge on the sense in which fiscal policy is "held constant." In regime M, the "constancy" of fiscal policy is quite specific: it eliminates any monetary effects on balance sheets. By neutralizing the fiscal consequences of monetary policy actions, this regime leaves the impression that, in Friedman's (1970) famous aphorism, "inflation is always and everywhere a monetary phenomenon." Of course, it is the *joint* behavior of monetary and fiscal policies that delivers this impression.

Regime M also delivers the fiscal counterpart to Friedman's monetarist adage: Ricardian equivalence.[m] A fiscal shock at t that reduces the surplus by one unit is financed

[m] Tobin (1980, p. 53) made this point: "Thus the Ricardian equivalence theorem is fundamental, perhaps indispensable, to monetarism."

initially by an expansion in nominal debt of P_t units. With inflation pinned down by expression (11), real debt also increases by P_t units. Higher real debt, through the fiscal rule, triggers higher future surpluses whose present value equals the original debt expansion. Even in this completely standard Ricardian experiment, it is the joint policy behavior—monetary policy's aggressive response to inflation and fiscal policy's passive adjustment of surpluses—that produces the irrelevance result.

2.1.2.2 Regime F

Consider the case in which fiscal policy is active, with exogenous surpluses, so $\gamma = 0$ to make the fiscal rule is $s_t = s^* + \varepsilon_t^F$. The solution for real debt is[n]

$$\frac{b_t}{R_t} = \frac{b^*}{R^*} + \sum_{j=1}^{\infty} \beta^j E_t \varepsilon_{t+j}^F \tag{12}$$

which implies that the value of debt at t depends on the expected present value of surpluses from $t + 1$ onward.

We can solve for inflation by combining this solution for b_t with the government's flow budget identity, noting that $B_{t-1}/P_t = \nu_t b_{t-1}$

$$\nu_t = \frac{(1-\beta)^{-1} s^* + \sum_{j=0}^{\infty} \beta^j E_t \varepsilon_{t+j}^F}{b_{t-1}} \tag{13}$$

where at t, b_{t-1} is predetermined, which produces the solution for the price level

$$P_t = \frac{B_{t-1}}{(1-\beta)^{-1} s^* + \sum_{j=0}^{\infty} \beta^j E_t \varepsilon_{t+j}^F} \tag{14}$$

News of lower surpluses raises the price level and reduces the value of outstanding debt. In contrast to regime M equilibria, in regime F *nominal* government debt is an important state variable.[o] Higher nominal debt or higher debt service raises the price level next period. These results reflect the impacts of higher nominal household wealth. Lower future surpluses—stemming from either lower taxes or higher transfers—or higher initial nominal assets raise households' demand for goods when there is no prospect that future taxes will rise to offset the higher wealth. Unlike regime M, now equilibrium inflation, as given by (13), depends explicitly on current and expected fiscal choices—through the steady-state surplus, s^*, and fiscal disturbances, $\sum_{j=0}^{\infty} \beta^j E_t \varepsilon_{t+j}^F$.

[n] To derive (12), define $\tilde{b}_t \equiv B_t / P_t R_t$ to write the flow government budget identity as $\tilde{b}_t + s_t = R_{t-1} \nu_t \tilde{b}_{t-1}$. Take expectations at $t - 1$, apply the Euler equation $\beta^{-1} = E_{t-1} R_{t-1} \nu_t$, iterate forward, and impose transversality to obtain (12).

[o] Debt is also a state variable in regime M because it contains information about future surpluses. But in M, changes in the *real* value of debt induce changes in expectations of future *real* government claims on private resources.

Expression (12) gives the real market value of debt. But in the absence of any stabilizing response of surpluses to real debt ($\gamma = 0$), debt's deviations from steady state are expected to grow over time at the real rate of interest, $1/\beta$, according to (10). Such growth in debt would violate the household's transversality condition, which is inconsistent with equilibrium. To reconcile these seemingly contradictory implications of the equilibrium, we need to understand the role that monetary policy plays in regime F.

Monetary policy ensures that actual debt, as opposed to expected debt, is stable by preventing interest payments on the debt from exploding and permitting surprise inflation to revalue government debt. In regime F, higher interest payments raise nominal wealth, increasing nominal aggregate demand, and future inflation, as both (13) and (14) indicate. To understand monetary policy behavior, substitute the solution for ν_t from (13) into the monetary policy rule, (7). To simplify the expression, assume that the policy shocks are *i.i.d.* so that

$$\frac{1}{R_t} - \frac{1}{R^*} = \frac{\alpha_\pi}{\beta}\left[\frac{\beta(1-\beta)^{-1}s^* + \beta\varepsilon_t^F}{b_{t-1}} - \frac{1}{R^*}\right] + \varepsilon_t^M \tag{15}$$

In response to a fiscal expansion—$\varepsilon_t^F < 0$—the central bank reduces $1/R_t$ by $\alpha_\pi \varepsilon_t^F$ to lean against the fiscally induced inflation. A serially uncorrelated fiscal disturbance leaves the market value of debt at its steady state, $b_{t+j}/R_{t+j} = b^*/R^*$ for $j \geq 0$. This greatly simplifies the time $t + 1$ version of (15) to yield

$$\frac{1}{R_{t+1}} - \frac{1}{R^*} = \frac{\alpha_\pi}{\beta}\left(\frac{1}{R_t} - \frac{1}{R^*}\right) \tag{16}$$

If monetary policy were to respond aggressively to inflation by setting $\alpha_\pi/\beta > 1$, $1/R$ would diverge to positive or negative infinity, both situations that violate lower bound conditions on the net, $R - 1$, nominal interest rate. Economically, these exploding paths stem from strong wealth effects that arise from ever-growing interest receipts to holders of government bonds. When $\alpha_\pi/\beta > 1$ the central bank raises the nominal interest rate by a factor that exceeds the real interest rate, which increases private agents' nominal wealth and inflation in the next period; this process repeats in subsequent periods. Active monetary policy essentially converts stable fiscally induced inflation into explosive paths.

Existence of equilibrium requires that the monetary reaction to inflation not be too strong—specifically, that $\alpha_\pi/\beta < 1$, what is called "passive monetary policy." A pegged nominal interest rate, $\alpha_\pi = 0$, is the easiest case to understand. By holding the nominal rate fixed at R^*, monetary policy prevents the fiscal expansion from affecting future inflation by fixing interest payments on the debt. A one-time reduction in s_t that is financed by new nominal bond sales raises P_t enough to keep B_t/P_t unchanged. But the higher price level also reduces the real value of existing nominal debt, B_{t-1}/P_t, and in doing so reduces the implicit real interest payments. In terms of the flow budget identity

$$\frac{b^*}{R^*} + s_t = \frac{B_{t-1}}{P_t}$$

where real debt remains at steady state because $\gamma = 0$ implies that expected surpluses are unchanged. The larger is the stock of outstanding debt, the less the price level must rise to keep the budget in balance.

More interesting results emerge when there is some monetary policy response to inflation—$0 < \alpha_\pi < \beta$.[p] When monetary policy tries to combat fiscal inflation by raising the nominal interest rate, inflation is both amplified and propagated. Pegging R_t forces all inflation from a fiscal shock to occur at the time of the shock. Raising R_t permits the inflation to persist and the more strongly monetary policy reacts to inflation, the longer the inflation lasts.

Difference equations (15) and (16) make the monetary policy impacts clear. When $\alpha_\pi = 0$, a shock to ε_t^F has no effect on the nominal interest rate. But the larger is α_π, though still less than β, the stronger are the effects of ε_t^F on future nominal interest rates and, through the Fisher relation, future inflation.

Even though the transitory fiscal expansion has no effect on real debt, higher nominal rates bring forth new nominal bond issuances that are proportional to the increases in the price level. Higher nominal debt coupled with higher interest on the debt increases interest payments that raise household nominal wealth in the future. Because future taxes do not rise to offset that wealth increase, aggregate demand and the price level rise in the future.

Expression (15) reveals that an exogenous monetary contraction—lower ε_t^M that raises R_t—triggers exactly the same macroeconomic effects as an exogenous fiscal expansion. Higher interest rates raise debt service and nominal wealth, which increases inflation in the future. In this simple model with a fixed real interest rate, only this perverse implication for monetary policy obtains. We shall discuss the effects of monetary policy contractions in a production economy with longer maturity debt in Section 2.2.[q]

2.2 The Role of Maturity Structure

Tobin (1963) discusses debt management in the context of the "monetary effect of the debt," contrasting this to the "direct fiscal effect" that is determined by the initial increase in the bond-financed deficit. The monetary effect stems from the maturity structure of the debt, which Tobin reasons outlasts the direct effect because it endures over the maturity horizon of the debt. Changes in the maturity composition of debt operate through

[p] Impulse responses to this case are considered in Section 2.3.

[q] The result that a monetary contraction raises future inflation is reminiscent of Sargent and Wallace's (1981) unpleasant monetarist arithmetic, but the mechanism is completely different. In Sargent and Wallace, tighter money today implies looser money in the future and the higher future inflation can feed back to reduce money demand today. Their result does not stem from wealth effects of monetary policy.

impacts on the size and composition of private wealth. Such changes can affect the macro economy, even if they do not entail changing the overall size of the debt. This section obtains closely related impacts from maturity structure in regime F.

The section introduces a full maturity structure of government debt in general form to derive the bond valuation equation and develop some intuition about the role that maturity plays in the endowment economy in regime F. It then uses a simple special case to make transparent the mechanisms at work in regime F.[r]

2.2.1 A General Maturity Structure

Let $B_t(t+j)$ denote the nominal quantity of zero-coupon bonds outstanding in period t that matures in period $t+j$ and let the dollar-price of those bonds be $Q_t(t+j)$. The government's flow budget identity at t is

$$B_{t-1}(t) - \sum_{j=1}^{\infty} Q_t(t+j)[B_t(t+j) - B_{t-1}(t+j)] = P_t s_t$$

In a constant-endowment economy, the bond-pricing equations are

$$Q_t(t+k) = \beta^k E_t \frac{P_t}{P_{t+k}} \tag{17}$$

for $k = 1, 2, \ldots$. These pricing equations imply the no-arbitrage condition that links the price of a k-period bond to the expected sequence of k 1-period bonds

$$Q_t(t+k) = E_t[Q_t(t+1)Q_{t+1}(t+2)\ldots Q_{t+k-1}(t+k)]$$

To derive the bond valuation equation with a general maturity structure, define

$$B_{t-1} \equiv B_{t-1}(t) + \sum_{j=1}^{\infty} Q_t(t+j)B_{t-1}(t+j)$$

as the portfolio of bonds outstanding at the end of period $t-1$ and rewrite the government budget identity as

$$\frac{B_{t-1}}{P_t} = Q_t(t+1)\frac{B_t}{P_t} + s_t$$

Iterating on this bond portfolio version of the constraint, taking expectations and imposing the bond-pricing relations and the consumer's transversality condition yields the valuation equation

$$\frac{B_{t-1}}{P_t} = \sum_{j=0}^{\infty} \beta^j E_t s_{t+j}$$

[r] These derivations draw on Cochrane (2001, 2014).

or, in terms of the underlying bonds

$$\frac{B_{t-1}(t)}{P_t} + \sum_{j=1}^{\infty} \beta^j E_t \frac{B_{t-1}(t+j)}{P_{t+j}} = \sum_{j=0}^{\infty} \beta^j E_t s_{t+j} \tag{18}$$

Use (18) to repeatedly substitute out future price levels to make explicit how maturity structure enters the valuation equation

$$\begin{aligned} \frac{B_{t-1}(t)}{P_t} = E_t \Bigg\{ & s_t + \beta \underbrace{\left[1 - \frac{B_{t-1}(t+1)}{B_t(t+1)}\right]}_{\text{weight on } t+1} s_{t+1} \\ & + \beta^2 \underbrace{\left\{1 - \left[\frac{B_{t-1}(t+2)}{B_{t+1}(t+2)} \frac{B_{t-1}(t+1)}{B_t(t+1)} \left(1 - \frac{B_t(t+2)}{B_{t+1}(t+2)}\right)\right]\right\}}_{\text{weight on } t+2} s_{t+2} + \ldots \Bigg\} \end{aligned} \tag{19}$$

We write this valuation equation more compactly by defining

$$\Lambda_t(t+k) \equiv \frac{B_t(t+k) - B_{t-1}(t+k)}{B_{t+k-1}(t+k)}$$

as newly issued debt that matures in period $t + k$ as a share of total outstanding debt in period $t + k - 1$ that matures at $t + k$. We can now define the maturity weight on the surplus at $t + k$, $L_{t,t+k}$, as depending recursively on these ratios

$$\begin{aligned} L_{t,t} &= 1 \\ L_{t,t+1} &= \Lambda_t(t+1) \\ L_{t,t+2} &= \Lambda_{t+1}(t+2) L_{t,t+1} + \Lambda_t(t+2) \\ L_{t,t+3} &= \Lambda_{t+2}(t+3) L_{t,t+2} + \Lambda_{t+1}(t+3) L_{t,t+1} + \Lambda_t(t+3) \\ &\vdots \\ L_{t,t+k} &= \sum_{j=0}^{k-1} \Lambda_{t+j}(t+k) L_{t,t+j} \end{aligned}$$

The compact form of valuation equation (19) is now

$$\frac{B_{t-1}(t)}{P_t} = \sum_{j=0}^{\infty} \beta^j E_t [L_{t,t+j} s_{t+j}] \tag{20}$$

Given a sequence of surpluses, $\{s_t\}$, discount factors and maturity determine the expected present value of surpluses. Shortening maturity (eg, reducing $\frac{B_{t-1}(t+1)}{B_t(t+1)}$) raises the weights on s_{t+1}, s_{t+2}, s_{t+3}, raising that present value—the backing of debt—and the value of debt. Shortening maturity of bonds due at $t+k$ raises weights on all $s_{t+j}, j \geq k$. In this sense, shortening maturity can offset a decline in surpluses.

Surprise changes in future maturity structure appear as innovations in the weights, $L_{t,t+j}$, in valuation equation (20). If primary surpluses are given, an unanticipated shortening of maturity of bonds held by the public would, by raising the value of outstanding debt, reduce the current price level. Viewed through the lens of the fiscal theory, the Federal Reserve's "operation twist" in 2011 would have a contractionary effect on the economy initially.[s] As the example to which we now turn illustrates, the lower price level at t would ultimately be offset by a higher future price level.

2.2.1.1 An Illustrative Example

To cleanly illustrate the role that changes in maturity structure play in determining the timing of inflation, we examine an example from Cochrane (2014). We use the same constant-endowment economy, but it operates only in periods $t=0, 1, 2$, and then ends; we set the real interest rate to zero, so the discount factor is $\beta=1$. The government issues one- and two-period nominal bonds at the beginning of time, $t=0$, denoted by $B_0(1)$ and $B_0(2)$, and uses surpluses in periods 1 and 2, s_1 and s_2, to retire the debt. At date $t=1$ the government may choose to issue new one-period debt, $B_1(2)$, so the change in debt at $t=1$ is $B_1(2)-B_0(2)$. The three potentially different quantities of bonds sell at nominal prices $Q_0(1), Q_0(2), Q_1(2)$ that obey (17) with $\beta=1$.[t]

Given initial choices of debt, $B_0(1)$ and $B_0(2)$, the government's budget identities in periods 1 and 2 are

$$B_0(1) = P_1 s_1 + Q_1(2)[B_1(2) - B_0(2)] \tag{21}$$

$$B_1(2) \; = P_2 s_2 \tag{22}$$

When primary surpluses are given at $\{s_1, s_2\}$, expression (22) immediately yields the price level in period 2 as

$$\frac{B_1(2)}{P_2} = s_2$$

because $B_1(2)$ is predetermined in period 2.

[s] The premise of the Fed's actions was that if short and long bonds are imperfect substitutes, then increasing demand for long bonds would reduce long-term interest rates. Lower long rates, it was hoped, would stimulate business investment and the housing market.

[t] We normalize the initial price level to be $P_0 = 1$.

Now impose the asset-pricing relations on the bond prices in the period 1 government budget identity, (21), to obtain the bond valuation equation

$$\frac{B_0(1)}{P_1} = s_1 + \left[\frac{B_1(2) - B_0(2)}{B_1(2)}\right] E_1 s_2$$

P_1 depends on the choice of newly issued bonds in period 1.

Solving for expected inflation and bond prices yields

$$E_0\left(\frac{1}{P_2}\right) = Q_0(2) = E_0\left(\frac{s_2}{B_1(2)}\right) = E_0\left[\frac{1}{B_0(2) + (B_1(2) - B_0(2))}\right] s_2$$

$$E_0\left(\frac{1}{P_1}\right) = Q_0(1) = \frac{E_0[s_1]}{B_0(1)} + \frac{1}{B_0(1)} E_0\left[\frac{B_1(2) - B_0(2)}{B_1(2)}\right] s_2$$

So the term structure of interest rates also depends on choices about maturity structure.

We can derive explicit solutions for the actual or realized price level at $t = 1$ in terms of innovations

$$B_0(1)(E_1 - E_0)\left(\frac{1}{P_1}\right) = (E_1 - E_0)s_1 + (E_1 - E_0)\left(\frac{B_1(2) - B_0(2)}{B_1(2)}\right) s_2$$

Surprise increases in the price level in period 1 depend negatively on innovations in time-1 and time-2 surpluses and on unexpected lengthening of the maturity of bonds due in period 2.

These derivations show that the government can achieve any path of the nominal term structure—and in this example, expected inflation—that it wishes by adjusting maturity structure. By unexpectedly selling less time-2 debt, the government reduces the claims to time-2 surpluses, which reduces the revenues that can be used to pay off period-1 bonds. This raises inflation in period 1. That increase in inflation comes from reducing $B_1(2)$, which lowers the price level in period 2, as seen from

$$(E_1 - E_0)\left(\frac{B_1(2)}{P_2}\right) = (E_1 - E_0)s_2$$

If s_2 is given, selling less $B_1(2)$ requires P_2 to fall.

2.2.2 A Useful Special Case

Suppose that the maturity structure declines at a constant rate $0 \leq \rho \leq 1$ each period so that the pattern of bonds issued at $t - 1$ obeys

$$B_{t-1}(t+j) = \rho^j B_{t-1}^m$$

where B_{t-1}^m is the portfolio of these specialized bonds in $t - 1$. When $\rho = 0$ all bonds are one period, whereas when $\rho = 1$ all bonds are consols. The average maturity of the portfolio is $1/(1 - \beta\rho)$.

With this specialization, the government's flow constraint is

$$B_{t-1}^m\left[1-\sum_{j=1}^{\infty}Q_t(t+j)\rho^j\right]=P_ts_t+B_t^m\sum_{j=1}^{\infty}Q_t(t+j)\rho^{j-1}$$

If we define the price of the bond portfolio as

$$P_t^m\equiv\sum_{j=1}^{\infty}Q_t(t+j)\rho^{j-1}$$

then the government's budget identity becomes

$$B_{t-1}^m(1+\rho P_t^m)=P_ts_t+P_t^mB_t^m \tag{23}$$

Bond portfolio prices obey the recursion

$$P_t^m=Q_t(t+1)[1+\rho E_tP_{t+1}^m]=R_t^{-1}[1+\rho E_tP_{t+1}^m] \tag{24}$$

This shows that a constant geometric decay rate in the maturity structure of zero-coupon bonds is equivalent to the interpretation of bonds that pay geometrically decaying coupon payments, as in Woodford (2001) and Eusepi and Preston (2013).

Let R_{t+1}^m denote the gross nominal return on the bond portfolio between t and $t+1$. Then $R_{t+1}^m=(1+\rho P_{t+1}^m)/P_t^m$ and the no-arbitrage condition implies that

$$\frac{1}{R_t}=\beta E_t\nu_{t+1}=E_t\left(\frac{1}{R_{t+1}^m}\right) \tag{25}$$

Combining (24) and (25) and iterating forward connects bond prices to expected paths of the short-term nominal interest rate and inflation

$$P_t^m=\sum_{j=0}^{\infty}\rho^jE_t\left(\prod_{i=0}^{j}R_{t+i}^{-1}\right)=\beta\sum_{j=0}^{\infty}(\beta\rho)^jE_t\left(\prod_{i=0}^{j}\nu_{t+i+1}\right) \tag{26}$$

2.3 Maturity Structure in Regime F

Ricardian equivalence in regime M makes the maturity structure of debt irrelevant for inflation, so in this section we focus solely on regime F. When surpluses are exogenous ($\gamma=0$), the debt valuation equation becomes[u]

[u] To derive (27), convert the nominal budget identity in (23) into a difference equation in the real value of debt, P^mB^m/P, impose pricing equations (24) and (25), using the fact that $\beta^{-1}=E_{t-1}[\nu_t(1+\rho P_t^m)/P_{t-1}^m]$, iterate forward, and impose the household's transversality condition for debt.

$$\frac{(1+\rho P_t^m)B_{t-1}^m}{P_t} = (1-\beta)^{-1} s^* + \sum_{j=0}^{\infty} \beta^j E_t \varepsilon_{t+j}^F \tag{27}$$

In contrast to the situation with only one-period debt ($\rho = 0$) when fiscal news appeared entirely in jumps in the price level, now there is an additional channel through which debt can be revalued: bond prices that reflect expected inflation over the entire duration of debt. News of lower future surpluses reduces the value of debt through both a higher P_t and a lower P_t^m. By (26), the lower bond price portends higher inflation and higher one-period nominal interest rates. The ultimate mix between current and future inflation is determined by the monetary policy rule. Long-term debt opens a new channel for monetary and fiscal policy to interact.

No-arbitrage condition (26) reveals a key aspect of regime F equilibria with long debt. With the simplified maturity structure, ρ determines the average maturity of the zero-coupon bond portfolio. A given future inflation rate has a larger impact on the price of bonds, the larger is ρ or the longer is the average maturity of debt. The maturity parameter serves as an additional discount factor, along with β, so more distant inflation rates have a smaller impact on bond prices than do rates in the near future. Of course, the date t expected present value of inflation influences only the price of bonds that are outstanding at the beginning of t, namely, B_{t-1}^m.

To understand monetary policy's influence on the timing of inflation, note that when monetary policy is passive, $\alpha_\pi/\beta < 1$, (9) implies that k-step-ahead expected inflation is

$$E_t \nu_{t+k} = \left(\frac{\alpha_\pi}{\beta}\right)^k (\nu_t - \nu^*) + \nu^*$$

which may be substituted into the pricing equation that links P_t^m to the term structure of inflation rates, (26), to yield[v]

$$\rho P_t^m = \sum_{j=1}^{\infty} (\beta\rho)^j \left\{ \prod_{i=0}^{j-1} \left[\left(\frac{\alpha_\pi}{\beta}\right)^{i+1} (\nu_t - \nu^*) + \nu^* \right] \right\}$$

Monetary policy's reaction to inflation—through α_π—interacts with the average maturity of debt—ρ—to determine how current inflation—ν_t, which is given by (13) in regime F—affects the price of bonds. More aggressive monetary policy and longer maturity debt both serve to amplify the impact of current inflation on bond prices, suggesting that higher α_π and higher ρ permit fiscal disturbances to have a smaller impact on current inflation at the cost of a larger impact on future inflation.

[v] Here we shut down the exogenous monetary policy shock, $\varepsilon_t^M \equiv 0$.

Consider two polar cases of passive monetary policy. When $\alpha_\pi = 0$, so the central bank pegs the nominal interest rate and bond prices at $\rho P_t^m = \beta\rho\nu^*/(1-\beta\rho\nu^*)$, the valuation expression becomes

$$\left(\frac{1}{1-\beta\rho\nu^*}\right)\nu_t b_{t-1}^m = (1-\beta)^{-1}s^* + \sum_{j=0}^{\infty}\beta^j E_t\varepsilon_{t+j}^F$$

where we define $b_{t-1}^m \equiv B_{t-1}^m/P_{t-1}$. In this case, expected inflation returns to target immediately, $E_t\nu_{t+j} = \nu^*$ for $j \geq 1$.

The second case is when monetary policy reacts as strongly as possible to inflation, while still remaining passive: $\alpha_\pi = \beta$.[w] Then $\rho P_t^m = \beta\rho\nu_t/(1-\beta\rho\nu_t)$ and the valuation equation is[x]

$$\left(\frac{\nu_t}{1-\beta\rho\nu_t}\right)b_{t-1}^m = (1-\beta)^{-1}s^* + \sum_{j=0}^{\infty}\beta^j E_t\varepsilon_{t+j}^F$$

Now inflation follows a martingale with $E_t\nu_{t+j} = \nu_t$ for $j \geq 1$.

The two polar cases are starkly different. By pegging the nominal interest rate, monetary policy anchors expected inflation on the steady-state (target) inflation rate and bond prices are constant. The full impact of a lower present value of surpluses must be absorbed by higher current inflation—lower ν_t—alone. But when monetary policy raises the nominal rate with current inflation by a proportion equal to the discount factor, higher current inflation is expected to persist indefinitely. Bond prices fall by the expected present value of that higher inflation rate, discounted at the rate $\beta\rho$. With the required change in inflation spread evenly over the term to maturity of outstanding debt, when fiscal news arrives, inflation needs to rise by far less than it does when bond prices are pegged. Of course, the "total"—present value—inflation effect of the fiscal shock is identical in the two cases. Although aggressive monetary policy cannot diminish the total inflationary impact, it can influence the timing of when inflation occurs.

We can consider both these polar cases and the intermediate case where $0 < \alpha_\pi < \beta$, by solving the model numerically in the presence of transfer shocks.[y] These are calibrated following Bi et al. (2013). We assume that the steady-state ratio of transfers to GDP is 0.18, government spending is 21% of GDP and taxes amount to 41% of GDP implying an (annualized) steady-state debt–GDP ratio of 50%. Transfers fluctuate according to an autoregressive process with persistence parameter of $\rho_z = 0.9$, and variance of $(0.005z^*)$.

[w] If monetary policy were to turn active, while fiscal policy remained active, then we would have an unstable equilibrium. The implications of temporarily being in such a regime are considered in Section 7.3.

[x] This result requires that $\beta\rho\nu_t < 1$ for all realizations of ν_t, so there cannot be "too much" deflation.

[y] The solution procedure follows Leith and Liu (2014), which relies on Chebyshev collocation methods and Gauss–Hermite quadrature to evaluate the expectations terms.

In this simple model with an active fiscal policy that does not respond to debt levels, the equilibrium outcome depends on the maturity of the debt stock and the responsiveness of monetary policy to inflation.

Fig. 1 plots the response to an increase in transfers. Each column represents a different value of the response of monetary policy to inflation. Monetary policy pegs the nominal rate in the first column, so the paths of all variables are the same across maturities: the entire adjustment occurs through surprise inflation in the initial period. In the second column $\alpha_\pi = 0.5$. Now differences emerge across maturities. With one-period debt the magnitude of the initial jump in inflation is the same as under a pegged interest rate because this is the price-level jump that is required to reduce the real value of debt to be consistent with lower surpluses. But the monetary policy reaction keeps inflation high for a prolonged period even though it is only the initial jump in inflation that serves to reduce the debt burden. As average maturity increases, the initial jump in inflation becomes smaller. A sustained rise in interest rates depresses bond prices, which allow the bond valuation equation to be satisfied at lower initial inflation rates. It is the surprise change in the *path* of inflation that occurs over the life of the maturing debt stock that reduces the real value of debt. With a positive value of α_π, any jump in inflation is sustained, which unexpectedly reduces the real returns that bondholders receive before that debt is rolled over. As we increase the responsiveness of the interest rate to inflation further to $\alpha_\pi = 0.9$, the surprise inflation needed to deflate the real value of debt remains unchanged for single-period debt, but is dramatically reduced for longer period debt. When $\alpha_\pi = 0.99$, as demonstrated analytically earlier, and $\rho > 0$, the rate of inflation follows a near-random walk, jumping to the level needed to satisfy the valuation equation.

The timing of the transfer shock—whether it is *i.i.d.* or persistent, realized immediately or in the future—does not matter beyond the change in the expected discounted value of surpluses that it produces. That present value must be financed with a path of inflation that combines current inflation surprises, and through bond prices, future inflation surprises, to ensure solvency. An anticipated increase in transfers produces surprise inflation today that reduces the current value of the outstanding debt stock, but whose value increase after the increase in transfers is realized.

This result foreshadows an important aspect of optimal policy, which Sections 4 and 5 explore: monetary policy can smooth the distortionary effects of fiscally induced inflation. The above analysis uses an endowment economy subjected to transfer shocks. That environment has the feature that under regime M, monetary policy can perfectly control inflation, while under regime F, prices are determined by the needs of fiscal solvency—the dichotomy across regimes that was emphasized in the original fiscal theory. The more general case breaks the dichotomy to produce interactions between monetary and fiscal policy in both policy regimes. This situation can arise even in the endowment economy when we consider government spending shocks rather than shocks to lump-sum transfers.

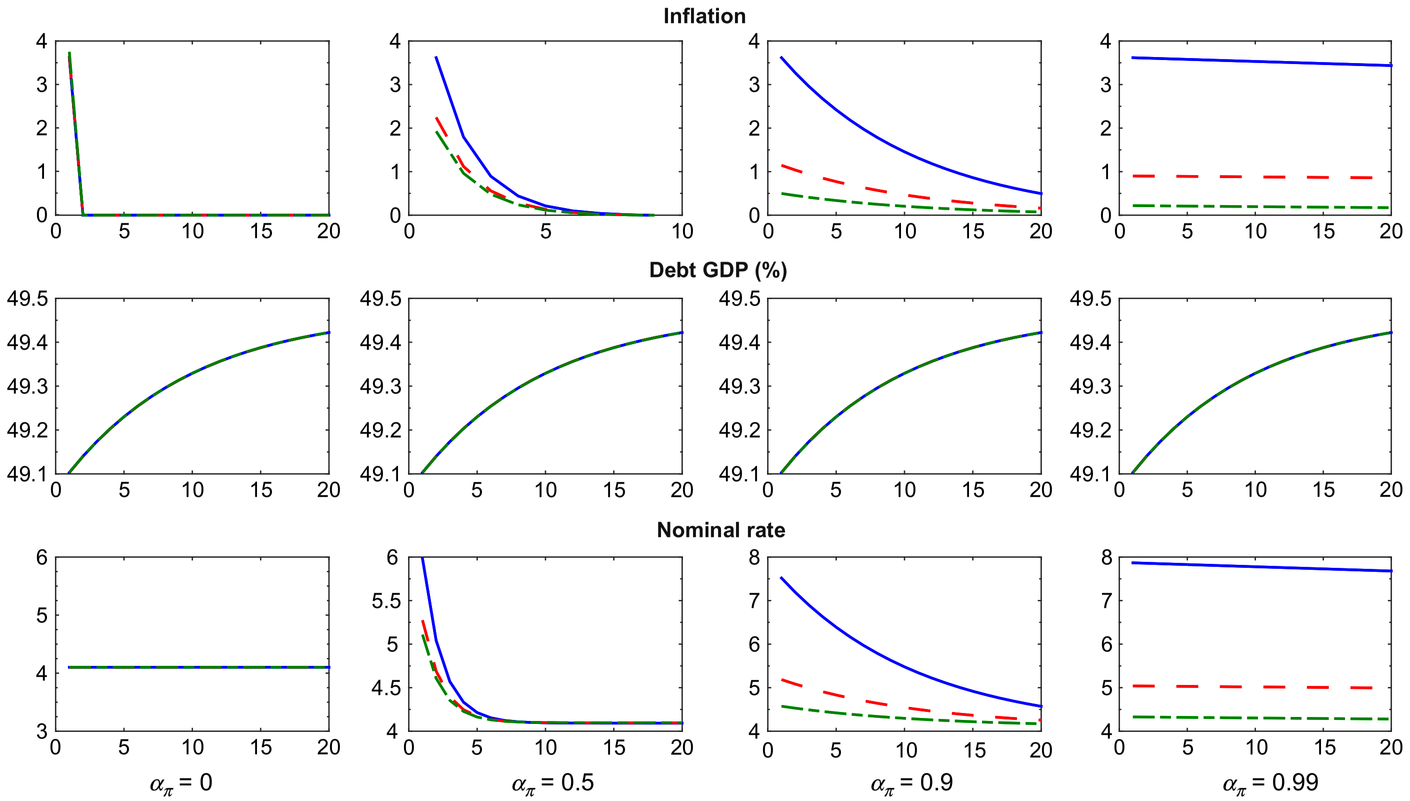


Fig. 1 Responses to an increase in transfers under alternative monetary policy rules and alternative maturity structures. 1-period debt (*solid lines*), 1-year debt (*dashed lines*), and 5-year debt (*dot-dashed lines*).

2.3.1 Increase in Government Spending

Government spending has implications for both monetary and fiscal policy. The direct impact on the government's finances is obvious. But given the resource constraint, $y = c_t + g_t$, variations in public consumption will have a one-for-one impact on private consumption which affects the stochastic discount factor. Through this channel government purchases carry additional effects on inflation and debt dynamics. Again we distinguish between the M and F regimes, although monetary and fiscal policy will interact under both.

2.3.1.1 Policy Under Regime M

When monetary policy is active and fiscal policy is passive, the analysis of the case of transfer shocks largely carries through, although with some additional monetary and fiscal interactions. Substituting the Fisher relation into the monetary policy rule yields the deflation dynamics[z]

$$v_t - v^* = \frac{\beta}{\alpha_\pi} E_t \left[\frac{u'(c_{t+1})}{u'(c_t)} v_{t+1} - v^* \right]$$

which can be solved forward as

$$v_t = \frac{\alpha_\pi - \beta}{\alpha_\pi} E_t \sum_{i=0}^{\infty} \left(\frac{\beta}{\alpha_\pi} \right)^i \frac{u'(c_{t+i})}{u'(c_t)} v^*$$

Inflation deviates from target in proportion to the deviations of the real interest rate path from steady state. Higher government spending raises the real interest rate and inflation.

Debt dynamics emerge from three distinct impacts of government spending: the direct effect on the fiscal surplus, the surprise inflation that arises in conjunction with the monetary policy rule, and movements in real interest rates. Monetary policy can insulate inflation from government spending shocks by reacting to real interest rates, as well as inflation, with the rule

$$\frac{1}{R_t} = \frac{1}{R^*} E_t \frac{u'(c_{t+1})}{u'(c_t)} + \alpha_\pi (\nu_t - \nu^*) \tag{28}$$

By this rule, the policymaker accommodates changes in the natural rate of interest caused by fluctuations in public consumption without deviating from the inflation target. To see this, combine this rule with the Fisher equation to get

[z] When the real interest rate can vary, the Fisher relation is

$$\frac{1}{R_t} = \beta E_t \frac{u'(c_{t+1})}{u'(c_t)} \nu_{t+1}$$

$$v_t - v^* = \frac{\beta}{\alpha_\pi} E_t \frac{u'(c_{t+1})}{u'(c_t)} (v_{t+1} - v^*)$$

Policy rule (28) implies that inflation/deflation is always equal to target, $v_t = v^*$. If the monetary policy rule does not respond to fiscal variables, inflation will be influenced by government spending shocks. Inflation can be insulated from fiscal shocks by allowing monetary policy to directly respond to the effects of fiscal policy on the natural rate of interest.

2.3.1.2 Policy Under Regime F

In regime F government spending shocks require jumps in inflation to satisfy the bond valuation equation[aa]

$$\begin{aligned}(1+\rho P_t^m)\frac{B_{t-1}^m}{P_t} &= E_t \sum_{i=0}^{\infty} \beta^i \frac{u'(c_{t+i})}{u'(c_t)} s_{t+i} \\ &= E_t \sum_{i=0}^{\infty} \beta^i \frac{u'(c_{t+i})}{u'(c_t)} s^* - E_t \sum_{i=0}^{\infty} \beta^i \frac{u'(c_{t+i})}{u'(c_t)} \varepsilon_{t+i}^G\end{aligned}$$

An increase in government spending increases the marginal utility of consumption, which increases real interest rates and requires a larger initial jump in inflation and drop in bond prices. Bond prices themselves are directly affected by the change in private consumption that arises when the government absorbs a larger share of resources, as the bond-pricing equation shows

$$P_t^m = \beta E_t (1 + \rho P_{t+1}^m) v_{t+1} \frac{u'(c_{t+1})}{u'(c_t)}$$

Bond prices fall initially and then gradually increase as the period of raised public consumption passes.

Adopting a specific form of utility, $u(c_t) = c_t^{1-\sigma}/(1-\sigma)$, with $\sigma = 2$, we can solve the model in the face of autocorrelated government spending shocks with $\rho_g = 0.9$, and variance of $0.005g^*$. As before, the stochastic model is solved nonlinearly using Chebyshev collocation methods (see Leith and Liu, 2014). Fig. 2 reflects the response to government spending shocks which are broadly consistent with the impacts of transfer shocks that appear in Fig. 1. The main difference is that the growth in consumption as government spending returns to steady state is equivalent to an increase in the real interest rate. But the main message that single-period debt requires an initial jump in inflation to stabilize debt and that this jump is unaffected by the description of the monetary policy parameter α_π remains. Once debt maturity extends beyond a single period, prolonging the initial

[aa] Shutting down shocks to lump-sum taxes and transfers, the surplus is defined as $s_t = \tau^* - z^* - g_t$, where $g_t = g^* \varepsilon_t^G$, and $\ln \varepsilon_t^g = \rho_g \ln \varepsilon_{t-1}^g + \xi_t$.

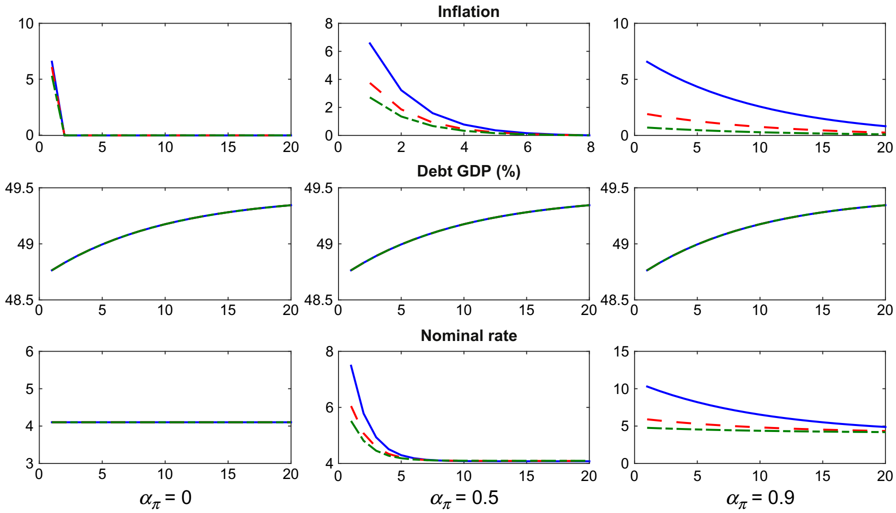


Fig. 2 Responses to an increase in government purchases under alternative monetary policy rules and alternative maturity structures. 1-period debt (*solid lines*), 1-year debt (*dashed lines*), and 5-year debt (*dot-dashed lines*).

jump in inflation can serve to reduce the magnitude of that initial jump. A sustained rise in inflation can also serve to satisfy the government's intertemporal budget identity through reducing bond prices. Essentially the inflation surprise is spread throughout the life to maturity of the outstanding debt stock.

3. PRODUCTION ECONOMIES WITH AD HOC POLICY RULES

The endowment economy is useful for understanding the mechanisms that underlie the fiscal theory. But the exogeneity of the real interest rate and the constancy of output limit a complete understanding of the theory and, in some cases, distort that understanding. We now turn to a conventional model in which inflation and output are determined jointly. In extending the analysis to the new Keynesian model we are widening the potential channels through which monetary and fiscal policy interact. To do so incrementally, we assume that taxes remain lump sum so that the effects of monetary policy on output do not affect the tax base to which a distortionary tax is applied. This means that the extra channel we are adding by introducing nominal inertia to a production economy is that monetary policy has influence over ex-ante real interest rates as well as nominal interest rates. This in turn means that the policymaker can ensure the bond valuation equation holds following fiscal shocks through a reduction in ex-ante real interest rates and not just ex-post real interest rates through inflation surprises.[ab] When we consider optimal policy in the new Keynesian model we shall allow taxes to distort behavior.

3.1 A Conventional New Keynesian Model

Endogenous output together with sticky prices allow both monetary policy and, in the case of regime F, fiscal policy to have real effects on the economy. We use a textbook version of a new Keynesian model of the kind that Woodford (2003) and Galí (2008) present. Because existing literature, including those two textbooks, thoroughly examines the nature of regime M equilibria, our exposition focuses exclusively on regime F.[ac]

The model's key features include: a representative consumer and firm; monopolistic competition in final goods; Calvo (1983) sticky prices in which a fraction $1-\phi$ of goods suppliers sets a new price each period; a cashless economy with one-period nominal bonds, B_t, that sell at price $1/R_t$, where R_t is also the monetary policy instrument; for now, government purchases are zero, so the aggregate resource constraint is $c_t = y_t$; an exogenous primary government surplus, s_t, with lump-sum taxes; and shocks only to monetary and fiscal policies.[ad] We solve a version of the model that is log-linearized around the deterministic steady state with zero inflation.

Let $\hat{x}_t \equiv \ln(x_t) - \ln(x^*)$ denote log deviations of a variable x_t from its steady-state value. Private-sector behavior reduces to a consumption-Euler equation

$$\hat{y}_t = E_t\hat{y}_{t+1} - \sigma(\hat{R}_t - E_t\hat{\pi}_{t+1}) \tag{29}$$

and a Phillips curve

$$\hat{\pi}_t = \beta E_t\hat{\pi}_{t+1} + \kappa\hat{y}_t \tag{30}$$

where $\sigma \equiv -\dfrac{u'(y^*)}{u''(y^*)y^*}$ is the intertemporal elasticity of substitution, $\omega \equiv \dfrac{w'(y^*)}{w''(y^*)y^*}$ is the elasticity of supply of goods, $\kappa \equiv \dfrac{(1-\phi)(1-\phi\beta)}{\phi}\dfrac{\omega+\sigma}{\sigma(\omega+\theta)}$ is the slope of the Phillips curve, and θ is the elasticity of substitution among differentiated goods. The parameters obey $0 < \beta < 1, \sigma > 0, \kappa > 0$.

[ab] By introducing this channel we could, in fact, turn off the revaluation effects stressed by the fiscal theory by assuming debt was solely real but still consider equilibria where monetary policy was passive and fiscal active. In this sense, as we widen the range of monetary and fiscal interactions, unconventional policy assignments do not necessarily require the revaluation mechanisms inherent in the fiscal theory to support determinate equilibria.

[ac] We draw from Woodford (1998a), but Kim (2003), Cochrane (2014), and Sims (2011) study closely related models.

[ad] Because these shocks have no effects on the natural rate of output, there is no distinction between deviations in output from steady state and the output gap.

3.1.1 Policy Rules

Monetary policy follows a conventional interest rate rule

$$\hat{R}_t = \alpha_\pi \hat{\pi}_t + \alpha_y \hat{y}_t + \varepsilon_t^M \tag{31}$$

and fiscal policy sets the surplus process, $\{\hat{s}_t\}$, exogenously, where $\hat{s}_t \equiv (s_t - s^*)/s^*$. By setting the surplus exogenously, we are implicitly assuming that taxes are lump sum so that any variations in real activity do not impact on the size of the tax base.

Policy choices must satisfy the flow budget identity, $\frac{1}{R_t}\frac{B_t}{P_t} + s_t = \frac{B_{t-1}}{P_t}$, which is linearized as

$$\hat{b}_t - \hat{R}_t + \left(\beta^{-1} - 1\right)\hat{s}_t = \beta^{-1}\left(\hat{b}_{t-1} - \hat{\pi}_t\right) \tag{32}$$

where b_t is real debt at the end of period t and π_t is the inflation rate between $t-1$ and t. Although this linearized budget identity does not appear to contain the steady-state debt-to-GDP ratio, the calibration of the surplus shock does implicitly capture the underlying steady-state level of debt.

3.1.2 Solving the Model in Regime F

The four-equation system—(29)–(32)—together with exogenous $\{\hat{s}_t\}$ yields solutions for $\{\hat{y}_t, \hat{\pi}_t, \hat{R}_t, \hat{b}_t\}$. Woodford (1998a) shows that a unique equilibrium requires that monetary policy react relatively weakly to inflation and output: α_π and α_y must satisfy

$$-1 - \frac{1+\beta}{\kappa}\alpha_y - \frac{2(1+\beta)}{\kappa\sigma} < \alpha_\pi < 1 - \frac{1-\beta}{\kappa}\alpha_y$$

For practical reasons, we restrict α_π's lower bound to 0. In this case, when monetary policy does not respond to output, this reduces to the condition that passive monetary policy requires $0 \le \alpha_\pi < 1$. In the analytical results that follow, we use this simplified policy rule; numerical results will bring the output response of monetary policy back in.

Substituting the simplified version of the monetary policy rule ($\alpha_y = 0$) into the government budget identity and iterating forward immediately yield several robust features of regime F equilibria

$$E_t\sum_{j=0}^{\infty}\beta^j\hat{\pi}_{t+j} = \left(\frac{1}{1-\alpha_\pi\beta}\right)\left[\hat{b}_{t-1} - (1-\beta)E_t\sum_{j=0}^{\infty}\beta^j\hat{s}_{t+j} + \beta E_t\sum_{j=0}^{\infty}\beta^j\varepsilon_{t+j}^M\right] \tag{33}$$

Although expression (33) is not an equilibrium solution to the model (since we still need to solve the path for inflation), it highlights several features that the solution displays. First, higher initial debt, a lower expected path of surpluses, or a higher expected path of the monetary shock all raise the present value of inflation. Second, a stronger response of monetary policy to inflation, but still consistent with existence of a bounded equilibrium, *amplifies* those inflationary effects. Dependence of inflation on the debt stock and surpluses is ubiquitous in regime F. Perversely, a higher path of the monetary shock or a higher value for α_π constitute a tightening of policy, yet they raise inflation.

In the flexible-price case, $\kappa = \infty$, so $\hat{y}_t \equiv 0$, and a solution for equilibrium inflation is immediate. This case collapses back to the endowment economy in Section 2.1.2.2 with a constant real rate and the simple Fisher relation $\hat{R}_t = E_t\hat{\pi}_{t+1}$. Combine the monetary policy rule with $\alpha_y = 0$ with the Fisher relation to solve for expected inflation

$$E_t\hat{\pi}_{t+j} = \alpha_\pi^j \hat{\pi}_t + \alpha_\pi^{j-1}\varepsilon_t^M + \alpha_\pi^{j-2}E_t\varepsilon_{t+1}^M + \cdots + \alpha_\pi E_t\varepsilon_{t+j-2}^M + E_t\varepsilon_{t+j-1}^M$$

and use this expression to replace expected inflation rates in (33). Equilibrium inflation is

$$\hat{\pi}_t = \hat{b}_{t-1} + \beta(1-\alpha_\pi\beta)E_t\sum_{j=0}^{\infty}\beta^j\varepsilon_{t+j}^M - (1-\beta)E_t\sum_{j=0}^{\infty}\beta^j\hat{s}_{t+j}$$

Actual inflation rises with initial debt, a higher path of the monetary policy shock, or a lower path for surpluses. The effects of surpluses on inflation are independent of the monetary policy choice of α_π, although we saw above that those fiscal effects on expected inflation are amplified by more aggressive monetary policy.

Solving the sticky-price new Keynesian model is more complicated. When $0 < \kappa < \infty$, both output and the real interest rate are endogenous. Defining the real interest rate as $\hat{r}_{t+j} \equiv \hat{R}_{t+j-1} - \hat{\pi}_{t+j}$, write the bond valuation equation as

$$\hat{\pi}_t - E_t\sum_{j=1}^{\infty}\beta^j\hat{r}_{t+j} = \hat{b}_{t-1} - (1-\beta)E_t\sum_{j=0}^{\infty}\beta^j\hat{s}_{t+j}$$

News about lower future surpluses shows up as a mix of higher current inflation and a lower path for the real interest rate. Lower real rates, in turn, transmit into higher output. Fiscal expansions have the old-Keynesian effects—higher real activity and inflation—and monetary policy behavior determines the split between them.

Combining the Euler equation, the Phillips curve and the monetary policy rule produce a second-order difference equation in inflation

$$E_t\hat{\pi}_{t+2} - \frac{1+\beta+\sigma\kappa}{\beta}E_t\hat{\pi}_{t+1} + \frac{1+\alpha_\pi\sigma\kappa}{\beta}\hat{\pi}_t = -\frac{\sigma\kappa}{\beta}\varepsilon_t^M$$

One can show that, given the restrictions on the underlying model parameters, this difference equation has two real roots, one inside $|\lambda_1| < 1$ and one outside $|\lambda_2 > 1|$ the unit circle, which yields the solution for expected inflation[ae]

[ae] Letting $\gamma_1 \equiv (1 + \beta + \sigma\kappa)/\beta$ and $\gamma_0 \equiv (1 + \alpha_\pi\sigma\kappa)/\beta$, the roots are $\lambda_1 = (1/2)(\gamma_1 - \sqrt{\gamma_1^2 - 4\gamma_0})$ and $\lambda_2 = (1/2)(\gamma_1 + \sqrt{\gamma_1^2 - 4\gamma_0})$. These derivations owe much to Tan (2015) who employs the techniques that Tan and Walker (2014) develop.

$$E_t\hat{\pi}_{t+1} = \lambda_1\hat{\pi}_t + (\beta\lambda_2)^{-1}\sigma\kappa E_t\sum_{j=0}^{\infty}\lambda_2^j\varepsilon_{t+j}^M \tag{34}$$

We can now solve for the j-step-ahead expectation of inflation by defining the operator $\mathcal{B}^{-j}x_t \equiv E_t x_{t+j}$ and iterating on (34)

$$\mathcal{B}^{-j}\hat{\pi}_t = \lambda_1^j\hat{\pi}_t + \frac{\sigma\kappa}{\lambda_2\beta}\frac{1}{1-\lambda_2^{-1}\mathcal{B}^{-1}}\left(\lambda_1^{j-1} + \lambda_1^{j-2}\mathcal{B}^{-1} + \cdots + \mathcal{B}^{-j+1}\right)\varepsilon_t^M$$

This yields the solution for expected discounted inflation that appears in (33)

$$E_t\sum_{j=0}^{\infty}\beta^j\hat{\pi}_{t+j} = \frac{1}{1-\lambda_1\beta}\hat{\pi}_t + \frac{\sigma\kappa}{\lambda_2(1-\lambda_1\beta)}\frac{1}{(1-\lambda_2^{-1}\mathcal{B}^{-1})(1-\beta\mathcal{B}^{-1})}\varepsilon_t^M$$

Using this expression for discounted inflation in (33) delivers a solution for equilibrium inflation

$$\begin{aligned}\hat{\pi}_t = {} & \left(\frac{1-\lambda_1\beta}{1-\alpha_\pi\beta}\right)\left[\hat{b}_{t-1} - \left(\frac{1-\beta}{1-\beta\mathcal{B}^{-1}}\right)\hat{s}_t\right] \\ & + \left[\frac{1-\lambda_1\beta}{1-\alpha_\pi\beta} - \frac{\sigma\kappa}{\lambda_2}\frac{1}{(1-\lambda_2^{-1}\mathcal{B}^{-1})}\right]\frac{1}{1-\beta\mathcal{B}^{-1}}\varepsilon_t^M\end{aligned} \tag{35}$$

It is straightforward to show how the monetary policy parameter affects inflation

$$\frac{\partial\lambda_1}{\partial\alpha_\pi} > 0, \quad \frac{\partial\lambda_2}{\partial\alpha_\pi} < 0, \quad \frac{\partial[\lambda_2(1-\lambda_1\beta)]}{\partial\alpha_\pi} < 0 \quad \frac{\partial\left(\frac{1-\lambda_1\beta}{1-\alpha_\pi\beta}\right)}{\partial\alpha_\pi} > 0$$

More aggressive monetary policy—larger α_π—affects the equilibrium in the following ways

- amplifies the impacts on inflation from outstanding debt and exogenous disturbances to monetary policy and surpluses
- makes the effects of these shocks on inflation more persistent.

Evidently, if fiscal policies set surpluses exogenously, monetary policy is impotent to offset fiscal effects on inflation. And adopting a more hawkish monetary policy stance has the perverse effect of amplifying and propagating the effects of shocks on inflation.

In this basic new Keynesian model, fiscal disturbances are transmitted to output through the path of the ex-ante real interest rate, as the consumption-Euler equation, (29), makes clear. Define the one-period real interest rate as $\hat{r}_t \equiv \hat{R}_t - E_t\hat{\pi}_{t+1}$. To simplify expressions, temporarily shut down the monetary policy shock, $\varepsilon_t^M \equiv 0$. Date the solution for inflation from (35) at t + 1, take expectations, and substitute the monetary policy rule for the interest rate. After some tedious algebra, the equilibrium real interest rate is

$$\hat{r}_t = \frac{(\alpha_\pi - \lambda_1)(1-\lambda_1\beta)}{1-\alpha_\pi\beta}\left[\hat{b}_{t-1} - (1-\beta)\sum_{j=0}^{\infty}\hat{s}_{t+j}\right]$$

The lead coefficient, $\alpha_\pi - \lambda_1$, depends on monetary policy behavior and on all the model parameters. Because its sign can be positive or negative, lower expected surpluses may lower or raise the short-term real interest rate on impact.

Substituting the monetary policy rule into the definition of the real interest rate and suppressing the monetary policy shock yield

$$\hat{r}_t = \alpha_\pi\hat{\pi}_t - E_t\hat{\pi}_{t+1}$$

Using the Phillips curve to eliminate inflationary expectations we obtain

$$\hat{r}_t = (\alpha_\pi - \beta^{-1})\hat{\pi}_t - \beta^{-1}\kappa\hat{y}_t$$

which shows that a given level of positive inflation and output deviations from steady state will be consistent with lower real interest rates the smaller is the monetary policy response to inflation. The intuition is very similar to that in the endowment economy: a passive monetary policy that responds to inflation generates a sustained rise in inflation which does not facilitate the stabilization of single-period debt. In the new Keynesian case such a policy response mitigates the reduction in debt service costs which are an additional channel through which the passive monetary policy stabilizes debt in a sticky-price economy.

3.2 Maturity Structure in Regime F

We introduce the simplified maturity structure that Section 2.2.2 describes, in which government debt maturity decays at the constant rate ρ each period, into the new Keynesian model of Section 3.1. The no-arbitrage condition links bond prices to the one-period nominal interest rate

$$\hat{P}_t^m = -\hat{R}_t + \beta\rho E_t\hat{P}_{t+1}^m$$

which implies the term structure relation

$$\begin{aligned}\hat{P}_t^m &= -E_t\sum_{j=0}^{\infty}(\beta\rho)^j\hat{R}_{t+j}\\ &= -\frac{1}{1-\beta\rho\mathcal{B}^{-1}}\left[\alpha_\pi\hat{\pi}_t + \varepsilon_t^M\right]\end{aligned}$$

where we have substituted the simpler monetary policy rule in for the nominal interest rate.

The government's flow budget identity is

$$\beta(1-\rho)\hat{P}_t^m + \beta\hat{b}_t^m + (1-\beta)\hat{s}_t + \hat{\pi}_t = \hat{b}_{t-1}^M \tag{36}$$

where we are defining $b_t^m \equiv B_t^m / P_t$ to be the real face value of outstanding debt.[af] Because bond prices depend on the expected infinite path of inflation and the monetary policy shock, analytical solutions along the lines of Section 3.1.2, though feasible, are cumbersome. For example, the analog to the discounted inflation expression, (33), is

$$\frac{1}{1-\beta\mathcal{B}^{-1}}\left[1-\frac{\alpha_\pi\beta(1-\rho)}{1-\beta\rho\mathcal{B}^{-1}}\right]\hat{\pi}_t = \hat{b}_{t-1}^m - \left(\frac{1-\beta}{1-\beta\mathcal{B}^{-1}}\right)\hat{s}_t + \frac{\beta(1-\rho)}{(1-\beta\mathcal{B}^{-1})(1-\beta\rho\mathcal{B}^{-1})}\varepsilon_t^M$$

which collapses to (33) when $\rho = 0$ so all debt is one period. The solution for equilibrium inflation, like that when there is only one-period debt in equation (35), depends on all the parameters of the model through the eigenvalues λ_1 and λ_2, but the analytical expression for inflation is too complex to offer useful intuition.

One-period debt makes the value of debt depend only on the current nominal interest rate and, through the monetary policy rule, current inflation. A maturity structure makes that value depend on the entire expected path of nominal interest rates. This gives monetary policy an expanded role in debt stabilization, allowing expected future monetary policy to affect the value of current debt. This additional channel operates through terms in $1/(1-\beta\rho\mathcal{B}^{-1})$ that create double infinite sums in the equilibrium solution.

3.2.1 Impacts of Fiscal Shocks

Figs. 3 and 4 illustrate the impacts of a serially correlated increase in the primary fiscal deficit financed by nominal bond sales.[ag] Fig. 3 maintains that all debt is one period to focus on how different monetary policy rules alter the impacts of a fiscal expansion.

When monetary policy pegs the nominal interest rate—$\alpha_\pi = \alpha_Y = 0$—it fixes the bond price, which front loads fiscal adjustments through current inflation and the real interest rate. Inflation rises, the real rate falls and output increases. Responses inherit the serial correlation properties of the fiscal disturbance. As monetary policy becomes progressively less passive, reacting more strongly to inflation and output, it amplifies and propagates the fiscal shock (dashed lines in Fig. 3). By reacting more strongly to inflation, monetary policy ensures that the real interest rate declines by less, tempering the short-run output increases.

The figure makes clear the role that debt plays in propagating shocks in regime F. Stronger and more persistent nominal interest rate increases transmit directly into stronger and more persistent growth in the nominal market value of debt.[ah] And persistently higher nominal debt keeps household nominal wealth and, therefore, nominal demand elevated, creating strong serial correlation in inflation and output. This internal

[af] The real market value is $P_t^m B_t^m / P_t$. To derive (36), we use the steady-state relationships $P^{m*} = 1/(\beta^{-1} - \rho)$ and $s^*/b^{m*} = (1-\beta)/(1-\beta\rho)$ in log-linearizing the government budget identity.

[ag] We calibrate the model to an annual frequency, setting $\beta = 0.95, \sigma = 1, \kappa = 0.3$. The surplus is $AR(1)$, $\hat{s}_t = \rho_{FP}\hat{s}_{t-1} + \varepsilon_t^F$, with $\rho_{FP} = 0.6$.

[ah] Growth in the nominal market value of debt is $P_t^m B_t^M / P_{t-1}^m B_{t-1}^m$.

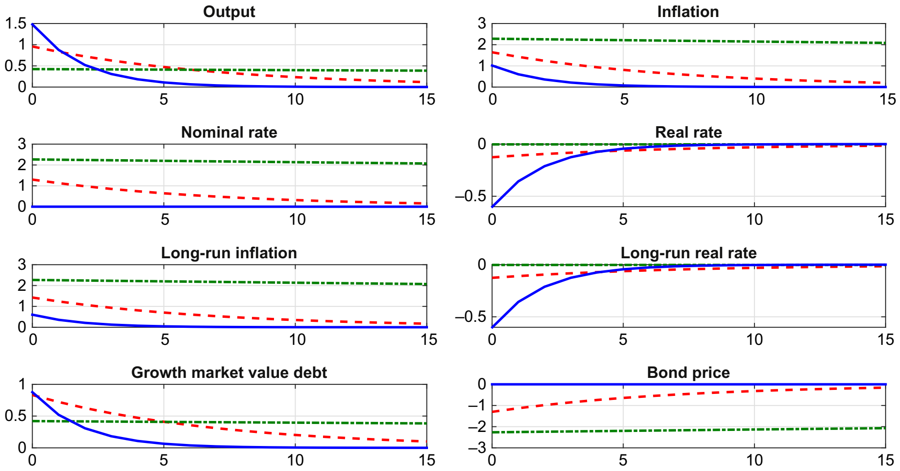


Fig. 3 Responses to a 20% increase in the initial deficit under alternative monetary policy rules when all debt is one period. Calibration reported in Footnote ag. $\alpha_\pi = \alpha_Y = 0$ (*solid lines*), $\alpha_\pi = \alpha_Y = 0.5$ (*dashed lines*), and $\alpha_\pi = 0.9$, $\alpha_Y = 0.5$ (*dot-dashed lines*).

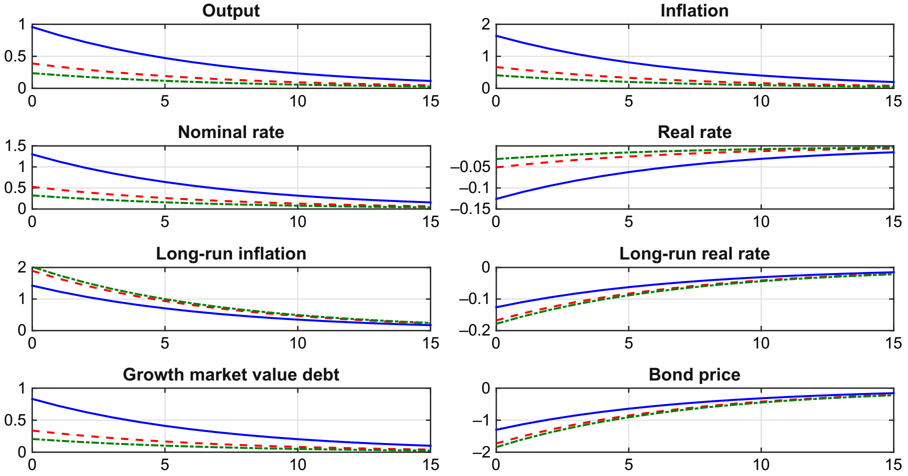


Fig. 4 Responses to a 20% increase in the initial deficit under alternative maturity structures. Calibration reported in Footnote ag with $\alpha_\pi = \alpha_Y = 0.5$. 1-year debt (*solid lines*), 5-year debt (*dashed lines*), and consol debt (*dot-dashed lines*).

propagation mechanism through government debt is absent from regime M, where higher debt carries with it the promise of higher taxes that eliminate wealth effects.

Fig. 4 holds the monetary policy rule fixed, setting $\alpha_\pi = \alpha_Y = 0.5$, to reveal how changes in maturity affect fiscal impacts. The figure contrasts one-period debt (solid lines) to an average of 5-year maturity (dashed lines) and consol debt (dot-dashed lines). Longer

Table 2 The fiscal shock initially raises the deficit by 20%

α_π	α_Y	Maturity	% due to $\hat{\pi}_t$	% due to $\hat{P}_t^m$	% due to $\hat{r}_{t+j}^m$
0	0	1 period	44	0	56
0.5	0.5	1 period	71	0	29
0.9	0.5	1 period	98	0	2
0.5	0.5	5 years	29	59	12
0.9	0.5	5 years	20.4	79.2	0.4
0.5	0.5	Consol	18	75	7
0.9	0.5	Consol	6	94	0

"% due to" are the ratios of the right-hand components of (38) to ξ_t, which is computed from the impulse response of $\hat{s}_{t+j}$, as described in the text. Calibration reported in Footnote ag.

maturities force more of the adjustment to higher deficits into lower bond prices, which push more of the impacts into low-frequency movements in long-run inflation and real interest rates.[ai]

Although short-run inflation is higher with one-period debt, in the long run inflation is lower with shorter maturity bonds. With long debt, bond prices reflect anticipated inflation rates farther into the future, in essence spreading inflationary effects over longer horizons. The cost of doing so is to raise the long-run inflation impacts of fiscal policy.

Another way to summarize the dynamic impacts of fiscal disturbances is to ask how a shock that raises primary deficits by a certain amount gets financed intertemporally, as a function of various model parameters. Underlying the calculations in Table 2 are two basic mechanisms that stabilize debt in the face of the surplus shock. First are the revaluation effects that we can summarize by examining the ex-post real return to holding government bonds in any period

$$r_t^m = \frac{(1+\rho P_t^m)}{P_{t-1}^m}\frac{1}{\pi_t}$$

or in linearized form

$$\hat{r}_t^m = \rho\beta\hat{P}_t^m - \hat{\pi}_t - \hat{P}_{t-1}^m$$

By contrasting this with the ex-ante returns the bond holders were expecting when they purchased the bonds in period $t-1$ we can identify the scale of the revaluation effects, which linearized, are

$$\hat{r}_t^m - E_{t-1}\hat{r}_t^m = -(\hat{\pi}_t - E_{t-1}\hat{\pi}_t) + \rho\beta(\hat{P}_t^m - E_{t-1}\hat{P}_t^m) \tag{37}$$

[ai] The long-term real interest rate, $\hat{r}_t^L$, comes from combining the bond-pricing equation and the Fisher relation to yield the recursion $\hat{r}_t^L = \hat{r}_t + \beta\rho E_t\hat{r}_{t+1}^L$. The long-run inflation rate, $\hat{\pi}_t^L$, which is the expected path of inflation discounted by $\beta\rho$, may be computed as $\hat{\pi}_t^L = -\hat{r}_t^L - \hat{P}_t^m$.

The first term on the right in (37) gives the losses suffered by bondholders due to surprise inflation in the initial period. The second term gives the losses suffered by holders of mature debt ($\rho > 0$) arising from jumps in bond prices caused by innovations to the expected future path of inflation. These latter revaluation effects are borne by the existing holders of government debt and arise for innovations to the path of inflation over the time to maturity of the debt stock they hold. In the sticky-price economy these effects can be complemented by reductions in the ex-ante real rates of return received by future bondholders, which reduce effective debt service costs to create an additional channel through which debt can be stabilized.[aj]

In the case of one-period debt it is only the surprise inflation in the initial period that reduces the real value of government debt. This is then combined with reductions in ex-ante real interest rates to stabilize debt. As α_π increases, there is less reliance on the latter effect and larger jumps in the initial rate of inflation are required to satisfy the bond valuation equation. When we move to longer period debt, there is an additional revaluation effect through the impact of innovations to the path of inflation on bond prices. With bond prices adjusting, we can have smaller, but more sustained, increases in inflation that reduce the real market value of debt. These continue to be combined with reductions in ex-ante real interest rates to satisfy the bond valuation equation with these debt service cost effects falling as monetary policy becomes less passive.

To see how this affects the decomposition of the adjustment required to stabilize the debt stock in the face of a surplus shock consider the evolution of the market value of government debt

$$\tilde{b}_t = r_t^m \tilde{b}_{t-1} - s_t$$

where $\tilde{b}_t \equiv \dfrac{P_t^m Bt}{P_t}$. This can be linearized as

$$\beta \hat{\tilde{b}}_t = \hat{r}_t^m + \hat{\tilde{b}}_{t-1} - (1-\beta)\hat{s}_t$$

Using the expected value of surpluses, $\xi_t \equiv (1-\beta)E_t \sum_{j=0}^{\infty} \beta^j \hat{s}_{t+j}$ which implies $(1-\beta)\hat{s}_t = \xi_t - \beta E_t \xi_{t+1}$, this becomes

$$\beta(\hat{\tilde{b}}_t - E_t \xi_{t+1}) - \hat{r}_t^m = \hat{\tilde{b}}_{t-1} - \xi_t$$

Iterating forward we obtain

[aj] An equivalent interpretation comes from thinking about the value of debt in the "forward" direction, as being determined by the expected present value of surpluses. Lower real interest rates raise real discount factors to increase the present value of a given stream of surpluses.

$$\begin{aligned} \xi_t &= \hat{\tilde{b}}_{t-1} + \hat{r}^m_t + E_t \sum_{j=1}^{\infty} \beta^j \hat{r}^m_{t+j} \\ &= \hat{\tilde{b}}_{t-1} - \hat{P}^m_{t-1} + \beta\rho\hat{P}^m_t - \hat{\pi}_t + E_t \sum_{j=1}^{\infty} \beta^j \hat{r}^m_{t+j} \end{aligned} \tag{38}$$

The required adjustment to a change in expected surpluses is made up of surprise changes in the returns to existing bond holders $\hat{r}^m_t$ as well as expected future returns on bond holdings, $E_t \sum_{j=1}^{\infty} \beta^j \hat{r}^m_{t+j}$. The former is made up of jumps in the initial rate of inflation combined with changes in bond prices to the extent that bonds have a maturity greater than one period, $\rho > 0$. The latter captures the reduction in ex-ante real interest rates which can occur in our sticky-price economy.

Table 2 computes the objects in (38) from impulse responses to a deficit innovation. When debt is single period, bond prices do not contribute to financing the deficit. If monetary policy pegs the nominal interest rate, current inflation and future real interest rates play nearly equally important roles. As monetary policy reacts more aggressively to inflation and output, real interest rate responses are tempered, and an increasing fraction of the adjustment occurs through inflation at the time of the fiscal innovation. Longer maturity debt brings bond prices into the adjustment process, and their role grows with both the maturity of debt and the aggressiveness of monetary policy. As a consequence, current inflation moves much less. Consol bonds, together with aggressive monetary policy, push nearly all the adjustment into bond prices, with contemporaneous inflation playing only a minimal role, as the last row of the table reports.

3.2.2 Impacts of Monetary Shocks

Section 2.1.2.2 describes the effects of exogenous monetary policy disturbances in an endowment economy under regime F. Because future surpluses do not adjust to neutralize the wealth effects of monetary policy, contractionary policy—a higher path for the nominal interest rate—raises household interest receipts and wealth, raising nominal aggregate demand. A similar phenomenon can arise in the new Keynesian model, though the dynamics are more interesting.

Fig. 5 reports the impacts of an exogenous monetary policy action that raises the nominal interest rate. To highlight the behavior of monetary policy in regime F, we consider three different monetary policy rules. A rule that does not respond to inflation (solid lines) raises the short-term real interest rate and depresses output in the short run. Despite the drop in output, inflation rises immediately, even in a model where the Phillips curve implies a strong positive relationship between output and inflation contemporaneously ($\kappa = 0.3$).

This seemingly anomalous outcome underscores the centrality of wealth effects in regime F. Higher nominal interest rates raise households' interest receipts in the future,

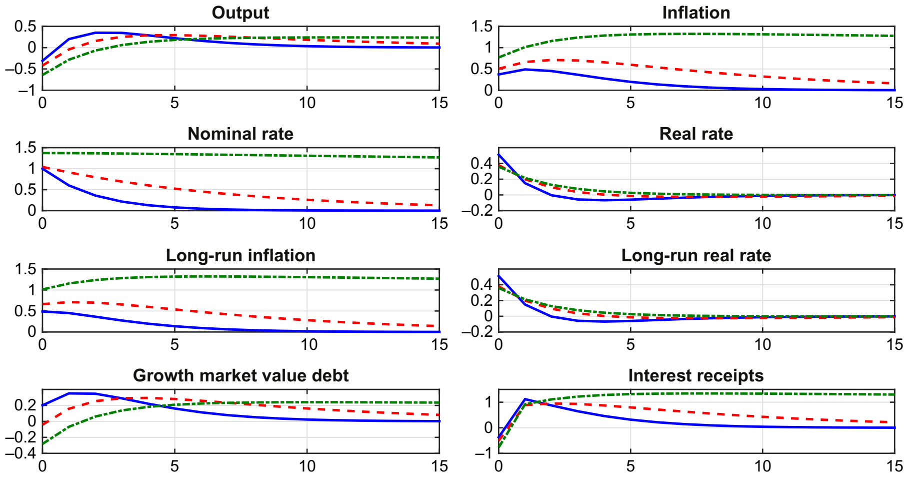


Fig. 5 Responses to a 1% monetary contraction under alternative monetary policy rules with only one-period government debt. Calibration reported in Footnote ag. The monetary policy shock follows the *AR*(1) process $\epsilon_t^M = \rho_{MP}\epsilon_{t-1}^M + \zeta_t^M$ with $\rho_{MP} = 0.6.\alpha_\pi = \alpha_Y = 0$ (*solid lines*), $\alpha_\pi = \alpha_Y = 0.5$ (*dashed lines*), and $\alpha_\pi = 0.9$, $\alpha_Y = 0.5$ (*dot-dashed lines*).

triggering an expectation of higher future demand and inflation.[ak] Through the Phillips curve, the higher expected inflation dominates the deflationary effects of lower output to raise inflation on impact. Expectations are critical to output effects as well. After an initial decline, output always eventually rises because the real interest rate declines at longer horizons.

More aggressive monetary policy behavior (dashed lines) transforms the transitory increase in the policy rate into larger and more persistent increases. Those higher nominal interest rates raise both the growth rate of the nominal market value of debt and real interest receipts. The resulting wealth effects raise and prolong the higher inflation.

That an exogenous monetary policy "contraction," which raises the nominal interest rate, also raises inflation may seem to contradict evidence from the monetary VAR literature. This pattern, dubbed the "price puzzle" by Eichenbaum (1992), is sometimes taken to indicate that monetary policy behavior is poorly identified, perhaps by misspecifying the central bank's information set, as Sims (1992) argues. Fig. 5 makes clear that there is nothing puzzling about the pattern from the perspective of the fiscal theory.

Introducing long debt makes impulse responses accord better with VAR evidence because bond prices absorb much of the monetary shock. Fig. 6 contrasts one-period (solid lines) with 5-year (dashed lines) and consol debt (dot-dashed lines). By reducing growth in the market value of debt, longer maturities attenuate the inflationary effects and make the short-run decline in output longer lasting. Inflation does eventually rise,

[ak] Real interest receipts are defined as $[(1+\rho P_t^m)/P_{t-1}^m](b_{t-1}^m/\pi_t)$.

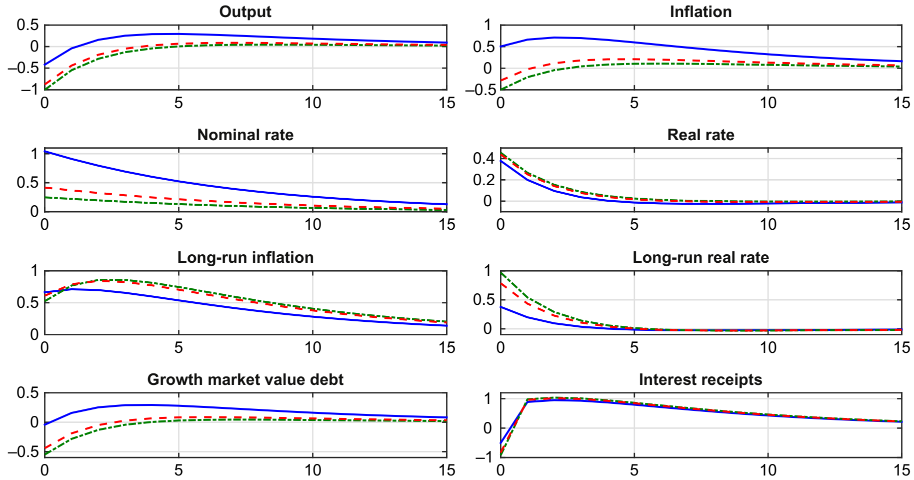


Fig. 6 Responses to a 1% monetary contraction under alternative maturity structures. Calibration reported in Footnote ag. The monetary policy shock follows the *AR*(1) process $\epsilon_t^M = \rho_{MP}\epsilon_{t-1}^M + \zeta_t^M$ with $\rho_{MP} = 0.6$ and $\alpha_\pi = \alpha_Y = 0.5$. 1-period debt (*solid lines*), 5-year debt (*dashed lines*), and consol debt (*dot-dashed lines*).

Table 3 A 1% monetary shock initially raises the short-term nominal interest rate

κ	σ	$\hat{\pi}_t$	$\hat{P}_t^m$	$\hat{r}_{t+j}^m$
0.3	1.0	−0.29	1.12	−0.83
∞	1.0	−1.54	1.54	0.0
0.1	1.0	−0.09	1.03	−0.94
0.3	5.0	−0.50	0.76	−0.26
0.3	0.5	−0.17	1.32	−1.15

"π_t" and "$\hat{P}_t^m$" are impacts of the monetary policy shock on contemporaneous inflation and bond prices; "$\hat{r}_{t+j}^m$" are the impacts on discounted real returns to bonds from expression (39). Calibration reported in Footnote ag plus $\alpha_\pi = \alpha_Y = 0.5$ and maturity set at 5 periods.

as it must if bond prices are lower. Sims (2011) calls the pattern of falling, then rising inflation following a monetary contraction "stepping on a rake."

While Fig. 6 shows how the response of short-run inflation to a monetary contraction varies with debt maturity, Table 3 reports how other model parameters affect this relationship. Following a monetary contraction, $\xi_t \equiv 0$ in expression (38), so if the monetary shock hits at time t, we have that

$$\hat{\pi}_t - \beta\rho\hat{P}_t^m - E_t\sum_{j=1}^{\infty}\beta^j\hat{r}_{t+j}^m = 0 \tag{39}$$

so the three sources of fiscal financing—higher current inflation, lower current bond prices, and lower future real bond returns—must sum to zero.

The first row of Table 3 shows that for the benchmark calibration with five-period average bond maturity, the monetary contraction initially lowers inflation along with the price of bonds, while it raises discounted real interest rates. As prices become more flexible ($\kappa \rightarrow \infty$), the impact on inflation becomes more pronounced, while that on real rates diminishes. A higher intertemporal elasticity of substitution ($\sigma \rightarrow 0$) pushes more of the adjustment into the future, reduces the effect on current inflation, and raises the impacts on bond prices and future real rates.

4. ENDOWMENT ECONOMIES WITH OPTIMAL MONETARY AND FISCAL POLICIES

In this section we turn to consider the nature of optimal policy in our simple endowment economy. In doing so we cut across various strands of the literature that addresses optimal monetary and fiscal policy issues.

4.1 Connections to the Optimal Policy Literature

We begin by considering Ramsey policies where the policymaker has an ability to make credible promises about how they will behave in the future, before turning to time-consistency issues below. We start by building on Sims' (2013) analysis. He considers a simple linearized model of tax smoothing under commitment in the face of transfer shocks and long-term debt. The policymaker can use costly inflation surprises as an alternative to distortionary taxation to ensure fiscal solvency. We extend that work in several ways. Specifically, we allow for a geometric maturity structure which nests single-period debt and consols as special cases, employ nonlinear model solution techniques, and allow for anticipated and unanticipated government spending shocks, in addition to transfer shocks. Nonlinear solutions allow us to consider the way in which the size of the debt stock, together with its maturity structure, influences the optimal combination of monetary and fiscal policy in debt stabilization. Innovations to the expected path for inflation can affect bond prices in a way which helps to satisfy the bond valuation equation even without any fiscal adjustment. These bond price movements are effective only if applied to a nonzero stock of outstanding liabilities so the optimal balance between inflation and tax financing of fiscal shocks depends on both the level of government debt and its maturity structure.

Without an ability to issue state-contingent debt or use inflation surprises to stabilize debt, Barro (1979) showed that debt and taxes should follow martingale processes to minimize the discounted value of tax distortions. While Barro did consider the impact of surprise inflation on the government's finances, these were treated as exogenous shocks rather than something that can be optimally employed to further reduce tax distortions. Lucas and Stokey (1983) is an equally influential paper that reaches quite different conclusions on the optimal response of tax rates to shocks. Lucas and Stokey consider an economy where the government can issue real state-contingent debt and show that it is optimal for a government to issue a portfolio of debt where the state-contingent returns to that debt isolate the government's finances from shocks so that there is no need for taxes to jump in the manner of Barro's tax-smoothing result. Instead, taxes are largely flat and inherit the dynamic properties of the exogenous shocks hitting the economy.

A large part of the post-Lucas and Stokey literature considers the implications of debt that is not state contingent, as well as ways of converting the payoffs from portfolios of nonstate-contingent debt into state-contingent payoffs. A key result is that when debt payoffs are not (or cannot be made) state contingent, then the optimal policy looks more like Barro's tax-smoothing result. Aiyagari et al. (2002) show this by assuming that debt is single period and noncontingent in a model otherwise identical to that of Lucas and Stokey. How might noncontingent debt instruments be made to mimic the payoffs that would be generated by state-contingent debt? Two approaches have been suggested in the literature. First, surprise inflation can render the real payoffs from risk-free nominal bonds state contingent. For example, Chari et al. (1994) use a model where surprise inflation is costless to show that the real contingencies in debt exploited by Lucas and Stokey could be created through monetary policy via surprise inflation when government debt is nominal. This underpins Sims' (2001) results in a model with costless inflation in which tax rates should be held constant to finance any fiscal shocks solely with surprise movements in inflation.

When we start to introduce a cost to surprise inflation, the optimal policy can be strikingly different. For a jointly determined optimal monetary and fiscal policy operating under commitment, Schmitt-Grohé and Uribe (2004) show that in a sticky-price stochastic production economy, even a miniscule degree of price stickiness will result, under the optimal policy, in a steady-state rate of inflation marginally less than zero, with negligible inflation volatility. In other words, although the optimal policy under flexible prices would be to follow the Friedman rule and use surprise inflation to create the desired state contingencies in the real payoffs from nominal debt, even a small amount of nominal inertia heavily tilts optimal policy toward zero inflation with little reliance on inflation surprises to insulate the government's finances from shocks. As in Benigno and Woodford (2004) and Schmitt-Grohé and Uribe (2004) return to the tax-smoothing results of Barro (1979) thanks to the effective loss of state-contingent returns to debt when prices are sticky. Sims (2013) argues that this may be due to the fact that Schmitt-Grohé and Uribe only consider single-period debt; with longer term debt the efficacy of using innovations to the expected path of inflation to affect bond prices would be enhanced. This is the first issue to which we turn: to what extent will the optimizing policymaker rely on fiscal theory-type revaluations of debt through innovations to the expected path of prices?

While the state contingencies in real bond payoffs can be generated through the impact of surprise inflation on nominal bonds, an alternative approach when bonds are real is to exploit variations in the yield curve to achieve the same contingencies for the government's whole bond portfolio. With single-period risk-free real bonds, Ramsey policy in the Lucas and Stokey model possesses a unit root as in Barro. Angeletos (2002) and Buera and Nicolini (2004) use the maturity structure of nonstate-contingent real bonds to render the overall portfolio state contingent. With two states for government spending, for example, a portfolio of positive short-term assets funded by issuing long-term debt can insulate the government's finances from government spending shocks. More generally, with a sufficiently rich maturity structure the policymaker can match the range of the stochastic shocks hitting the economy and achieve this hedging. The second broad optimal policy question we consider is: what is the role of debt management in insulating the government's finances from shocks?

Having looked at the ability of the Ramsey policymaker to both hedge against shocks and utilize monetary policy as a debt stabilization tool when complete hedging is not possible, we turn to consider the time-inconsistency problem inherent in such policies. We find that constraining policy to be time consistent radically affects the policymaker's ability to hedge against fiscal shocks and generates serious "debt stabilization bias" problems, as in Leith and Wren-Lewis (2013), that are akin to the inflationary bias problems analyzed in the context of monetary economies.

We begin by considering the role inflation surprises play in optimal policy in our simple endowment economy with a geometrically declining maturity structure. We then generalize these results to a more general maturity structure and consider the role of debt management in hedging for fiscal shocks. We then turn to a simple example where complete hedging is feasible.

4.2 The Model

We follow Sims (2013) in defining the inverse of inflation as $\nu_t = \pi_t^{-1}$, and assuming the policymaker's objective function is given by

$$-E_0 \frac{1}{2} \sum_{t=0}^{\infty} \beta^t \left[\tau_t^2 + \theta(\nu_t - 1)^2 \right]$$

which the policymaker maximizes subject to the constraints given by the resource constraint in our endowment economy,

$$y = c_t + g_t$$

the bond valuation equation (after assuming a specific form for per-period utility, $u(c_t) = \frac{c_t^{1-\sigma}}{1-\sigma}$)

$$\beta E_t \frac{(1+\rho P^m_{t+1})}{P^m_t}\nu_{t+1}\left(\frac{c_{t+1}}{c_t}\right)^{-\sigma}=1$$

the government's flow budget identity

$$b_t P^m_t=(1+\rho P^m_t)b_{t-1}\nu_t+g_t-\tau_t-z_t$$

and the associated transversality condition

$$\lim_{j\to\infty} E_t\left(\prod_{i=0}^{j}\frac{1}{R^m_{t+i+1}\nu_{t+i+1}}\right)\frac{P^m_{t+j}B^m_{t+j}}{P_{t+j}}\geq 0$$

where $R^m_{t+1}\equiv(1-\rho P^m_{t+1})/P^m_t$, and government spending and/or transfers follow exogenous stochastic processes. Our adopted objective function is clearly ad hoc in the context of our simple endowment economy. However, it can easily be motivated as capturing the trade-off between the costs of tax vs inflation financing in richer production economies. Indeed, many of the insights this analysis offers will reappear when considering optimal policy in a fully microfounded economy subject to distortionary taxation and nominal inertia in Section 5.

4.3 Ramsey Policy

We analyze the time-inconsistent Ramsey policy for our endowment economy given the policymaker's objective function by forming the following Lagrangian

$$\begin{aligned}L_t=E_0\frac{1}{2}\sum_{t=0}^{\infty}\beta^t[&-\frac{1}{2}(\tau_t^2+\theta(\nu_t-1)^2)\\&+\mu_t\left(\beta E_t\frac{(1+\rho P^m_{t+1})}{P^m_t}\nu_{t+1}\left(\frac{c_{t+1}}{c_t}\right)^{-\sigma}-1\right)\\&+\lambda_t(b_tP^m_t-(1+\rho P^m_t)b_{t-1}\nu_t-g_t-z_t+\tau_t)]\end{aligned}$$

which yields the first-order conditions

$$\begin{aligned}&\tau_t:-\tau_t+\lambda_t=0\\&\nu_t:-\theta(\nu_t-1)+\mu_{t-1}\frac{(1+\rho P^m_t)}{P^m_{t-1}}\left(\frac{c_t}{c_{t-1}}\right)^{-\sigma}-(1+\rho P^m_t)\lambda_t b_{t-1}=0\\&P^m_t:-\frac{\mu_t}{P^m_t}+\mu_{t-1}\rho\frac{\nu_t}{P^m_{t-1}}\left(\frac{c_t}{c_{t-1}}\right)^{-\sigma}+\lambda_t(b_t-\rho\nu_t b_{t-1})=0\\&b_t:\lambda_tP^m_t-\beta E_t(1+\rho P^m_{t+1})\nu_{t+1}\lambda_{t+1}=0\end{aligned}$$

Defining $\tilde{\mu}_t\equiv\dfrac{\mu_t}{P^m_t c_t^{-\sigma}}$ the system to be solved for $\{P^m_t,\tilde{\mu}_t,\nu_t,\tau_t,b_t,c_t\}$ is given by

$$-\theta(\nu_t - 1) + \widetilde{\mu}_{t-1}(1+\rho P_t^m)c_t^{-\sigma} - (1+\rho P_t^m)\tau_t b_{t-1} = 0$$

$$\tau_t b_t - \widetilde{\mu}_t c_t^{-\sigma} - \rho\nu_t(\tau_t b_{t-1} - \widetilde{\mu}_{t-1}c_t^{-\sigma}) = 0$$

$$\tau_t P_t^m - \beta E_t(1+\rho P_{t+1}^m)\nu_{t+1}\tau_{t+1} = 0$$

$$\beta E_t \frac{(1+\rho P_{t+1}^m)}{P_t^m}\left(\frac{c_{t+1}}{c_t}\right)^{-\sigma}\nu_{t+1} - 1 = 0$$

$$b_t P_t^m - (1+\rho P_t^m)b_{t-1}\nu_t - g_t + \tau_t - z_t = 0$$

$$g_t - (1-\rho_g)g^* - \rho_g g_{t-1} - \varepsilon_t^g = 0$$

$$z_t - (1-\rho_z)z^* - \rho_z z_{t-1} - \varepsilon_t^z = 0$$

$$y - c_t - g_t = 0$$

with two exogenous shocks describing the evolution of government consumption, g_t, and transfers, z_t and two endogenous state variables, $\widetilde{\mu}_{t-1}$ and b_{t-1}, where the former captures the history dependence in policymaking under commitment.

To obtain some intuition for how policy operates under commitment, it is helpful to consider three polar cases. First, where inflation is costless, so that $\theta = 0$. Second, where inflation is so costly that the economy can be considered to be real, $\theta \to \infty$. Third, we allow inflation to be costly $\theta > 0$, but assume that taxes have reached the peak of the Laffer curve so that they are no longer available to engage in tax smoothing and instead are held constant, $\tau_t = \overline{\tau}$.

4.3.1 Costless Inflation

In the former case, where inflation is costless ($\theta = 0$), the first two first-order conditions imply

$$\widetilde{\mu}_{t-1}c_t^{-\sigma} = \tau_t b_{t-1}$$

and

$$\tau_t b_t - \widetilde{\mu}_t c_t^{-\sigma} = \rho\nu_t(\tau_t b_{t-1} - \widetilde{\mu}_{t-1}c_t^{-\sigma})$$

Substituting the first into the second, lagging one period, and comparing the first condition yield

$$\tau_t = \left(\frac{c_t}{c_{t-1}}\right)^{-\sigma}\tau_{t-1}$$

In the absence of government spending shocks (the only source of variation in private consumption in our simple endowment economy) taxes are unchanged. But taxes are higher whenever government spending is higher. In the case of transfer shocks, inflation jumps to satisfy the bond valuation equation and this is a pure case of the fiscal theory. But when bonds have a maturity beyond a single period, there are an infinite number of patterns of inflation which can satisfy this, due to the impact inflation has on bond prices. While there is a unique required discounted magnitude of surprise inflation needed to satisfy the government debt valuation condition, there are a variety of paths which can achieve that magnitude. When the fiscal shock is a shock to government consumption, this affects real interest rates so that even though inflation can costlessly stabilize debt at its initial steady-state level, there is still tilting of tax rates: during periods of high real interest rates, it is desirable to suffer the short-run costs of higher taxation to avoid the longer run costs of supporting the higher steady-state level of debt that would emerge when higher interest rates raise the rate of debt accumulation. In this case it is only because of the commitment to honor the past promises not to deflate away the government's outstanding liabilities that there are positive tax rates at all.

4.3.2 Real Economy

In the second case, inflation is so costly it would never be used under the optimal policy, $\theta \rightarrow \infty$ and $\nu_t = 1$. As a result, we rely on jumps in the tax rate to satisfy government solvency and we return to a world of pure tax smoothing, where the tax rate follows the path implied by the first-order condition

$$\tau_t P_t^m = \beta E_t (1 + \rho P_{t+1}^m) \tau_{t+1}$$

Under a perfect foresight equilibrium this reduces to

$$\frac{\tau_t}{c_t^{-\sigma}} = \frac{\tau_{t+1}}{c_{t+1}^{-\sigma}}$$

This tax rate is constant in the face of transfer shocks, but will be tilted in the presence of government spending shocks—the tax rate at t is higher (lower) when public consumption is anticipated to rise (fall). The fact that it is purely forward-looking captures the usual tax-smoothing result that the tax rate will jump to the level required to satisfy the government's budget identity, although we have tilting in the tax rate to capture changes in real interest rates induced by government spending shocks. Eventually, the tax rate will achieve a new long-run value consistent with servicing the new steady-state level of debt.

4.3.3 Intermediate Case

In the intermediate case where $0 < \theta < \infty$, the tax-smoothing condition remains as above, but is combined with a pattern of inflation described by

$$-\theta(\nu_t - 1) + \frac{\mu_{t-1}}{P^m_{t-1}}(1 + \rho P^m_t) - (1 + \rho P^m_t)\tau_t b_{t-1} = 0$$

$$\tau_t b_t - \frac{\mu_t}{P^m_t} - \rho\nu_t\left(\tau_t b_{t-1} - \frac{\mu_{t-1}}{P^m_{t-1}}\right) = 0$$

$$b_t P^m_t - (1 + \rho P^m_t) b_{t-1}\nu_t - g_t - z_t + \tau_t = 0$$

which will deliver initial jumps in inflation, bond prices, and tax rates to ensure fiscal solvency. These first-order conditions also imply that gross inflation returns to 1 in steady state, so the optimal commitment policy makes any inflation only temporary. But there is a continuum of steady-state debt levels, each with an associated optimal tax rate, that are consistent with the steady state of the first-order conditions under commitment.

When we consider a variant on the third case where taxes are no longer available for tax smoothing, either for political reasons or because the tax rate has reached the peak of the Laffer curve, the relevant optimality conditions become

$$\lambda_t P^m_t - \beta E_t(1 + \rho P^m_{t+1})\nu_{t+1}\lambda_{t+1} = 0$$

$$-\theta(\nu_t - 1) + \tilde{\mu}_{t-1}(1 + \rho P^m_t)c_t^{-\sigma} - (1 + \rho P^m_t)\lambda_t b_{t-1} = 0$$

$$\lambda_t b_t - \tilde{\mu}_t c_t^{-\sigma} - \rho\nu_t(\lambda_t b_{t-1} - \tilde{\mu}_{t-1} c_t^{-\sigma}) = 0$$

where the tax rate is fixed at $\overline{\tau}$.

Here the unit root in government debt is no longer present because taxes cannot adjust to support a new steady-state debt level, and inflation cannot influence future surpluses. Instead, inflation must be adjusted to ensure fiscal solvency by returning debt to the steady-state level consistent with the unchanged tax rate. The pattern of inflation also depends on the maturity structure of the inherited debt stock. To see this more clearly we consider the perfect foresight solution in the face of a transfers shock in which the first-order condition for debt implies that $\lambda_t = \lambda_{t+1}$ since $g_t = g^*$. Combining the second and third conditions yields

$$\nu_t(\nu_t - 1) = \left[1 + (\rho P^m_t)^{-1}\right]\beta\nu_{t+1}(\nu_{t+1} - 1)$$

which describes the dynamics of inflation. Inflation rises following a fiscal shock that would otherwise make debt initially higher and then decline toward its steady-state value. The rate of convergence depends on the inverse of the maturity parameter multiplied by the bond price, which initially falls, but then recovers as the period of inflation passes. When $\rho = 0$ the inflation only occurs in the initial period, but becomes more protracted the longer is the maturity of government debt. Similar inflation dynamics are observed when taxes are smoothed, although the magnitude of the initial jump in inflation will be reduced to the extent that tax rates rise to stabilize debt at a higher level in the face of a given shock.

4.4 Numerical Results

The grid-based approach to solving the stochastic version of the model under the simple rules works well when the economy has a well-defined steady state to which it returns. With commitment policies the model enters a new steady state following the realization of a shock, which makes the model difficult to solve using these techniques. For this reason, when considering commitment we restrict attention to perfect foresight equilibrium paths following an initial shock. These paths are computed as follows. We guess the new steady-state value of debt and solve the steady state of the Ramsey problem conditional on that guess. This serves as a terminal condition on the model solution 800 periods in the future. The Ramsey first-order conditions are then solved for 800 periods conditional on this guess for the ultimate steady state. If the solution exhibits a discontinuity between the final period of the solution and the imposed terminal condition, the steady-state guess is revised. This process continues until the guessed new steady state is indeed the steady state to which the economy now settles.

We begin by considering the same transfers shock considered above for various degrees of maturity and different initial debt-to-GDP ratios. The autocorrelated shock to transfers reduces the discounted value of future surpluses and requires a monetary and/or fiscal adjustment. These adjustments are plotted in Fig. 7 for various initial debt-to-GDP ratios and debt maturities. The first column starts from an initial debt-to-GDP ratio of zero. When debt is initially zero and the initial tax rate of $\tau = 0.39$ can support the initial level of transfers and public consumption, under the optimal policy there is no inflation, regardless of the maturity of debt. This is due to the fact that surprise changes in inflation or bond prices only help satisfy the government's intertemporal budget identity if there is already an initial debt stock for them to act on. Even though the debt that will be issued as a result of the transfer shock is of different maturities across the experiments reported in the first column of the figure, this will not affect the optimal policy response to the transfers shock when there is initially no debt. The tax rate jumps to a permanently higher level to support a higher steady-state debt level, as under Barro's (1979) original tax-smoothing result.

The second column begins from an initial steady state with a debt-to-GDP ratio of 25% (and a supporting initial tax rate of $\tau = 0.4$). Now there is mild use of inflation to offset the effects of the transfers shock. Inflation is smaller but more sustained the longer is the average maturity of debt. As maturity lengthens, inflation surprises play an increasingly important role in stabilizing debt, with smaller adjustments in taxes. At higher debt levels, the role of inflation and maturity grows in importance as substitutes for distorting taxes. Ultimately, the increase in inflation is unwound (it serves no purpose as the initial debt stock matures) and there is a permanent increase in both the debt stock and tax rates. These examples underscore that optimal policy is highly state dependent, particularly with respect to the level and maturity of debt at the time the shock hits.

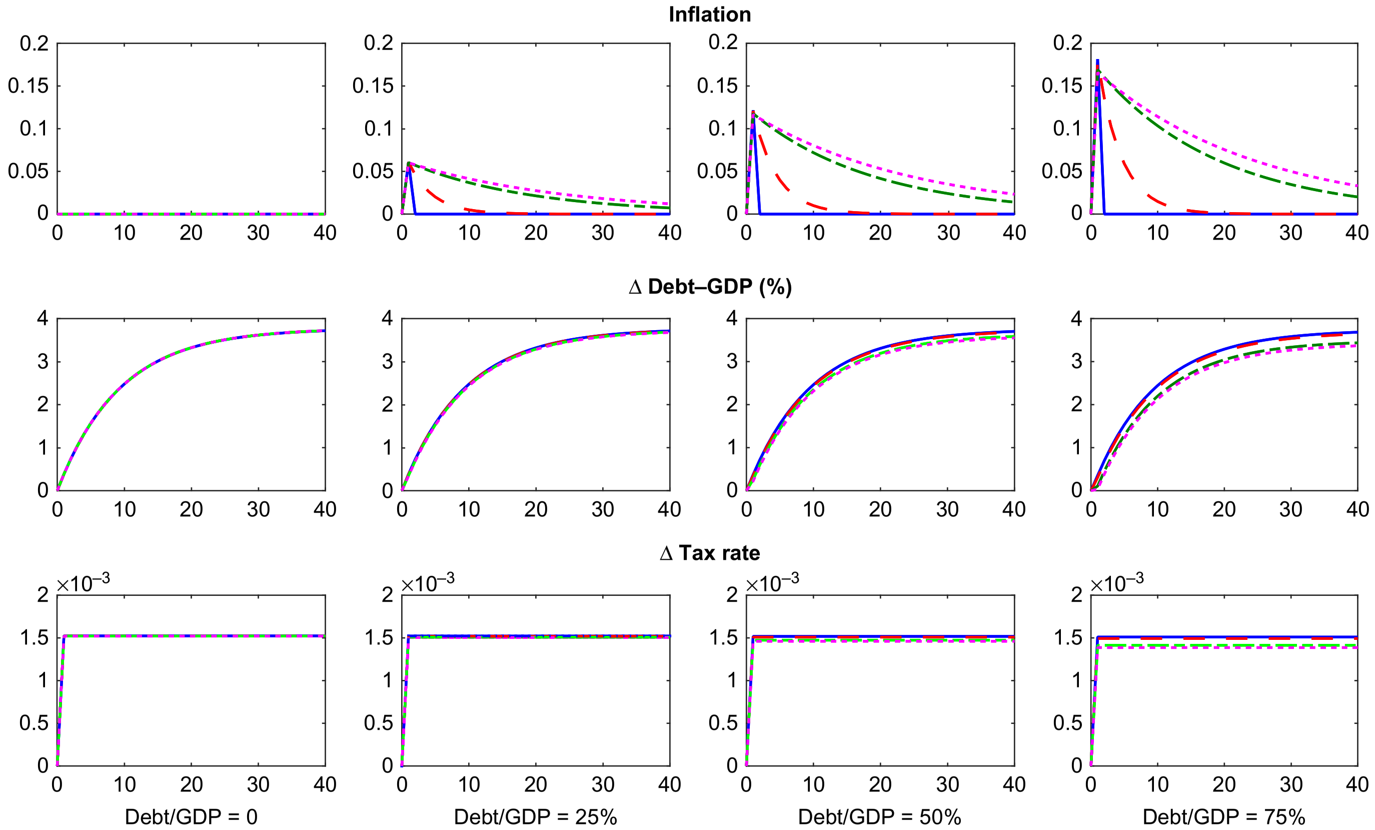


Fig. 7 Optimal policy in response to higher transfers with different debt levels and maturities. 1-period debt (*solid lines*), 1-year debt (*dashed lines*), 4-year debt (*dot-dashed lines*), and 5-year debt (*dotted lines*).

When we turn to government spending shocks in Fig. 8, the story is similar except that now, through the stochastic discount factor, public consumption tilts the optimal path of taxes and affects the magnitude of the fiscal and inflation adjustments needed to satisfy the debt valuation equation. With no initial stock of debt, the subsequent debt maturity structure is irrelevant and the optimal policy does not generate any inflation. But for a positive initial debt level, the spike in inflation for one-period debt is several orders of magnitude larger than for the portfolio of bonds with an average maturity of 8 years. With only short debt, the inflation is immediately eliminated, while the slight rise in inflation is sustained in the presence of longer term debt. Sustained inflation decreases bond prices that reduce the value of debt to for the more mature bonds, permitting the policymaker to reduce the required jump in the tax rate needed to support the higher level of steady-state debt. Interestingly, the higher tax rates during the period of raised public consumption end up reducing the new steady-state level of debt so that the new steady-state tax rate is actually lower than before the shock. This contrasts to the case of the transfer shock where debt levels were raised following the shock.

Fig. 9 reports optimal responses to news of a sustained increase in government spending 5 years in the future. Initially inflation falls and the tax rate jumps down in support of a debt level that is ultimately lower, despite the increase in government spending. This occurs because the policymaker raises the tax rate for the duration of the rise in public consumption to avoid the rapid accumulation of government debt in a period when real interest rates are relatively high. Bond prices rise as the anticipated increase in government spending approaches and then drop dramatically when the spending is realized.

In this experiment the cost of inflation is quite high, $\theta = 10$. A lower cost would lead to greater reliance on the use of monetary policy and innovations in the anticipated path of prices to stabilize debt. As we show later, even this relatively conservative weight on the costs of inflation still generates a sizeable endogenous inflation bias when we consider time-consistent policy.

4.5 Ramsey Policy with a General Maturity Structure

Although the geometrically declining maturity structure is a tractable and plausible description of the profile of government debt for many economies, it is useful to broaden the analysis with a more general description of the maturity structure. This generalization refines the description of the role of optimal inflation surprises in stabilizing debt and begins to consider the role of debt management in insulating the government's finances from fiscal shocks. We employ Cochrane's (2001) notation, allowing the bond valuation equation to be written as in (19) in Section 2.2.1. The government's optimization problem becomes

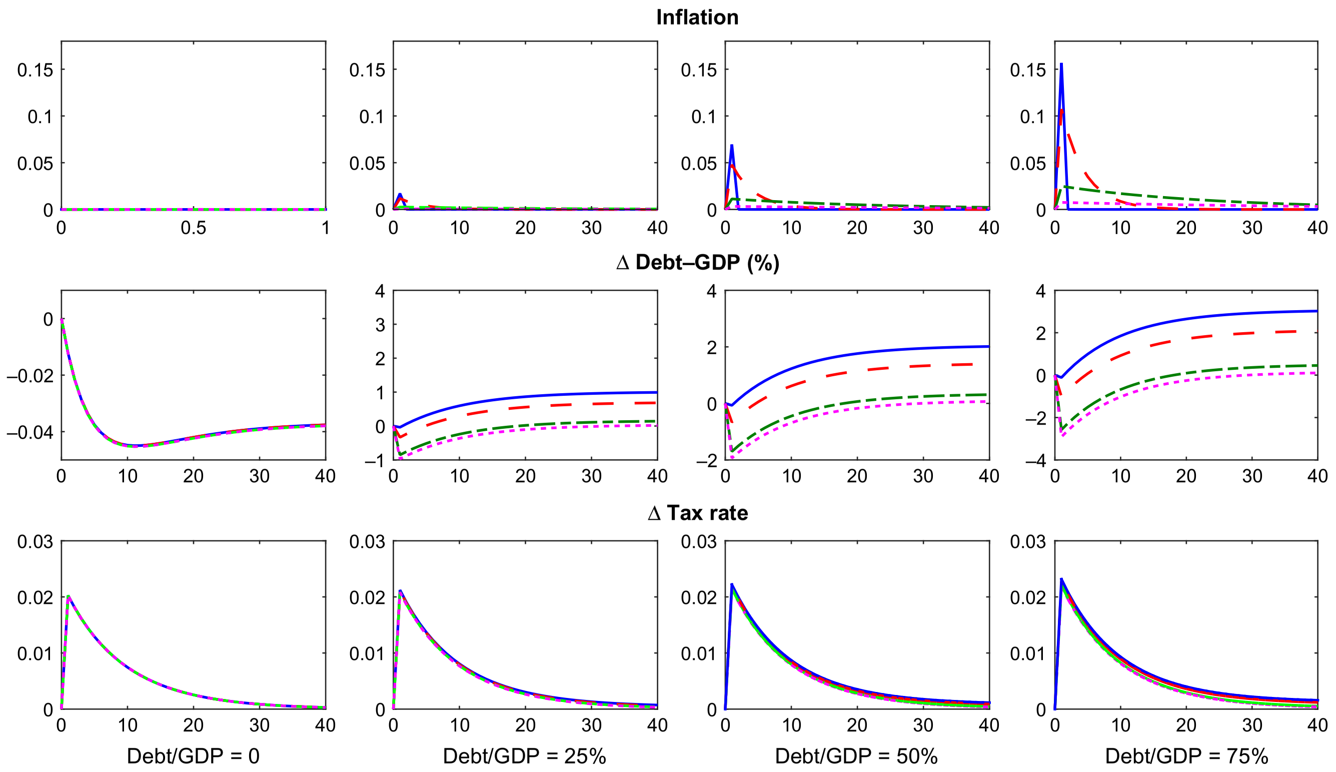


Fig. 8 Optimal policy in response to higher government spending with different debt levels and maturities. 1-period debt (*solid lines*), 1-year debt (*dashed lines*), 4-year debt (*dotted–dashed lines*), and 5-year debt (*dotted lines*).

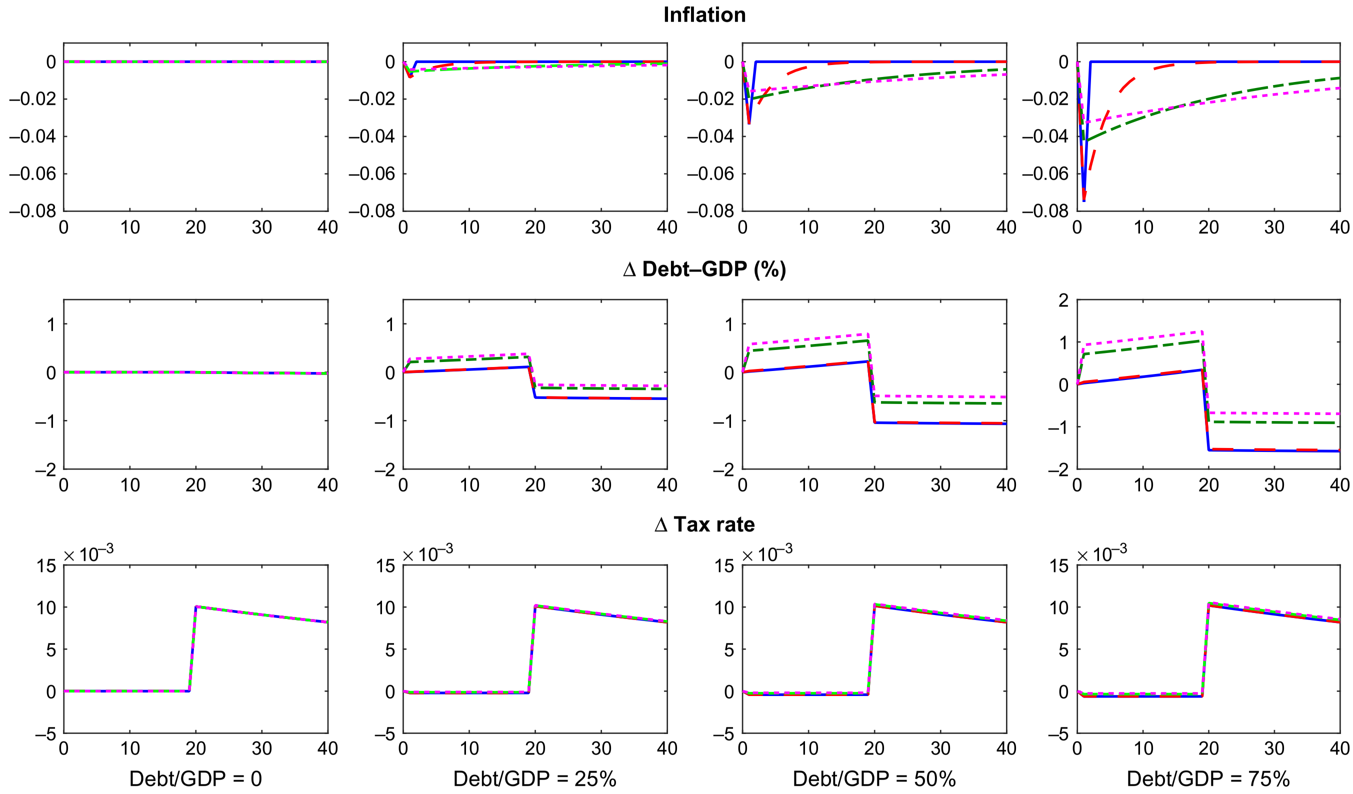


Fig. 9 Optimal policy in response to an anticipated increase in government spending with different debt levels and maturities. 1-period debt (*solid lines*), 1-year debt (*dashed lines*), 4-year debt (*dotted–dashed lines*), and 5-year debt (*dotted lines*).

$$L_0 = E_0 \sum_{t=0}^{\infty} \beta^t \left[-\frac{1}{2}(\tau_t^2 + \theta(\nu_t - 1)^2) \right.$$
$$\left. + \lambda_t \left(-\sum_{j=0}^{\infty} E_t \left[\beta^j u'(c_{t+j}) \prod_{s=0}^{j} \nu_{t+s} \right] \left[\frac{B_t(t+j)}{P_{t-1}} - \frac{B_{t-1}(t+j)}{P_{t-1}} \right] - u'(c_t)(\tau_t - g_t - z_t) \right) \right]$$

The first-order condition for taxation is

$$-\tau_t = u'(c_t)\lambda_t$$

The debt management problem optimally chooses the maturity structure of debt issued in period t which is repayable at future dates, $B_t(t + j)$, to yield the optimality condition

$$-\beta^t \lambda_t \beta^j E_t u'(c_{t+j}) \prod_{s=0}^{j} \nu_{t+s} \frac{1}{P_{t-1}} = -\beta^{t+1} E_t \lambda_{t+1} \beta^{j-1} u'(c_{t+j}) \prod_{s=0}^{j} \nu_{t+s} \frac{1}{P_{t-1}}$$

which can be simplified as

$$\frac{\tau_t}{u'(c_t)} E_t u'(c_{t+j}) \prod_{s=0}^{j} \nu_{t+s} = E_t \frac{\tau_{t+1}}{u'(c_{t+1})} u'(c_{t+j}) \prod_{s=0}^{j} \nu_{t+s}$$

which implies

$$E_t \left[\left[\frac{u'(c_t)}{u'(c_{t+1})} \tau_{t+1} - \tau_t \right] \frac{u'(c_{t+j})}{u'(c_t)} \frac{P_{t-1}}{P_{t+j}} \right] = 0$$

The covariance between the payoff of debt instrument of maturity j periods and next period's tax rate is zero (Bohn, 1990). This is the hedging across states that Angeletos (2002) and Buera and Nicolini (2004) explore. By structuring debt in this way the policymaker minimizes the fiscal and monetary adjustments required in the face of shocks; those policy adjustments then depend on the magnitude and maturity of the outstanding debt stock. To see how debt management can mitigate the need for adjusting tax rates and generating inflation in the face of fiscal shocks, we construct a simple example in the following section where the policymaker can completely insulate the government's finances from government spending shocks.

The final first-order condition is for deflation

$$-\beta^t \theta(\nu_t - 1)\nu_t + \sum_{i=0}^{t} \beta^i \lambda_i \left(-\sum_{j=0}^{\infty} \left[\beta^j u'(c_{i+j}) \prod_{s=0}^{j} \nu_{i+s} \right] \left[\frac{B_i(i+j)}{P_{i-1}} - \frac{B_{i-1}(i+j)}{P_{i-1}} \right] \right)$$

This can be combined with the condition for debt management and quasi-differenced to obtain, under perfect foresight

$$(\nu_t - 1)\nu_t = \beta(\nu_{t+1} - 1)\nu_{t+1} + \theta^{-1}\lambda_0 u'(c_t)\left[\frac{B_{-1}(t)}{P_t}\right]$$

This expression highlights more clearly the link between inflation and the maturity structure of the predetermined debt stock than does the geometrically declining maturity structure. The inflation dynamics under the optimal policy are in a very similar form to the nonlinear new Keynesian Phillips curve when price stickiness results from Rotemberg (1982) quadratic adjustment costs. The key difference is that the forcing variable is the element of the predetermined debt stock that matures in period t. Deflation/inflation anticipates the rate at which the debt stock issued at time $t = -1$ when the plan was formulated, matures. This makes current inflation reflect the discounted value of future debt as it matures. As debt matures, the effectiveness of inflation diminishes and inflation falls: the optimal rate of inflation jumps and gradually erodes until all the initial outstanding debt stock has matured. Notice that this Ramsey plan for inflation is only affected by debt dated at time $t = -1$, and the maturity structure of debt issued after this initial period is irrelevant in a perfect foresight environment. Future maturities will affect the government's ability to insure against fiscal shocks in a stochastic environment. We can see this latter point more clearly by considering a simple example.

4.6 Commitment and Hedging

Angeletos (2002) and Buera and Nicolini (2004) argue that debt maturity should be structured to insure the economy against shocks by having the government issue long-term liabilities, but hold an almost offsetting portfolio of short-term assets (the net difference being the government's overall level of indebtedness). In the face of fluctuating spending needs and interest rates, bond prices adjust to help finance debt without requiring any change in taxation. In these papers the short and long positions are constant over time, so that they do not require active management, although numerically they are extremely large positions (for example, five or six times the value of GDP in Buera and Nicolini, 2004). This approach amounts to another way to introduce the contingency in overall debt payments even though these individual assets/liabilities are not state contingent.

To construct a simple example of the use of debt management for hedging purposes we consider an environment where taxes and transfers are at their steady-state values ($\tau_{t+j} = \tau^*$ and $z_{t+j} = z^*$). Government spending can either take the value of $g^h > g^*$, with probability 1/2, or $g^l < g^*$ with complementary probability. Government debt takes the form of a single-period bond of quantity b^s issued in period t, repayable in period $t + 1$, and a portfolio of longer term bonds of geometrically declining maturity, so that the quantity of debt issued in period t maturing in period $t + j$ is $\rho^j b^m$. With a single *i.i.d.* shock all that is required for complete hedging is that the maturity structure contains both one- and two-period debt to enable us to perfectly hedge, as in Buera and Nicolini. With additional *i.i.d.* shock processes, complete hedging is not possible, as we would require some persistence in the shock process and longer term debt. Because we wish to contrast this case with a scenario where a time-consistent policymaker seeks to use debt management for the purposes of hedging and mitigating time-consistency problems, we allow for a combination of longer term bonds and short-term bonds in which varying proportions of the two types can act as a proxy for changes in average debt maturity. In this example, transfer shocks, which amount to shocks that do not directly affect bond prices and interest rates, cannot be completed hedged, although movements in inflation as part of the optimal policy response could provide some hedging opportunities.

Generalizing the Ramsey policy considered above to include a single-period nominal bond as well as the portfolio of bonds with geometrically declining maturity, the system of first-order conditions to be solved as part of the Ramsey problem is

$$-\theta(\nu_t-1)+\widetilde{\mu}_{t-1}(1+\rho P_t^m)c_t^{-\sigma}+\widetilde{\gamma}_{t-1}c_t^{-\sigma}-(1+\rho P_t^m)\tau_t b_{t-1}-\tau_t b_{t-1}^s=0$$

$$\tau_t b_t-\widetilde{\mu}_t c_t^{-\sigma}-\rho\nu_t(\tau_t b_{t-1}-\widetilde{\mu}_{t-1}c_t^{-\sigma})=0\tau_t b_t^s-\widetilde{\gamma}_t c_t^{-\sigma}=0$$

$$\tau_t P_t^m-\beta E_t(1+\rho P_{t+1}^m)\nu_{t+1}\tau_{t+1}=0$$

$$\tau_t P_t^s-\beta E_t\nu_{t+1}\tau_{t+1}=0$$

$$\beta E_t\frac{(1+\rho P_{t+1}^m)}{P_t^m}\left(\frac{c_{t+1}}{c_t}\right)^{-\sigma}\nu_{t+1}-1=0$$

$$\beta E_t\left(\frac{c_{t+1}}{c_t}\right)^{-\sigma}\nu_{t+1}-P_t^s=0$$

$$b_t P_t^m+b_t^s P_t^s-(1+\rho P_t^m)b_{t-1}\nu_t-b_{t-1}^s\nu_t-g_t-z^*+\tau_t=0$$

$$g_t=g^i,\quad i=h,l\,\text{with prob }1/2$$

where $\widetilde{\mu}_{t-1}=\frac{\mu_{t-1}}{P_{t-1}^m c_{t-1}^{-\sigma}},\widetilde{\gamma}_{t-1}=\frac{\gamma_{t-1}}{P_{t-1}^s c_{t-1}^{-\sigma}}$, and γ_t is the Lagrange multiplier associated with the pricing of single-period bonds, $P_t^s=\beta E_t\left(\frac{c_{t+1}}{c_t}\right)^{-\sigma}\nu_{t+1}$. There are four state variables—$\widetilde{\mu}_{t-1},\widetilde{\gamma}_{t-1},b_t,b_t^s$—the first two of which capture the history dependence in policymaking under commitment. Despite the complexity of these first-order conditions, the policymaker can fulfill this Ramsey program with a constant tax rate and no inflation by buying an appropriate quantity of single-period assets paid for by issuing longer term bonds. Shocks to public consumption then induce fluctuations in the prices of these assets/liabilities which perfectly insulate the government's finances.

With *i.i.d.* fluctuations in government spending, the current level of spending is also a state variable: we are either in the high- or in the low-government spending regime and may exit that regime with a probability of 1/2 each period.

The pricing equation for geometrically declining coupon bonds is

$$P_t^m = \beta E_t (1 + \rho P_{t+1}^m) \left(\frac{c_{t+1}}{c_t} \right)^{-\sigma} \nu_{t+1}$$

With government spending fluctuating between high and low states, bond prices will fluctuate depending on the spending state. Define $u_{ij} = \frac{u'(1-g^i)}{u'(1-g^j)} = \frac{(1-g^i)^{-\sigma}}{(1-g^j)^{-\sigma}}$, $i,j = l,h$, and $i \neq j$ bond prices in spending regime i, $i = h,l$ are given by

$$\begin{aligned} P_i^m &= \beta \frac{1}{2} (1 + \rho P_i^m) + \beta \frac{1}{2} (1 + \rho P_j^m) u_{ji} \\ &= A_i + B_i P_j^m \end{aligned}$$

where $A_i = (1 - \frac{1}{2}\beta\rho)^{-1} (\frac{1}{2}\beta + \frac{1}{2}\beta u_{ji})$ and $B_i = (1 - \frac{1}{2}\beta\rho)^{-1} \frac{1}{2}\beta\rho u_{ji}$, $i, j = l, h$, and $i \neq j$, which can be solved as

$$P_i^m = \frac{A_i + B_i A_j}{1 - B_i B_j}$$

For one-period debt this reduces to

$$P_i^s = \frac{1}{2}\beta + \frac{1}{2}\beta u_{ji}$$

Optimal hedging uses these fluctuations in bond prices to construct portfolio of government debt that negates the need to vary taxes or induce inflation surprises, despite the random movements in government consumption.

The flow budget identity conditional on the government spending regime, but with constant tax rates and no inflation, is

$$P_i^m b^m + P_i^s b^s = (1 + \rho P_i^m) b^m + b^s - (\tau^* - g^i - z^*)$$

We choose b^m and b^s to ensure this equation holds regardless of the government spending regime, so that the government does not need to issue or retire debt as it moves between low and high spending regimes. This portfolio is given by

$$\begin{bmatrix} b^m \\ b^s \end{bmatrix} = - \begin{bmatrix} P_i^m (1-\rho) - 1 & P_i^s - 1 \\ P_j^m (1-\rho) - 1 & P_j^s - 1 \end{bmatrix}^{-1} \begin{bmatrix} \tau^* - g^i - z^* \\ \tau^* - g^j - z^* \end{bmatrix}$$

We can achieve the same portfolio by considering the debt valuation equation in a given period, which is contingent on the government spending state. If government spending is currently high, that equation is

$$b^s(u'(c^h)) + b^m(u'(c^h)) + \sum_{j=1}^{\infty}(\rho\beta)^j\left[\frac{1}{2}u'(c^l) + \frac{1}{2}u'(c^h)\right]b^m$$
$$= u'(c^h)(\tau^* - g^h - z^*) + \sum_{j=1}^{\infty}\beta^j\left[\frac{1}{2}u'(c^l)(\tau^* - g^l - z^*) + \frac{1}{2}u'(c^h)(\tau^* - g^h - z^*)\right]$$

and if government spending is low it is

$$b^s(u'(c^l)) + b^m(u'(c^l)) + \sum_{j=1}^{\infty}(\rho\beta)^j\left[\frac{1}{2}u'(c^l) + \frac{1}{2}u'(c^h)\right]b^m$$
$$= u'(c^l)(\tau^* - g^l - z^*) + \sum_{j=1}^{\infty}\beta^j\left[\frac{1}{2}u'(c^l)(\tau^* - g^l - z^*) + \frac{1}{2}u'(c^h)(\tau^* - g^h - z^*)\right]$$

subtracting one from the other implies

$$[b^s + b^m](u'(c^h) - u'(c^l)) = u'(c^h)(\tau^* - g^h - z^*) - u'(c^l)(\tau^* - g^l - z^*) \tag{40}$$

Without any change in taxation or inflation, government solvency is ensured, provided that debt maturing in the current period has the value implied by this equation. Assuming a sufficiently low level of net indebtedness, the primary budget will swing between deficit and surplus as government spending moves from high to low regimes, implying that the right side of (40) is negative. Since $u'(c^h) > u'(c^l)$, this condition requires that the Ramsey policymaker buys short-term assets to such an extent that $b^s < -b^m$. The budget identity is insulated from the effects of government spending shocks, which can be absorbed by bond prices without any need to issue new debt, change taxes, or generate inflation surprises.

The size of the longer term liabilities must, equivalently, satisfy the solvency conditions conditional on the current level of government consumption. For example

$$b^s(u'(c^h)) + b^m(u'(c^h)) + \sum_{j=1}^{\infty}(\rho\beta)^j\left[\frac{1}{2}u'(c^l) + \frac{1}{2}u'(c^h)\right]b^m$$
$$= u'(c^h)(\tau^* - g^h - z^*) + \sum_{j=1}^{\infty}\beta^j\left[\frac{1}{2}u'(c^l)(\tau^* - g^l - z^*) + \frac{1}{2}u'(c^h)(\tau^* - g^h - z^*)\right]$$

which can be written as

$$\frac{\rho\beta}{1-\rho\beta}\left[\frac{1}{2}u'(c^l) + \frac{1}{2}u'(c^h)\right]b^s + b^s u'(c^h) + b^m u'(c^h)$$
$$= \frac{\beta}{1-\beta}\left[\frac{1}{2}u'(c^l)(\tau^* - g^l - z^*) + \frac{1}{2}u'(c^h)(\tau^* - g^h - z^*)\right] + u'(c^h)(\tau^* - g^h - z^*)$$

This expression can either define the steady-state level of long-term debt given the tax rate or the tax rate given the long-term debt stock. Either interpretation is consistent with a steady-state solution to the Ramsey tax-smoothing plan where the solution of the remainder of the Ramsey problem is $\tau_t = \tau^*$, $z_t = z^* = 0.18y$, $\nu_t = 1$, $\frac{\gamma_i}{P_i^s} = \tau^* b^s$, $\frac{\mu_i}{P_i^m} = \tau^* b^m$, $i = h, l$ with probability of 1/2. In other words, the steady-state tax rate can support the average level of government spending, steady-state transfers, and the steady-state net debt stock, while fluctuations in bond prices mitigate the need for further tax adjustments to compensate for fluctuations in government spending.

Fig. 10 reveals the pattern of bond returns and the underlying asset positions for a series of random draws across the two spending regimes. The figure's bottom right panel describes a particular realization of the government spending shocks. Despite these movements in spending the budget identity can be satisfied with a constant tax rate and no inflation surprises by buying short-term assets that are funded by issuing longer term debt. The portfolio that achieves this implies that the government holds short-term assets of around 22% of GDP, with longer term liabilities of around 70% of GDP and a net debt of around 48%. Although large, these positions are less than those typically found for richer stochastic processes, where positions often exceed the economy's total endowment by several factors (Buera and Nicolini, 2004). Since the ability to hedge relies on variation in the yield curve, having longer term liabilities to set against the short-term assets is most effective. Then a portfolio of single-period assets matched with 1-year liabilities requires far more short-term assets, compared to a portfolio made up of the same assets and bonds

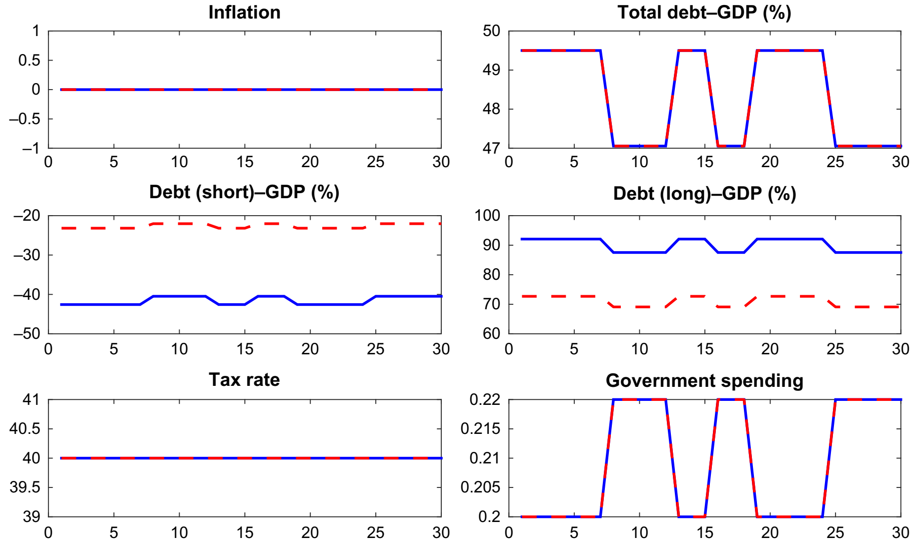


Fig. 10 Optimal hedging under commitment. 1-year debt (*solid lines*) and 5-year debt (*dashed lines*).

with an average maturity of 5 years. Hedging in this way implies that a positive shock to government spending, which raises the primary deficit, actually leads to a reduction in the value of government indebtedness, rather than to an increase. This is a general prediction of models that have achieved financial market completeness which Marcet and Scott (2009) use as the basis of an empirical test, but the data strongly reject.

Faraglia et al. (2008) extend Buera and Nicolini's analysis to move away from an endowment economy to consider a production economy with capital. This makes the size of the extreme portfolio positions even larger, and now the liability/asset positions are no longer constant, but highly volatile, possibly even reversing the issue-long-buy-short recommendation. Because yield premia are not very volatile, they are therefore not very effective as a source of insurance. They then consider what happens if the government is unsure about the specification of some element of the model. The sensitivity of results to small changes in model specification means that it is often better to run a balanced budget than run the risk of getting the portfolio composition wrong. Similarly, even modest transaction costs would make it undesirable to construct such huge portfolios.

4.7 Discretion

A large part of the literature that extends Lucas and Stokey's (1983) analysis focuses on the importance of having access to state-contingent debt either directly or by using inflation surprises and debt management to render state dependent the real payoffs from government debt. When the policymaker can replicate the Ramsey policy in Lucas and Stokey through such devices, there remains the issue of whether the underlying policy is time consistent. In the original Lucas and Stokey model, the Ramsey policy can be made time consistent by adhering to a particular debt maturity structure. Lucas and Stokey then conjecture that allowing debt to be nominal would make the policy problem trivial: positive debt would be costlessly deflated by positive surprise inflation and negative debt would be adjusted by surprise deflation to the level sufficient to support the first-best allocation (the interest on the debt paying for government consumption, consistent with any fiscal taxes/ subsidies required by offset other market distortions). This reasoning suggests that the only interesting case is when the outstanding debt stock is zero.

Persson et al. (1987) initiated a debate exploring the Lucas and Stokey conjecture.[al] Alvarez et al. (2004) conclude that the Lucas and Stokey structure of state-contingent indexed debt, in combination with a condition that net nominal debt is zero so that government debt liabilities equal the stock of money, can ensure the time consistency of the original Lucas and Stokey Ramsey policy in a monetary economy that follows the Friedman rule. As Persson et al. (2006) note, these conditions essentially reduce the monetary version of the Lucas and Stokey economy to its real version.

[al] Persson et al. (2006) chart the course of this debate.

Bohn (1988) argues that in issuing nominal debt the policymaker trades off the ability to use inflation surprises as a hedging device when debt is nominal against the inflation bias that a positive stock of debt creates. In models where the problem is not constructed to mimic the Lucas and Stokey Ramsey policy, the time-consistent policy typically implies a mean reverting steady-state level of debt. Debt can be positive or negative, depending on the nature of the time-inconsistency problem. The issue of the time consistency of policy is also dependent on the cost of inflation surprises. Persson et al. (2006) use beginning- rather than end-of-period money balances in the provision of liquidity services to make unexpected inflation costly, which allows them to construct a time-consistent portfolio of indexed and nominal debt. Martin (2009) adopts the cash–credit good distinction in Lucas and Stokey to generate a cost to inflation which is then balanced against the gains from using inflation to reduce the value of single-period nominal debt. This generates a mean reverting steady-state level of debt under discretion, rather than the random walk in steady-state debt, which is a feature of the Ramsey tax-smoothing policy without state-contingent debt. Martin (2011) combines the Lagos and Wright (2005) monetary search model with fiscal policy and explores the time-consistency problem to find that the welfare costs of an inability to commit are small. This conclusion likely reflects the nature of the costs of surprise inflation; as noted earlier, when Schmitt-Grohé and Uribe (2004) introduce even a tiny degree of nominal inertia, the time-inconsistent Ramsey policy tilts very firmly in favor of price stability, away from the Friedman rule and the use of inflation surprises.

We now turn to consider the impact on the balance between monetary and fiscal policy of constraining the policymaker to be time consistent. We continue to use the endowment economy where inflation is assumed to be costly as a shortcut to introducing nominal inertia.

The policymaker cannot make credible promises about how they will behave in the future in order to improve policy trade-offs today. However, even in this simple model there is an endogenous state variable in the form of government debt, so that policy actions today will affect future expectations through the level of debt that the policy bequeaths to the future. We define the auxiliary variable

$$M(b_{t-1},g_{t-1}) = (1+\rho P_t^m)\nu_t(c_t)^{-\sigma}$$

to write the Bellman equation of the associated policy problem as

$$\begin{aligned} V(b_{t-1},g_{t-1}) = &-\frac{1}{2}(\tau_t^2+\theta(\nu_t-1)^2)+\beta E_t V(b_t,g_t) \\ &+\mu_t\left(\beta\frac{c_t^\sigma}{P_t^m}E_t M(b_t,g_t)-1\right) \\ &+\lambda_t(b_t P_t^m-(1+\rho P_t^m)b_{t-1}\nu_t-g_t-z_t+\tau_t) \end{aligned}$$

We have replaced the expectations in the bond-pricing equation with the auxiliary variable to indicate that the policymaker cannot influence those expectations directly by making policy commitments. But those expectations are a function of the state variables. We take government spending and transfers to be exogenous autoregressive processes.

The implies the first-order conditions

$$
\begin{aligned}
\tau_t &: -\tau_t + \lambda_t = 0 \\
\nu_t &: -\theta(\nu_t - 1) - \lambda_t(1 + \rho P_t^m) b_{t-1} = 0 \\
P_t^m &: -\frac{\mu_t}{P_t^m} + \lambda_t(b_t - \rho b_{t-1}\nu_t) = 0 \\
b_t &: \frac{\mu_t}{P_t^m} c_t^\sigma \beta E_t \frac{\partial M(b_t, g_t)}{\partial b_t} + \lambda_t P_t^m + \beta E_t \frac{\partial V(b_t, g_t)}{\partial b_t} = 0
\end{aligned}
$$

From the envelope theorem

$$
\frac{\partial V(b_{t-1}, g_{t-1})}{\partial b_{t-1}} = -(1 + \rho P_t^m)\nu_t \lambda_t
$$

which can be led one period and substituted into the first-order condition for government debt

$$
\frac{\mu_t}{P_t^m} c_t^\sigma \beta E_t \frac{\partial M(b_t, g_t)}{\partial b_t} + \lambda_t P_t^m - \beta E_t(1 + \rho P_{t+1}^m)\nu_{t+1}\tau_{t+1} = 0
$$

Combining the condition for the bond price P_t^m with the Fisher equation implies

$$
\frac{\mu_t}{P_t^m} = \lambda_t(b_t - \rho \nu_t b_{t-1})
$$

which can be used to eliminate $\frac{\mu_t}{P_t^m}$ from the condition for debt. The system to be solved for $\{P_t^m, \nu_t, \tau_t, b_t, g_t\}$ is

$$
\begin{aligned}
\nu_t &: -\theta(\nu_t - 1) - \tau_t(1 + \rho P_t^m) b_{t-1} = 0 \\
b_t &: \tau_t(b_t - \rho \nu_t b_{t-1}) \beta c_t^\sigma E_t \frac{\partial M(b_t, g_t)}{\partial b_t} + \tau_t P_t^m - \beta E_t(1 + \rho P_{t+1}^m)\nu_{t+1}\tau_{t+1} = 0
\end{aligned}
$$

along with the bond-pricing equation and the government's budget constraint.

The first-order condition for inflation is now

$$
-\theta(\nu_t - 1) = (1 + \rho P_t^m) b_{t-1} \tau_t
$$

Under commitment, inflation persisted only for as long as the maturity structure of the predetermined debt stock at the time a shock hit. Under time-consistent policy, outside of the policymaker's bliss point (of zero inflation and no taxation), with a nonzero debt stock there will always be a state-dependent mix of taxation and inflation. A positive stock of debt delivers positive inflation, regardless of the maturity structure of that debt. This reflects the inflation bias inherent in the time-consistent policy in the presence of nominal debt.

We can see some more differences between discretion and commitment by contrasting the equivalent expressions describing the evolution of the tax rate. Under commitment we obtain the standard tax-smoothing result adjusted for the tilting implied by variations in the stochastic discount factor

$$\tau_t P_t^m = \beta E_t (1 + \rho P_{t+1}^m) \nu_{t+1} \tau_{t+1}$$

The equivalent condition under discretion is

$$\tau_t P_t^m = \beta E_t (1 + \rho P_{t+1}^m) \nu_{t+1} \tau_{t+1} - \tau_t (b_t - \rho \nu_t b_{t-1}) \beta c_t^\sigma E_t \frac{\partial M(b_t, g_t)}{\partial b_t}$$

The additional term captures the effects of the tax rate on expectations of inflation and bond prices through the level of debt carried into the future. Increased debt raises expected inflation and lowers expected bond prices, so $E_t \frac{\partial M(b_t, g_t)}{\partial b_t} < 0$. This captures the debt-contingent nature of the time-consistency problem facing the policymaker. As debt levels rise the policymaker faces a greater temptation to utilize surprise inflation to reduce the debt burden. Economic agents anticipate this and raise their inflationary expectations until the temptation to induce surprises is offset. However, unlike in the standard analysis of the inflationary bias problem this bias is not static since the policymaker can raise additional distortionary taxes to reduce debt and its associated inflation. Therefore the additional term in the above expression raises the tax rate above the level implied by the tax-smoothing condition observed under commitment. Where the tax rate under commitment was carefully constructed to allow debt levels to permanently rise, under discretion the tax rate prevents debt from rising permanently.[am] Moreover, the rate at which the policymaker reduces debt under discretion depends crucially on the term, $(b_t - \rho \nu_t b_{t-1})$ which in turn depends on the maturity structure of the debt stock. Effectively the lower bond prices mean the policymaker must issue more bonds to finance a given deficit, but pays less to buy back the existing debt stock. As debt maturity is increased this latter effect comes to dominate the former and the speed of debt reduction is reduced. Therefore, in contrast to the random walk in steady-state debt observed under commitment, the time-consistent policymaker returns debt to a steady-state value that is very close to zero, but slightly negative where the speed of adjustment depends crucially on average debt maturity. This cannot be seen entirely analytically, so we

[am] Calvo and Guidotti (1992) label this the "debt aversion" effect and Leith and Wren-Lewis (2013) call it the "debt stabilization bias."

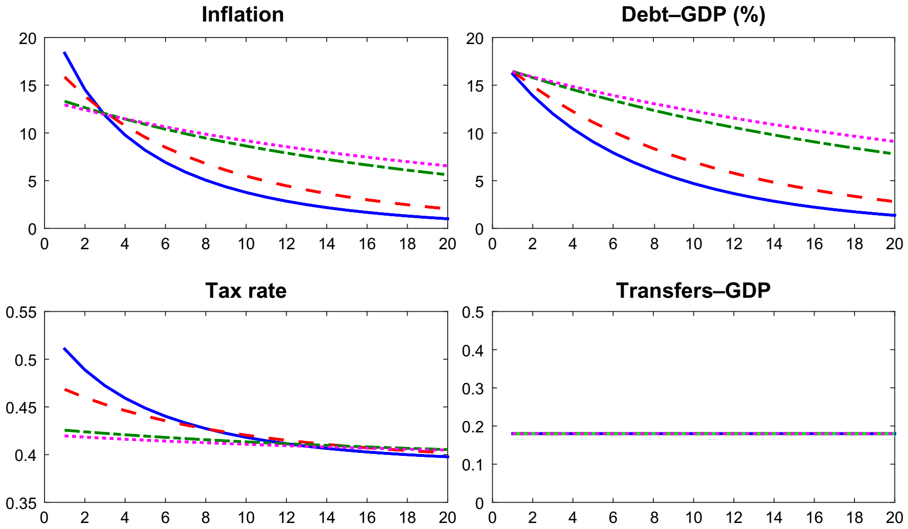


Fig. 11 Optimal time-consistent policy when debt is above its steady-state level. 1-period debt (*solid lines*), 1-year debt (*dashed lines*), 5-year debt (*dotted–dashed lines*), and 7-year debt (*dotted lines*).

need to analyze the numerical solution to the time-consistent policy problem to gain further insight.

The numerical solution under discretion is radically different from that under commitment (Fig. 11). Under commitment, policy allows the steady-state level of debt to follow a random walk and the use of inflation to offset shocks is relatively modest. Under discretion there is a unique steady state at which the policy supporting the steady-state debt level is time consistent, and this occurs at a slightly negative debt stock with a mild deflation. The negative steady-state debt stock falls far short of the negative debt levels that would be needed to support the first-best allocation—that is, the stock of government-held assets generates interest income sufficient to pay for all transfers and government spending without levying any distortionary taxes. Private-sector expectations ensure that the policymaker does not accumulate such a level of assets. Bondholders know that once the government has accumulated a positive stock of assets, it has an incentive to introduce surprise deflation to increase the real value of those assets. This knowledge reduces agents' inflation expectations until the policymaker no longer wishes to introduce such deflationary surprises. Accumulating more assets would then worsen this incentive to deflate confronting the policymaker with a trade-off between accumulating assets to reduce tax rates and the expected deflation that the accumulation of assets implies. In the steady state a balance is struck with a mild deflation and small negative debt stock, although both are extremely close to zero.

At positive debt levels there is a significant desire to reduce debt through inflation surprises. Economic agents anticipate this and raise their inflationary expectations.

Positive debt levels raise inflation in a highly nonlinear way because they introduce a state-dependent inflationary bias which can be very large. Even modest debt-to-GDP ratios can imply double-digit inflation. This is a surprising outcome since the same model and parameterization under commitment imply no inflation at all in the absence of shocks and only small inflation with shocks and positive debt levels.

As noted earlier, the policymaker's desire to mitigate this bias leads to a deviation from tax smoothing where the policymaker raises distortionary taxation above the tax-smoothing level to not only stabilize debt but reduce it toward its steady-state value. Debt maturity lessens this debt stabilization bias problem so that for a given debt-to-GDP ratio inflation is lower, the longer is debt maturity. The debt stabilization bias is heavily dependent on the magnitude of the government debt stock. When debt is high, the efficacy of surprise inflation—either current inflation or through bond prices—is also much higher and this raises the government's incentives to use this device to stabilize debt. As a result the debt stabilization bias rises dramatically with debt levels.

In the absence of innovations to the fiscal surplus, this higher inflation does not actually stabilize debt. As in the original inflation bias problem, there is a pure cost in the form of higher inflation which does not generate any reduction in debt.[an] But unlike the original inflation bias problem, in our case the magnitude of the bias is endogenous and depends on the size and maturity of the government debt. The policymaker can choose to reduce debt through taxation to gradually reduce the bias. Under discretion the reduction in debt can be a quite rapid, particularly when the debt stock is large and of short maturity. The costs of the policymaker being unable to commit in this context are not that debt is unstable, but that the policymaker too aggressively returns government debt to its steady-state level following shocks. This message resonates when thinking about actual fiscal austerity policies in many countries after the 2008 global financial crisis.

4.8 Debt Management under Discretion

The above results highlight the time-consistency issues created by nominal debt. The existing optimal policy literature also considers time-consistency issues in relation to debt management issues. Specifically, in the Lucas and Stokey model with state-contingent debt, the maturity structure is key in ensuring that the Ramsey policy described in Lucas and Stokey is time consistent. At the same time, the optimal hedging analysis shows that the maturity structure can create a portfolio of government bonds that features the right state-contingent payoffs even when the underlying bonds are not state contingent. In the context of a real model, Debortoli et al. (2014) also allow the government to hold

[an] Analogously, in Barro and Gordon (1983) this additional inflation does not reduce unemployment.

short-term assets and longer term liabilities (which are individually not state contingent), but require the policy to be time-consistent. They show that the optimal policy results in a relatively flat maturity structure that offsets the costs of not being able to commit even though this removes the tilting in maturity that is beneficial in terms of insurance effects.

To assess the trade-offs between optimal hedging and time consistency, we use the same model that delivered complete hedging of government expenditure shocks under commitment and solve that model under discretion. In introducing single-period bonds to the time-consistent policy problem we need to define an additional auxiliary variable

$$N(b_{t-1}, b_{t-1}^s, g_{t-1}) = \nu_t (c_t)^{-\sigma}$$

All expectations are now a function of three state variables, longer term bonds, b_{t-1}, single-period bonds, b_{t-1}^s and government spending, g_{t-1}, which will either equal $0.22y$ in the high spending regime, or $0.2y$ in the low spending case.

The policy problem is

$$\begin{aligned} V(b_{t-1}, b_{t-1}^s, g_{t-1}) = & -\frac{1}{2}(\tau_t^2 + \theta(\nu_t - 1)^2) + \beta E_t V(b_t, b_t^s, g_t) \\ & + \mu_t (\beta \frac{c_t^\sigma}{P_t^m} E_t M(b_t, b_t^s, g_t) - 1) \\ & + \gamma_t (\beta E_t \frac{c_t^\sigma}{P_t^s} E_t N(b_t, b_t^s, g_t) - 1) \\ & + \lambda_t (b_t P_t^m + b_t^s P_t^s - (1 + \rho P_t^m) b_{t-1} \nu_t - b_{t-1}^s \nu_t - g_t + \tau_t - z_t) \end{aligned}$$

which has an additional constraint associated with the pricing of short-term bonds, and the government's flow budget identity contains both single-period and declining coupon bonds. After applying the envelope theorem this implies the first-order conditions. For inflation

$$-\theta(\nu_t - 1) = \tau_t [(1 + \rho P_t^m) b_{t-1} + b_{t-1}^s]$$

The level of inflation depends on the total level of indebtedness across short and long bonds, so that a positive level of net indebtedness implies an inflationary bias. As before, this bias serves no purpose in terms of reducing the real debt burden, but reflects economic agents' expectations that if inflation were any lower, the policy would be tempted to introduce a surprise inflation to facilitate debt reduction.

The tax-smoothing conditions are

$$\begin{aligned} \tau_t P_t^m = & \beta E_t (1 + \rho P_{t+1}^m) \nu_{t+1} \tau_{t+1} - \tau_t (b_t - \rho \nu_t b_{t-1}) \beta c_t^\sigma E_t \frac{\partial M(b_t, b_t^s, g_t)}{\partial b_t} \\ & - \tau_t b_t^s \beta c_t^\sigma E_t \frac{\partial N(b_t, b_t^s, g_t)}{\partial b_t} \end{aligned}$$

and

$$\tau_t P_t^s = \beta E_t \nu_{t+1} \tau_{t+1} - \tau_t (b_t - \rho \nu_t b_{t-1}) \beta c_t^\sigma E_t \frac{\partial M(b_t, b_t^s, g_t)}{\partial b_t^s}$$
$$- \tau_t b_t^s \beta c_t^\sigma E_t \frac{\partial N(b_t, b_t^s, g_t)}{\partial b_t^s}$$

The first two terms of these expressions reflect the same tax-smoothing conditions found under commitment, where the choice of short-term assets and longer term bonds could satisfy these conditions while perfectly insulating the government's finances from the fluctuations in government spending. The final two terms in each condition capture the impact that another unit of short or long debt has on long- and short-term bond prices through the impact of debt on inflation expectations. These effects highlight the incentives that the policymaker has to reduce indebtedness to reduce inflation, given the inflationary bias problem created by a positive stock of government debt. The magnitude of the effect of reducing either short- or long-term debt by one bond may vary depending on the relative proportions of the two bonds. In other words, by varying the relative proportions of single period and longer term debt, the policymaker can vary the average debt maturity and thereby influence the inflationary bias problem implied by a given level of indebtedness.

Solving the model without switching in government spending generates a steady state with near-zero debt and inflation (Fig. 12). Introducing government spending switches induces fluctuations in all variables. The movements in spending are largely matched with movements in tax rates (even though these could have been eliminated by issuing an

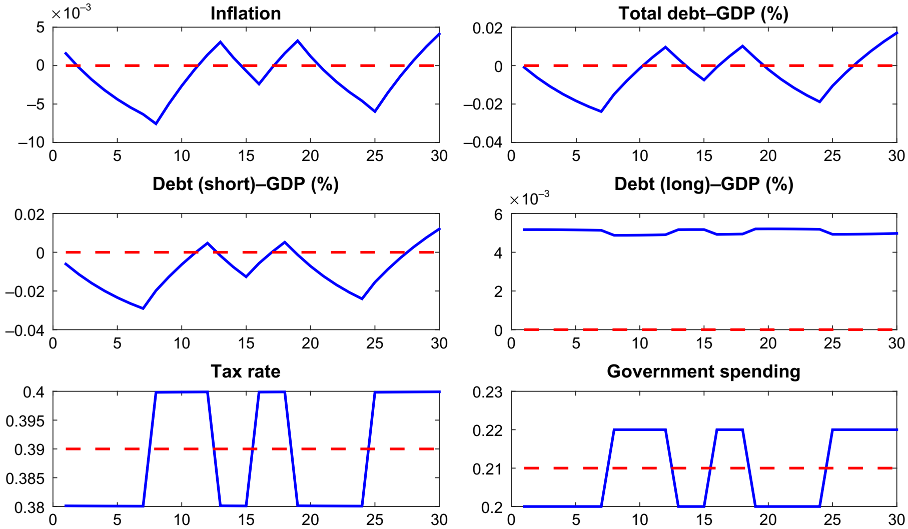


Fig. 12 Hedging under discretion. With government spending switching (*solid lines*) and without government spending switching (*dashed lines*).

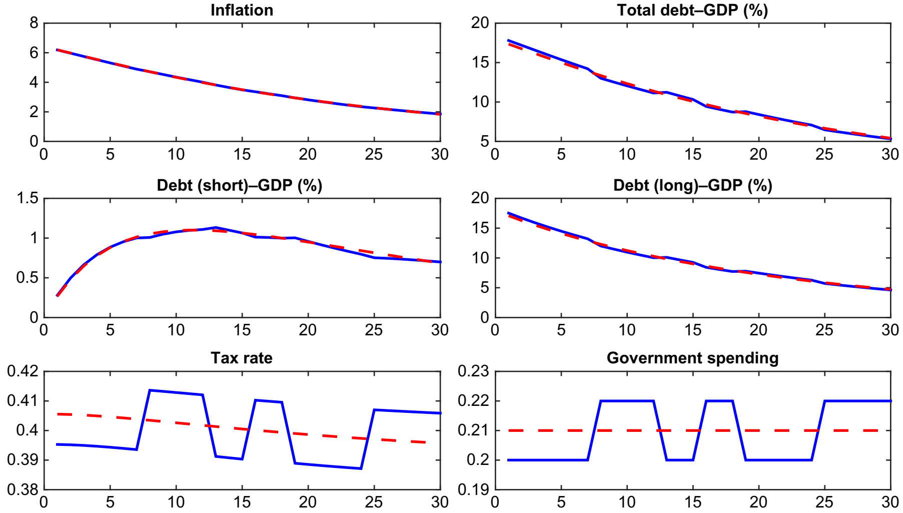


Fig. 13 Hedging and time-consistent policy. With government spending switching (*solid lines*) and without government spending switching (*dashed lines*).

appropriately constructed portfolio of short-term assets and longer term liabilities), although with some increase in the debt/deficit when we are in the high spending regime. The stochastic steady-state asset and liability positions are only slightly positive for assets, and slightly negative for liabilities, but quite distant from the magnitude of the positions required for perfect hedging. Inflation follows the level of indebtedness, giving rise to a positive (negative) inflation bias when the level of indebtedness is positive (negative).

Starting from a positive level of indebtedness, Fig. 13 plots the mix of short- and long-term debt as the economy transitions toward the stochastic steady-state. Calvo and Guidotti's (1992) debt aversion appears as the policymaker fairly rapidly reduces indebtedness in an attempt to eliminate the inflationary bias that debt induces. The fluctuations in debt induced by the changing spending regime are small relative to the general debt dynamics implied by the transition to steady state. The fact that the single-period debt does not rise dramatically when overall indebtedness increases implies that there is an effective lengthening of maturity as overall debt levels increase. This echoes the results of Calvo and Guidotti, which are also discussed in Missale (1999).

5. PRODUCTION ECONOMIES WITH OPTIMAL MONETARY AND FISCAL POLICIES

5.1 The Model

Until now our analysis of optimal policy has been based on a simple flexible price endowment economy, where we have captured the costs of inflation and distortionary taxation by adding quadratic terms in these variables to the policymaker's objective function. We now attempt to generalize these results by considering a production economy where households supply labor to imperfectly competitive firms who are subject to quadratic costs in changing prices as in Rotemberg (1982). The government levies a tax on sales to finance exogenous processes for transfers and government consumption. The policymaker aims to maximize the utility of the representative household. This section therefore endogenizes the welfare costs of both inflation and distortionary taxation. We also widen the scope for monetary and fiscal policy interactions because monetary policy not only generates revaluations of government bonds but also affects real debt service costs and the size of the tax base. Changes in distortionary taxation not only influence the government's budget identity, but they also affect production decisions and have a direct cost-push effect on inflation.

This basic setup is similar to that in Benigno and Woodford (2004) and Schmitt-Grohé and Uribe (2004) but with some differences.[ao] We model price stickiness using Rotemberg's (1996) adjustment costs rather than Calvo (1983) pricing because this reduces the number of state variables when solving the model nonlinearly. We also consider a richer maturity structure rather than single-period bonds.

5.1.1 Households

There is a continuum of households of size one. We assume complete asset markets so that through risk sharing households face the same budget constraint. The typical household seeks to maximize

$$E_0 \sum_{t=0}^{\infty} \beta^t \left(\frac{c_t^{1-\sigma}}{1-\sigma} - \frac{N_t^{1+\varphi}}{1+\varphi} \right)$$

where c and N are a consumption aggregate and labor supply, respectively. The consumption basket is made up of a continuum of differentiated products, $c_t = (\int_0^1 c(j)_t^{\epsilon-1/(\epsilon)} dj)^{\epsilon/(\epsilon-1)}$, and the basket of public consumption takes the same form.

The budget constraint at time t is given by

$$\int_0^1 P_t(j) c_t(j) dj + P_t^m B_t^m = \Pi_t + (1 + \rho P_t^m) B_{t-1}^m + W_t N_t + Z_t \tag{41}$$

where $P_t(j)$ is the price of variety j, Π is the representative household's share of profits in the imperfectly competitive firms (after tax), W are wages, and Z are lump-sum transfers and the bonds the household can invest in are the geometrically declining coupon bonds used above.

We maximize utility subject to the budget constraint (41) to obtain the optimal allocation of consumption across time and the associated pricing of declining coupon bonds

[ao] Leeper and Zhou (2013) study a linear-quadratic version of this setup.

$$\beta E_t\left[\left(\frac{c_t}{c_{t+1}}\right)^{\sigma}\left(\frac{P_t}{P_{t+1}}\right)\left(1+\rho P_{t+1}^{m}\right)\right]=P_t^{m}$$

Notice that when these reduce to single-period bonds, $\rho=0$, the price of these bonds is $P_t^m=R_t^{-1}$.

The second first-order condition relates to the labor supply decision

$$\left(\frac{W_t}{P_t}\right)=N_t^{\varphi}c_t^{\sigma}$$

5.1.2 Firms

Firms produce output using to a linear production function, $y(j)_t=AN(j)_t$, where $a_t=\ln(A_t)$ is time varying and stochastic, such that the real marginal costs of production are $mc_t=\frac{W_t}{P_tA_t}$. Household demand for their product is given by $y(j)_t=\left(\frac{P(j)_t}{P_t}\right)^{-\epsilon}y_t$ and firms are also subject to quadratic adjustment costs in changing prices

$$v_t^jP_t=\frac{\phi}{2}\left(\frac{p_t(j)}{\pi^{*}p_{t-1}(j)}-1\right)^2P_ty_t$$

where $\pi^{*}=1$ is the steady-state gross inflation rate. In a symmetric equilibrium where $p_t(j)=P_t$ the first-order condition for firms' profit maximization implies

$$(1-\theta)(1-\tau_t)+\theta mc_t-\phi\frac{\pi_t}{\pi^{*}}\left(\frac{\pi_t}{\pi^{*}}-1\right)+\phi\beta E_t\left(\frac{c_t}{c_{t+1}}\right)^{\sigma}\frac{\pi_{t+1}}{\pi^{*}}\frac{y_{t+1}}{y_t}\left(\frac{\pi_{t+1}}{\pi^{*}}-1\right)=0$$

which is the nonlinear version of the Phillips curve and includes the effects of a distortionary tax on sales revenues, τ_t.

5.1.3 Equilibrium

Goods market clearing requires, for each good j

$$y(j)_t=c(j)_t+g(j)_t+v(j)_t$$

which allows us to write

$$y_t\left[1-\frac{\phi}{2}\left(\frac{\pi_t}{\pi^{*}}-1\right)^2\right]=c_t+g_t$$

There is also market clearing in the bonds market where the longer term bond portfolio evolves according to the government's budget identity which we now describe.

5.1.4 Government Budget Identity

Combining the series of the representative consumer's flow budget constraints, (41), and noting the equivalence between factor incomes and national output, we obtain the government's flow budget identity

$$P_t^m b_t = (1+\rho P_t^m)\frac{b_{t-1}}{\pi_t} - y_t\tau_t + g_t - z_t$$

where real debt is defined as $b_t \equiv \frac{B_t^M}{P_t}$.

5.2 Commitment Policy in the New Keynesian Model

Setting up the Lagrangian

$$\begin{aligned} L_t \; &= E_0\sum_{t=0}^{\infty}\beta^t\Bigg[\left(\frac{c_t^{1-\sigma}}{1-\sigma}-\frac{N_t^{1+\varphi}}{1+\varphi}\right)+\lambda_{1t}\left(y_t\left(1-\frac{\phi}{2}\left(\frac{\pi_t}{\pi*}-1\right)^2\right)-c_t-g_t\right) \\ &\quad +\lambda_{2t}\left(\beta\left(\frac{c_t}{c_{t+1}}\right)^{\sigma}\left(\frac{P_t}{P_{t+1}}\right)(1+\rho P_{t+1}^m)-P_t^m\right) \\ &\quad +\lambda_{3t}\left((1-\theta)(1-\tau_t)+\theta y_t^{\varphi}c_t^{\sigma}A_t^{-1-\varphi}-\phi\pi_t(\pi_t-1)+\phi\beta\left(\frac{c_t}{c_{t+1}}\right)^{\sigma}\pi_{t+1}\frac{y_{t+1}}{y_t}(\pi_{t+1}-1)\right) \\ &\quad +\lambda_{4t}\left(P_t^M b_t-(1+\rho P_t^M)\frac{b_{t-1}}{\pi_t}+y_t\tau_t-g_t-tr_t\right)\Bigg] \end{aligned}$$

and differentiating with respect to $\{c_t, y_t, \tau_t, P_t^m, b_t^m, \pi_t\}$ yield the first-order conditions for the Ramsey program. Those conditions are sufficiently complex to afford little additional insight that was not already gained from the analysis of the comparable problem for our simple endowment economy. But when we solve the model numerically, several interesting results relating to the optimal monetary and fiscal policy mix emerge.

5.3 Numerical Results

The first experiment considers a transfers shock at different initial levels of debt (Fig. 14).[ap] Transfers start at 18% of GDP and then increase with an autocorrelated shock, but do not respond further to GDP. When, as in the first column, the initial debt level is zero the maturity structure of the debt issued after the shock has hit is irrelevant. There is an initial one-period burst in inflation caused by the rise in the tax rate and not fully offset by the tightening of monetary policy. Then a coordinated use of monetary and fiscal policy stabilizes debt at its new steady-state level. The tax rate does not jump immediately to its new steady state, but follows a dynamic path which captures the movement in the real interest rate in the sticky-price economy, while monetary policy ensures that inflation is zero outside of the initial period.

Moving to column 2, at a higher initial debt level radically different policy responses emerge that depend on debt levels and maturity structures. As in Leith and Wren-Lewis (2013) with single-period debt and a sufficiently high debt stock, the transfers shock

[ap] In all cases we solve the model nonlinearly under perfect foresight following an initial perturbation from the steady state.

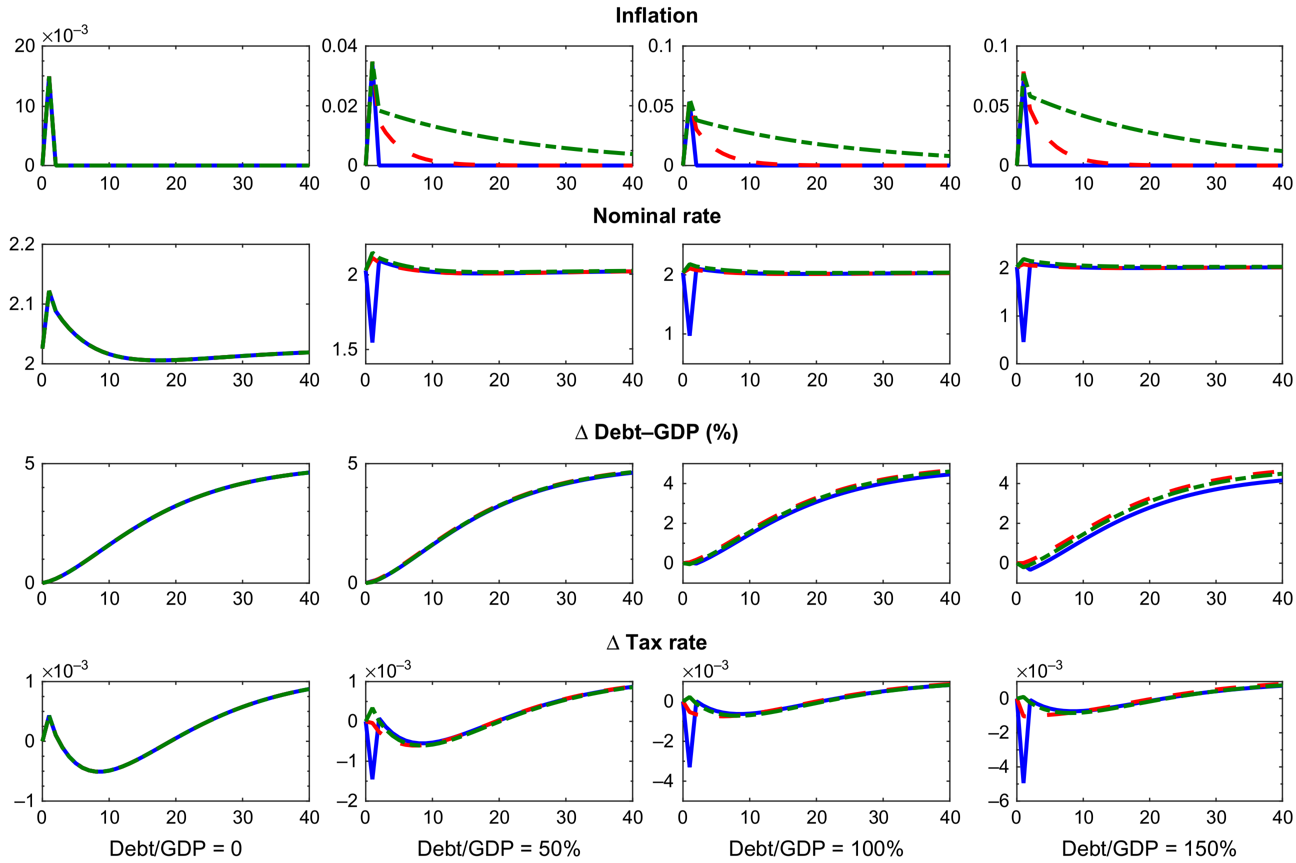


Fig. 14 Optimal policy response to higher transfers with different debt levels and maturities. 1-period debt (*solid lines*), 1-year debt (*dashed lines*), and 4-year debt (*dotted–dashed lines*).

results in the policymaker relaxing monetary policy to reduce debt service costs and fuel the initial burst in inflation. Monetary policy stabilizes the debt—just as in the fiscal theory—while tax rates fall to moderate the rise in inflation. Thereafter a combination of monetary and fiscal policy stabilizes the debt without generating any further inflation. When the debt is of longer term maturity (1 or 5 years), the initial policy response is quite different, with a tighter monetary policy and higher tax rates. The initial rise in inflation extends beyond the first period to help stabilize debt through reduced bond prices.

We now turn to the government spending shock in Fig. 15. The first column sets the initial tax rate at $\tau = 0.39$, sufficient to pay for both the initial value of transfers and public consumption, so there is no debt. In this case, as in the simple endowment economy, debt maturity does not matter and the policy response is the same regardless of the maturity of the debt. Unlike the endowment economy, there is surprise inflation, but this plays no direct role in stabilizing debt. Here the inflation reflects initial jumps in tax rates and interest rates that deliver the optimal balance between monetary and fiscal policy. There is a tax-smoothing jump in taxation that would fuel inflation, but which is offset by a tighter monetary policy that makes inflation zero after the initial period. As private consumption recovers, the tax rate rises, and ultimately there is a high tax rate to support an increased level of debt.

As we increase the initial level of debt, maturity structure generates differences in policy responses. As before, longer maturity delivers a smaller, but more sustained increase in inflation that stabilizes debt by reducing bond prices. But there are differences in the policy mix behind this result. When initial debt to GDP is just under 50%, with only single-period debt the policymaker actually cuts taxes to reduce the inflationary consequences of the government spending shock.

At higher initial debt, more radical differences in the policy mix arise across maturities. Sticky prices mean that not only surprises in the path of inflation influence debt dynamics: the policymaker can also influence real ex-ante interest rates and, through the Phillips curve, the size of the tax base. At a debt level near 100%, we observe a substantial fall in both tax rates and interest rates when debt is only single period. This amounts to a reversal of the conventional assignment of monetary and fiscal policy: monetary policy acts to stabilize debt by cutting real interest rates, while fiscal policy mitigates the inflationary consequences of this by reducing tax rates. For an average debt maturity of 5 years we retain the conventional assignment, with tax rates rising and monetary policy tightening to offset the rise in inflation that higher tax rates would generate.

5.4 An Independent Central Bank

Two key features of jointly optimal policy are worth highlighting. First, price-level control, which is typically a feature of optimal monetary policy in the new Keynesian model, is absent in the presence of fiscal policy and the associated tax-smoothing objective.

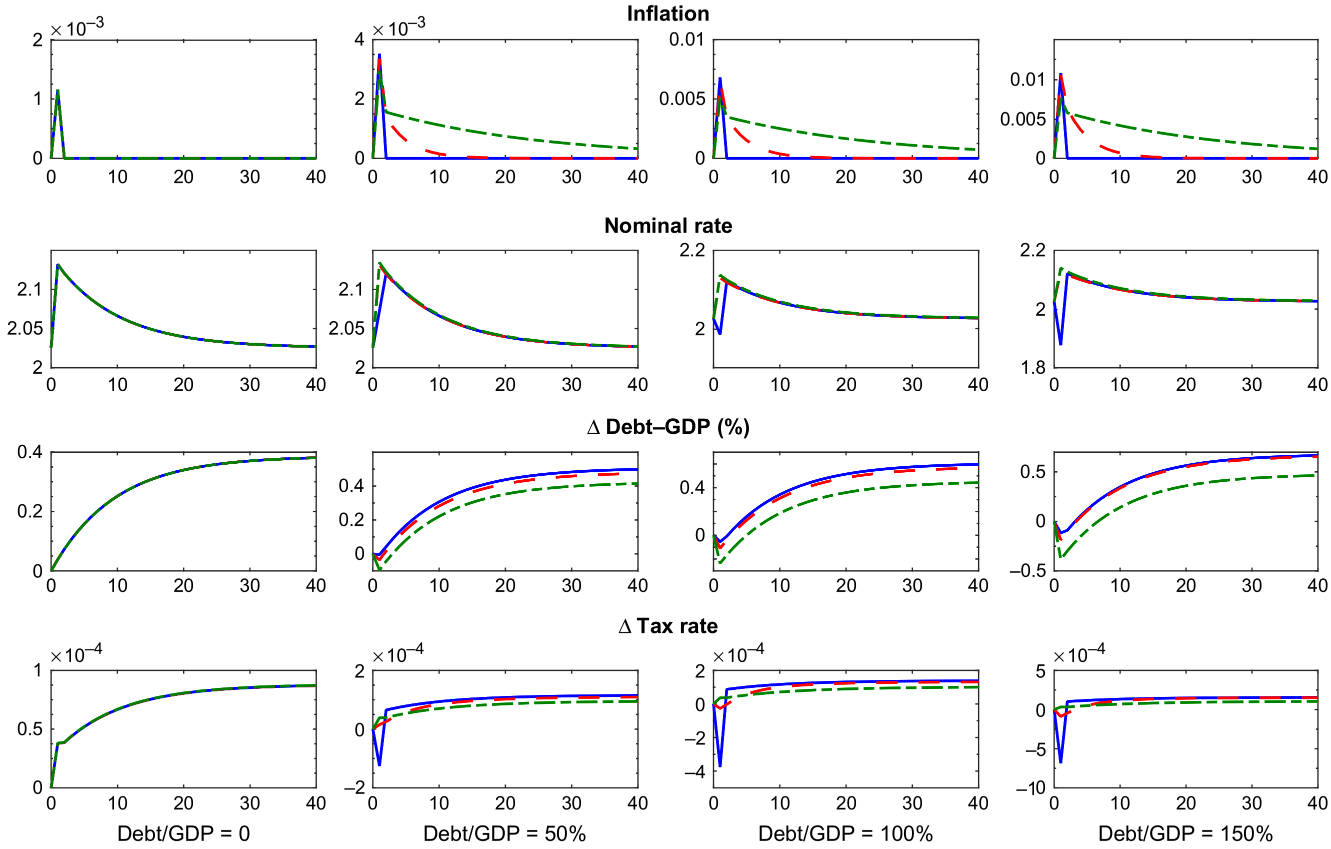


Fig. 15 Optimal policy response to an increase in government spending with different debt levels and maturities. 1-period debt (*solid lines*), 1-year debt (*dashed lines*), and 4-year debt (*dotted–dashed lines*).

Typical analyses have policymakers commit not only to return inflation to target after a shock hits but also to return the *price level* back to its preshock level. This commitment reduces inflation expectations and improves the trade-off between stabilization of inflation and the real economy. When fiscal policy enters the picture, the initial inflation becomes a desirable means of stabilizing debt through the revaluation effects that are a distinguishing feature of the fiscal theory.

Second, the policy mix depends on the size and maturity of government debt. With short maturity and high debt levels, optimal policy reverses the usual policy assignment—raising taxes and interest rates in the face of higher transfers or government consumption—and instead cuts interest rates to reduce debt interest dynamics and cuts taxes to offset the inflation that the relaxation in monetary policy would otherwise induce. Many economists would be uncomfortable with using monetary policy as a tool of fiscal stabilization in this way and would argue in favor of independent central banks to avoid this policy mix.

We assess the implications of independent monetary policy by deriving the optimal fiscal policy conditional on a given monetary policy rule. We assume that the central bank follows a simple Taylor rule with a coefficient on inflation of $\alpha_\pi = 1.5$. The fiscal authority faces the same optimization described earlier, but with the additional constraint that monetary policy follows this rule. Fig. 16 reports that the policy response to higher government spending exhibits some notable differences from the outcome when monetary and fiscal policies are jointly optimal. Inflation's increase is far more prolonged under an independent central bank. When monetary and fiscal policy operate cooperatively, even for the largest stock of debt we analyzed, inflation is less than half that observed when decoupling monetary from fiscal policy. This gives rise to the second surprising result. The active independent monetary policy results in the fiscal policymaker *cutting* rather than raising taxes in response to the government spending shock. The magnitude of the tax cut increases with the stock of debt, but does not vary much across maturities. Optimal fiscal policy counteracts the higher debt service costs that active monetary policy generates by cutting tax rates. This offsets the increase in inflation and under the policy rule mitigates the rise in real interest rates. Because this action is more important the higher the debt, the magnitude of the tax cuts increases with rising debt levels. Similar inflation paths across all debt levels imply that the value of longer maturity debt gets reduced through revaluation effects by more than the other maturities. This also has the implication that the spillovers from monetary policy shocks to the government's finances are likely to be greater at higher and longer maturity debt levels.

These results point to the ubiquity of a central feature of the fiscal theory—debt revaluation through surprise changes in inflation and bond prices. Whether policies are jointly optimal or optimal fiscal policy is constrained by an independent central bank, debt revaluation continues to characterize optimal policy behavior.

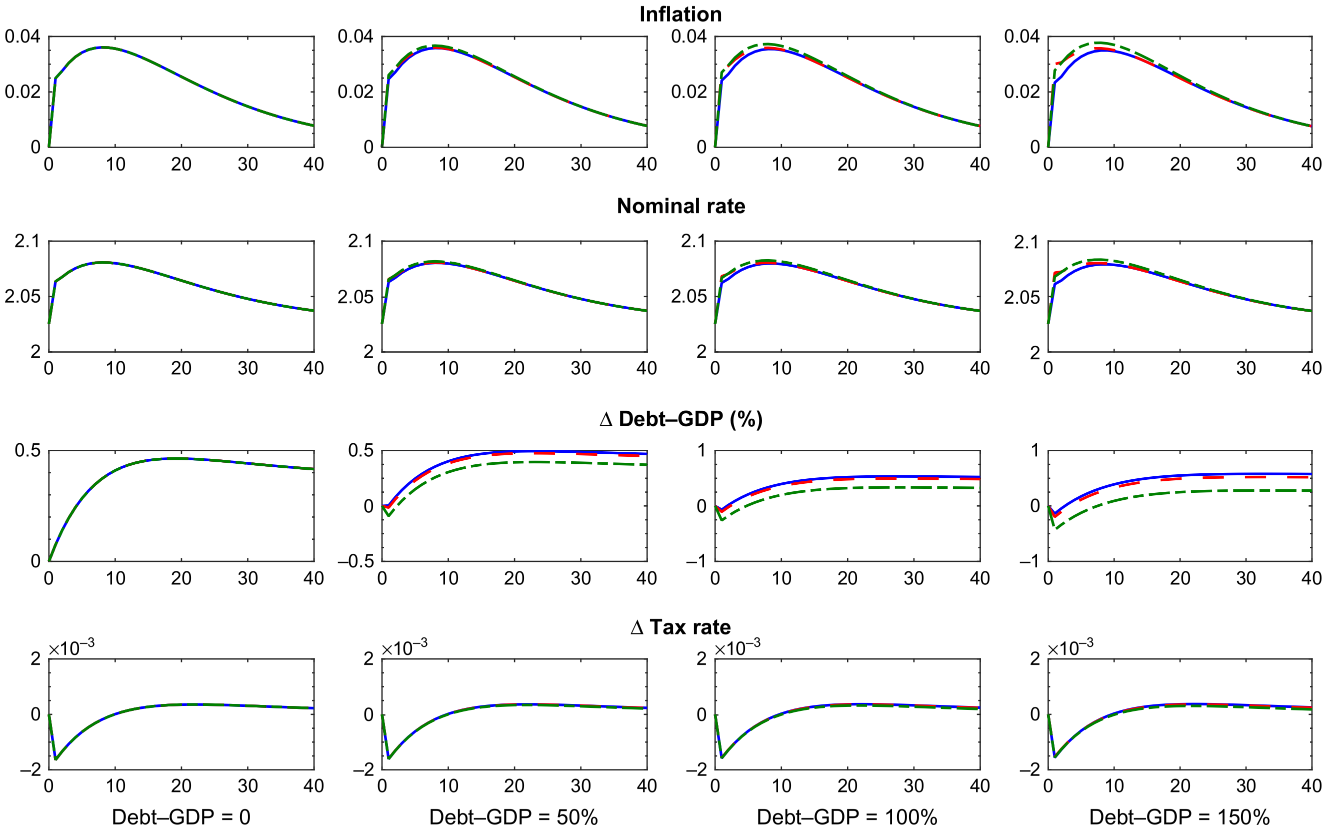


Fig. 16 Optimal fiscal policy response to an increase in government spending with an independent central bank. 1-period debt (*solid lines*), 1-year debt (*dashed lines*), and 4-year debt (*dotted–dashed lines*).

5.5 Discretion in the New Keynesian Economy

This subsection turns to optimal discretionary policy, following the setup in Leeper et al. (2015a). That setup employs a new Keynesian model in which the tax applies to labor income rather than sales revenue and government spending is treated as an endogenous policy instrument rather than an exogenous stream of purchases that need to be financed. There are no transfers. The policy under discretion is a set of decision rules for $\{c_t, y_t, \pi_t, b_t, \tau_t, g_t, P_t^M\}$ that maximize

$$V(b_{t-1}, A_t) = \max\left\{\frac{c_t^{1-\sigma}}{1-\sigma} + \chi\frac{g_t^{1-\sigma_g}}{1-\sigma_g} - \frac{(y_t/A_t)^{1+\varphi}}{1+\varphi} + \beta E_t[V(b_t, A_{t+1})]\right\}$$

subject to the resource constraint

$$y_t\left(1 - \frac{\phi}{2}\left(\frac{\pi_t}{\pi^*} - 1\right)^2\right) - c_t - g_t$$

the Phillips curve

$$(1-\epsilon) + \epsilon(1-\tau_t)^{-1} y_t^{\varphi} c_t^{\sigma} A_t^{-1-\varphi} - \phi\frac{\pi_t}{\pi^*}\left(\frac{\pi_t}{\pi^*} - 1\right) + \phi\beta c_t^{\sigma} y_t^{-1} E_t\left[c_{t+1}^{-\sigma}\frac{\pi_{t+1}}{\pi^*}\left(\frac{\pi_{t+1}}{\pi^*} - 1\right)\right] = 0$$

and the government's budget identity

$$\begin{aligned}
&\beta E_t\left[\left(\frac{c_t}{c_{t+1}}\right)^{\sigma}\left(\frac{P_t}{P_{t+1}}\right)\left(1+\rho P_{t+1}^M\right)\right] b_t \\
&= \left\{1 + \rho\beta E_t\left[\left(\frac{c_t}{c_{t+1}}\right)^{\sigma}\left(\frac{P_t}{P_{t+1}}\right)\left(1+\rho P_{t+1}^M\right)\right]\right\}\frac{b_{t-1}}{\pi_t} \\
&- \left(\frac{\tau_t}{1-\tau_t}\right)\left(\frac{y_t}{A_t}\right)^{1+\varphi} c_t^{\sigma} + g_t
\end{aligned}$$

where we have used the bond-pricing equation to eliminate the current value of the portfolio of bonds.

Leeper et al. (2015a) solve the nonlinear system consisting of seven first-order conditions and the three constraints to yield the time-consistent optimal policy using the Chebyshev collocation method. In contrast to the case of commitment where steady-state inflation is zero, discretion implies a steady state with a mildly negative debt stock and a mild deflation. Fig. 17 shows that starting from high debt levels produces significant policy differences across differing bond maturities. These impulse responses reflect the time-consistent adjustment from a high debt level to the ultimate steady-state debt level, which is slightly negative. The most notable element in these dynamic paths is the very high levels of inflation. This inflation does not serve to reduce the real value of debt; instead, it reflects the state-dependent inflationary bias problem generated by high

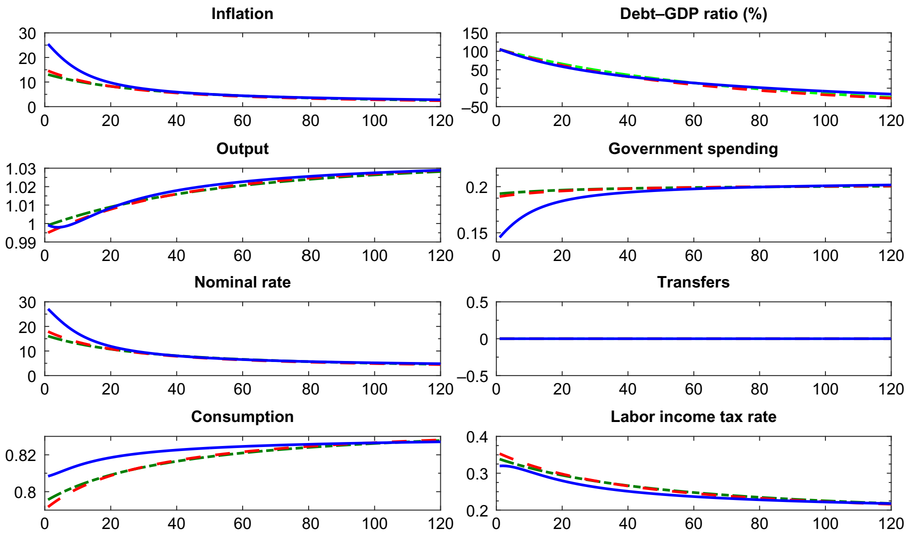


Fig. 17 New Keynesian model under discretionary policy. 1-year debt (*solid lines*), 5-year debt (*dashed lines*), and 8-year debt (*dotted–dashed lines*).

debt levels. When debt levels are raised, the policymaker faces a temptation to use surprise inflation or surprise reductions in bond prices to reduce the real value of government debt. Knowing this, economic agents raise their inflationary expectations until this temptation is no longer present. At empirically plausible debt levels, this temptation is very strong and very high rates of inflation are required to ensure the policy remains time consistent. The shorter the debt maturity, the greater the temptation to inflate and reduce debt levels quickly—what we label "the debt stabilization bias." The steady-state economy eventually achieves a small negative long-run optimal value for debt and a slight undershooting of the inflation target. This falls far short of the accumulated level of assets that would be needed to finance government consumption and eliminate tax and other distortions.

6. EMPIRICAL CONSIDERATIONS

The chapter's emphasis to this point reflects the bulk of the literature on the fiscal theory in its theoretical focus. This section discusses a set of empirical considerations that arise from work on monetary and fiscal interactions. First, we briefly explain why it is difficult to distinguish whether time series data were generated by regime M or by regime F. Then we turn to both reduced form and structural evidence about the prevailing policy regime, including work on regime-switching policies. We end the section by clarifying some common misperceptions about the nature of equilibrium under regime F.

6.1 Distinguishing Regimes M and F

It is well established that regimes M and F can generate equivalent (or nearly equivalent) equilibrium processes. Cochrane (1999) discusses this point and Woodford's (1999) comments on Cochrane's paper elaborate on the issue in some detail. Leeper and Walker (2013) display a simple theoretical example in which the two regimes are observationally equivalent.

Observational equivalence of the two regimes may be surprising. After all, Sections 2 and 3 went to great length to show that monetary and fiscal disturbances produce strikingly different dynamic responses in the two regimes. To understand the equivalence, consider the linearized new Keynesian model that Section 3.1 describes. That model's economic state in period t is the triple $X_t \equiv (\varepsilon_t^M, \varepsilon_t^F, \hat{b}_{t-1})$ and in regime M, each endogenous variable—including the policy variables $\hat{R}_t$ and $\hat{s}_t$—is a linear function of X_t in equilibrium. But those mappings from X_t to the policy variables are consistent with regime F policy behavior: the interest rate depends only on ε_t^M and the surplus depends on ε_t^F.[aq]

Some critics argue that this equivalence result renders the fiscal theory "untestable" and therefore empirically vacuous. Naturally, *equivalence* implies that the conventional view—regime M—is also "untestable." But the critics' nihilism is unwarranted. Observational equivalence merely implies that *in the absence of identifying restrictions* it is impossible to discern which regime produced observed data. But this is nearly a truism. No set of simple correlations—among debt, deficits, inflation, and interest rates—can tell us whether the underlying policy behavior comes from regime M or regime F.[ar]

Yet correlation-based "tests" of the fiscal theory abound in the literature. Canzoneri et al. (2001b) argue that if a positive shock to surpluses both raises future surpluses and lowers the real value of government debt, regime M prevails; if the positive surplus shock raises the value of debt, then regime F prevails. Cochrane (1999) succinctly explains why this is not a "test" of regime. Like any asset, government debt has both a "backward-looking" and a "forward-looking" representation. Let $b_t \equiv B_t/P_t$ denote the real market value of debt. Debt's law of motion—the budget identity—yields the backward view

$$b_{t+1} = r_{t+1}(b_t - s_t)$$

where $r_{t+1} \equiv R_t P_t / P_{t+1}$ is ex-post real return on bonds between t and $t + 1$ and s_t is the primary surplus at t. Higher s_t seems to imply a lower value for debt at $t + 1$. But the forward view, which determines the asset value of debt yields

[aq] If the economy starts with an initial level of debt, the $\{\hat{s}_t\}$ process must be chosen to be consistent with that level.

[ar] Much of the evidence that Friedman and Schwartz (1963a,b) compiled in favor of the quantity theory sought to show that erratic monetary policy drove nominal income movements. But that evidence came from efforts to *identify* "exogenous" or "autonomous" changes in the money stock, as Sims (1972) later showed. Friedman and Schwartz recognized that reduced form correlations alone cannot establish causality.

$$b_t = E_t \sum_{j=0}^{\infty} \left(\frac{1}{r}\right)^j s_{t+j} \tag{42}$$

to suggest that a persistent increase in surpluses raises the value of debt.[as] Evidently, manipulations of identities do not impose enough structure to distinguish between regimes.

A second branch of the correlation-based "testing" literature follows Bohn (1998) in using limited information techniques to estimate

$$s_t = \gamma b_{t-1} + \delta' Z_t + \varepsilon_t^F \tag{43}$$

where s_t is the primary surplus at t, b_{t-1} is the real value of government debt at $t-1$, Z_t is a vector of control variables, and ε_t^F is a possibly serially correlated disturbance. This line of work interprets estimates of (43) as descriptions of fiscal policy behavior.[at] When $\hat{\gamma} > 0$, researchers infer that fiscal behavior is passive, while if $\hat{\gamma} >$ net real interest rate, fiscal policy reacts sufficiently to stabilize debt. Based on such estimates, researchers conclude the economy resides in regime M, so the fiscal theory does not apply.[au]

Missing from this analysis is the bond valuation equation, which is an equilibrium condition that holds regardless of the prevailing policy regime. As condition (42) makes clear, b_{t-1} must be positively correlated with future surpluses *in any equilibrium*. When (43) is estimated without imposing this equilibrium condition, estimates of γ are subject to simultaneous equations bias.

Leeper and Li (2015) use a linearized variant on the endowment economy in Section 2 to study the nature of the simultaneity bias. If the policy disturbance is serially uncorrelated or a lagged dependent variable is added to the regression in (43), then the limited information procedure is valid only if the underlying monetary and fiscal policies are in regime M. Serious biases can arise when data are equilibria in regime F. The sign and severity of bias in $\hat{\gamma}$ depend on monetary policy behavior: the weaker is the reaction of monetary policy to inflation, the stronger is the positive bias. In periods like the aftermath of the 2008 financial crisis, when central banks pegged the nominal interest rate, estimates of γ are more likely to imply a strong response of surpluses to debt. This finding is consistent with Bohn's (1998) estimates, which rarely find evidence that the surplus response is weak.

There are two natural solutions to the simultaneous equations bias. The first is to impose the bond valuation equation on estimates of the fiscal rule, as Chung and

[as] For convenience, (42) assumes a constant real return.

[at] See, for example, Mendoza and Ostry (2008). Ghosh et al. (2012) employ such estimates to compute a country's "fiscal space." Woodford (1999) raises issues with this interpretation.

[au] Canzoneri et al. (2001b) estimate an unrestricted bivariate VAR for the primary surplus and the real value of debt, a technique that is equivalent to estimating a version of (43).

Leeper (2007) and Hur (2013) do in a structural VAR, and estimate monetary and fiscal rules jointly. The second solution is to estimate a fully specified DSGE model.

6.2 Some Suggestive Empirical Evidence

A complete account of empirical evidence about policy regime is beyond the scope of this chapter, so we will briefly recount two kinds of evidence that regime F has prevailed in some historic periods. The first is suggestive evidence that points to empirical facts that are consistent with regime F; then we turn to more formal econometric analysis.

Cochrane (1999) was the first to suggest that U.S. post-World War II inflation could be interpreted through the lens of the fiscal theory. He stresses that readily available fiscal data do not line up well with the theoretical concepts and constructs a data series for the real market value of government debt, from which he infers two different real primary surplus series. Not surprisingly, substantial differences emerge between the primary surplus and conventionally measured surplus (inclusive of debt service), particularly in periods of high debt or high interest rates. He further contrasts his computed surplus series with the Treasury's reported net-of-interest surplus, which does not account for capital gains and losses incurred from bond transactions. Cochrane's calculations make the broad methodological point that scrutiny of regime F equilibria requires careful data construction.

But Cochrane's substantive contribution lies in interpreting the data correlations. He specifies an exogenous—regime F—process for primary surpluses from which he computes the real value of debt as the present value of those artificial surpluses. Processes are chosen to match correlations in the data. Simulations produce observed gross movements in post-war U.S. inflation when the equilibrium price-level sequence emerges from the debt valuation equation.[av] As it happens, the chosen processes would pass either the Bohn (1998) or the Canzoneri et al. (2001b) "test" that those authors claim refutes the fiscal theory. Cochrane's analysis illustrates the difficulties in distinguishing between regimes M and F.[aw]

Woodford (2001) argues that Federal Reserve policy from before World War II until the Treasury-Fed Accord in March 1951 is a clear example in which monetary policy was explicitly assigned the task of maintaining the value of government debt, as it is in regime F. Beginning in April 1942, as Woodford writes

[av] Shim (1984) is an early effort to use VAR analysis to find cross-country evidence of a link between fiscal deficit innovations and inflation.

[aw] Cochrane (2011b) uses the government debt valuation condition to interpret monetary and fiscal policy actions in the wake of the 2008 global recession. He argues that recent policy developments suggest that in coming years the equilibrium condition is likely to have a stronger influence on economies than it has in the past.

> *The yield on ninety-day Treasury bills was pegged at 3/8 of a percent; this peg was maintained through June 1947, and … until that point the price of bills was completely fixed, as the Treasury offered both to buy and sell bills at that price. An intention was also announced of supporting one-year Treasury certificates at a price corresponding to a 7/8 percent annual yield; this policy continued after 1947, though at a slightly higher yield. Finally, the prices of twenty-five-year Treasury bonds were supported at a price corresponding to a 2 and 1/2 percent annual yield; this price floor was maintained up until the time of the "Accord."*
>
> **(Woodford, 2001, pp. 672–673)**

Woodford, however, seems to regard regime F as the exception, arising during wartime and in special circumstances when monetary policy is subordinated to fiscal needs.

Loyo (1999) uses Brazil in the late 1970s and the early 1980s as an example where the fiscal consequences of monetary policy led to explosive inflation. His case does not fall into either of the two regimes in which a determinate bounded equilibrium exists. Instead, Loyo argues that a combination of active fiscal policy and active monetary policy that aggressively sought to combat inflation by raising interest rates strongly in response to inflation produced exactly the phenomenon that Section 3.2.2 describes. Higher interest rates raised bondholders' interest receipts which, in the absence of commensurately higher taxes, raised wealth and aggregate demand. Higher demand increased inflation still further, to which monetary policy responded by raising interest rates, setting off an explosive cycle that produced double-digit inflation rates *per month*. Importantly, this hyperinflation arose with no appreciable change in real seigniorage revenues, as Loyo documents. Loyo's work illustrates a theme that runs through the chapter. If fiscal behavior is active, refusing to raise surpluses to stabilize government debt, more aggressive inflation fighting by the central bank exacerbates the problem: when monetary policy is passive, it amplifies shocks more as it becomes more active; if it is active, those shocks lead to ever-increasing inflation. An alternative monetary policy rule—one that merely pegged the nominal interest rate, for example—would have prevented the explosive inflation.

As of 2015, Brazil may be poised to rerun the experience that Loyo describes. Brazil's 1988 Constitution mandates that government benefits are indexed to inflation, effectively putting 90% of expenditures out of the legislature's reach. With sizeable tax adjustments apparently politically unviable, the gross-of-interest budget deficit reached over 10% of GDP in 2015. Consumer price inflation rose steadily through the year to breach double digits by year-end, despite the Banco Central do Brasil's aggressive antiinflationary efforts that raised the policy interest rate to 14.25% in the second half of 2015 (Banco Central do Brasil, 2015). As *The Economist* (2016) put it: "Fiscal dominance has left arcane discussions among economic theorists and burst onto newspaper columns." As in the period that Loyo studies, rising inflation is driven by the combination of active fiscal behavior and single-minded inflation targeting by the central bank. Coupling that fiscal behavior with passive monetary policy, as in regime F, would not generate explosive inflation rates.

Another recurring theme of the chapter's theory is that debt revaluation effects are a ubiquitous feature of both ad hoc and optimal policy rules. Sims (2013) calculates that since 1960 the surprise gains and losses on U.S. government debt as a percentage of GDP are similar in magnitude to the fluctuations in the deficit relative to GDP: debt revaluations are an important aspect of monetary–fiscal dynamics.[ax] Similarly, Akitoby et al. (2014) calculate that there would be substantial reductions in debt-to-GDP ratios for several developed economies from raising inflation targets to 6%. But Hilscher et al. (2014) argue that it is important to account for the maturity structure of the debt which is actually held by the private sector when undertaking such calculations, concluding that for the United States this may be lower than the maturity of the overall debt stock. Sections 4.4 and 5.3 found that the efficacy of using revaluation effects as a tool of optimal policy increases with both the size and the maturity of the outstanding debt stock. This suggests that the recent increase in debt-to-GDP ratios in most advanced economies raises the likelihood that such revaluation effects may become an increasingly important feature of policy. This does not establish that revaluation effects of the magnitude that Sims reports can come only from regime F-style policies. Instead, it points toward an important source of fiscal financing that formal macro models must confront.

6.3 Some Formal Empirical Evidence

Sims (1998) argues that to assess which part of the policy space—regime M or F—is empirically relevant, it is essential to embed alternative descriptions of policy within a general equilibrium model before taking them to the data. This leads to a more direct attack on the empirical problem of discerning policy regime, as well as the possibility of "testing" which regime is most consistent with observed data.

Leeper and Sims (1994) is an early attempt to estimate a DSGE model with a complete specification of monetary and fiscal policy. Real and nominal rigidities made the analogs to regimes M and F lie in a complicated geometry and the numerical search algorithm had to traverse regions of the parameter space in which either no equilibrium exists or the equilibrium is indeterminate—both cases where the likelihood function is not defined. These difficulties prevented the paper from reaching a conclusion about which policy combination yielded the best fit.[ay]

Bayesian estimation methods have permitted researchers to overcome some of the limitations of earlier work to make progress on the question of the prevailing regime. Expanding on the money-only specification of Smets and Wouters (2007), the models

[ax] See also Taylor (1995), King (1995), and Hall and Sargent (2011) for discussions of and estimates of revaluation effects.

[ay] Leeper (1989) is an even earlier effort that uses a calibrated DSGE model to ask whether impulse response functions from regime M or regime F best match empirical responses. When agents are endowed with foresight about future fiscal actions, there is weak evidence in favor of regime F.

fill in fiscal details and impose the government's budget identity to estimate monetary and fiscal behavior jointly with private behavior. Traum and Yang (2011) impose priors that are centered on either regime M or regime F for various subperiods of U.S. data from 1955 to 2007 and find that the data least prefer the parameter space associated with regime F.

Using a simpler new Keynesian model, but with a maturity structure for government bonds, Tan (2014) argues that rejection of regime F stems from a test procedure that Geweke (2010) calls the "strong interpretation." The strong interpretation takes literally all the cross-equation restrictions of a fully specified dynamic general equilibrium model, which necessarily includes any and all possible sources of misspecification. When Tan employs the methods that DeJong et al. (1996) and Del Negro and Schorfheide (2004) developed, which take the DSGE model as a prior for a VAR, he finds that data no longer strongly prefer regime M. Tan argues that tests of model fit that are robust to misspecification no longer find compelling support for one regime over the other.

Leeper et al. (2015b) estimate medium-scale models that include additional fiscal details—government consumption that may complement or substitute for private consumption, a maturity structure for government debt, explicit rules for several fiscal instruments, and steady-state distorting taxes. For U.S. data covering 1955–2014, even under the strong interpretation, marginal data densities suggest nearly equivalent fits under the two regimes for the full sample and for pre- and post-Volcker subsamples. Details of model specification are as important as policy rules for determining the relative fit of the two regimes.

That paper also reports estimated revaluation effects that arise from government spending expansions that are initially financed by selling debt (partially reproduced in Table 4). These are analogous to the first two columns in Table 2, but the estimated model also includes many other sources of financing—capital, labor and consumption

Table 4 Reports 90% credible intervals around posterior modes

	% due to $\hat{\pi}_t$	% due to $\hat{P}_t^m$
1955q1–2014q2		
Regime M	[0.3,0.6]	[8.2,13.6]
Regime F	[0.5,0.8]	[11.8,17.0]
1955q1–1979q4		
Regime M	[−0.3,0.3]	[0.7,12.7]
Regime F	[0.6,1.2]	[18.4,29.9]
1982q1–2007q4		
Regime M	[0.1, 0.4]	[7.3,14.2]
Regime F	[0.1,0.9]	[13.2,22.9]

"% due to" are the ratios of the analogs to the right-hand components of (38) to ξ_t, which are computed from the impulse response to a shock to government spending.
Source: Leeper, E.M., Traum, N., Walker, T.B., 2015b. Clearing up the fiscal multiplier morass. NBER Working Paper No. 21433, July.

tax revenues, real interest rates, government transfers, and endogenous government spending. Over the full sample and the post-Volcker subsample, the 90% credible intervals display substantial overlap for both inflation and bond prices, suggesting no large differences in revaluation effects in the two regime. Intervals do not overlap in the pre-Volcker period, with larger revaluation effects in regime F for both components.

Both the theory in this chapter and the empirical evidence just cited make clear that revaluation effects that stabilize the value of government bonds are not solely the preserve of regime F. Even in the endowment economy with policy described by simple rules in Section 2, monetary policy and government spending shocks both induce revaluation effects in the two policy regimes. Optimal policy exercises show that it is desirable to use a combination of surprise inflation and tax smoothing to stabilize the economy in the face of fiscal shocks, blurring the lines between the M and F regimes. Such exercises also suggest that the balance between inflationary and fiscal financing is also highly state dependent. In richer production economies subject to nominal inertia, the range of monetary and fiscal policy interactions is far wider: monetary and fiscal policy jointly determine the extent to which there are inflation surprises, movements in real interest rates and bond prices and changes in the tax base. The relative magnitudes of these effects, though, depend on the nature of the policy regime and on the level and maturity of the debt stock.

6.4 Regime-Switching Policies

A growing body of work estimates Markov-switching policy rules and embeds them in otherwise conventional DSGE models. Davig and Leeper (2006) find recurring switches between active and passive monetary and fiscal rules, with some periods in which both policies are active or passive. In a rational expectations model in which agents are endowed with knowledge of the policy process, no single monetary–fiscal mix determines the nature of the equilibrium. Instead, expectations of future policy regimes spill-over to affect the current equilibrium. In a new Keynesian model with lump-sum taxes, Davig and Leeper show that even if regime M currently prevails, a tax cut can produce quantitatively important increases in output and the price level. The effects are still larger conditional on being in regime F.

Gonzalez-Astudillo (2013) uses limited information Bayesian methods to estimate a new Keynesian model with monetary and fiscal policy rules whose coefficients are time varying and interdependent. He finds that monetary policy switches more frequently than fiscal policy—a result that contrasts with findings from Markov-switching models—and that the policies are interdependent. But other findings align closely to models with recurring Markov switching: a monetary contraction reduces inflation in the short run, but raises it over longer horizons; lump-sum tax changes always affect output and inflation.

Kliem et al. (2016) find some provocative reduced-form support for time-varying fiscal effects. Using U.S. data from 1900 to 2011, they discovered that the low-frequency correlation between inflation and the fiscal stance—defined as the ratio of primary deficits to government debt—is significantly positive most of the time until 1980 when it becomes zero. They attribute the shift in correlation to a change in monetary policy behavior.

Those authors extend their analysis in Kliem et al. (2015) to include Germany and Italy and to interpret their findings with an estimated DSGE model. Germany never exhibits a significant low-frequency correlation between fiscal stance and inflation, while in Italy the correlation is positive until the Banca d' Italia gained its independence in the 1990s.

Bianchi (2012) and Bianchi and Ilut (2014) estimate a simple new Keynesian model with fiscal policy, habits, and inflation inertia and that also allows for switches in monetary and fiscal policy rules. Bianchi permits a circular movement across three regimes where policy can transition from the conventional assignment (active monetary policy/passive fiscal policy) through the fiscal theory assignment of passive monetary/active fiscal policy, to an unstable regime where both monetary and fiscal policy are active. He finds that the 1960s and 1970s featured a combination of passive monetary and active fiscal policy, before the Volcker disinflation resulted in a combination of active monetary and fiscal policies. Only around 1990 did fiscal policy turn passive. Bianchi and Ilut model a slightly different set of policy transitions that allows the two stable regimes (active monetary/passive fiscal and passive monetary/active fiscal) to briefly transition through the unstable, doubly active, regime. In their estimates, regime F prevails until before monetary policy turns active in 1979 and fiscal policy turns passive shortly afterward (by 1982). These papers suggest that regime M, though not always in place historically, has been the predominant regime in the United States from at least the early 1990s until the financial crisis.

Chen et al. (2015) build on this work in two ways. First, they allow additional permutations of policy in which monetary and fiscal policy may be simultaneously passive and they make the nature of transitions across regimes less restrictive. Their estimates find that the switch to regime M after the Volcker disinflation is far less certain, with both monetary and fiscal policy repeatedly falling outside regime M, even in the recent data.

Second, Chen et al. move away from ad hoc rules for policy to permit monetary and, in some exercises, fiscal policy to be chosen optimally. Monetary policy turns out to be both optimal and time-consistent, but with switches in the degree of anti-inflation conservatism. Those switches imply that monetary policy was not only less conservative in the 1970s, but also intermittently during the 1960s and briefly after the financial market turmoil from the stock market crash of 1987, the Russian default in 1998, and the dot-com crash. At the same time, fiscal policy can rarely be described as optimal (except in the early 1990s), and instead tends to move between an active and passive rule. For the bulk of the period between 1954 and the 2008 financial crisis, fiscal policy was primarily active with the only sustained periods of passive fiscal policy from the late 1950s until the late 1960s, between 1995 and 2000, and briefly between 2005 and the financial crisis. These estimates imply that regime M is the exception rather than the norm.

More subtle findings in Chen et al. emerge from examining the roles of the maturity structure and the level of debt in determining optimal policy. Sections 4.4 and 5.3 found that the Ramsey plan does resemble regime M in periods when debt levels are low and maturity is long: monetary policy was tightened to stabilize inflation in the face of a government spending shock, while tax rates were raised to stabilize debt. But as debt levels rise, especially when maturity is short, policy assignments get reversed: monetary policy responds weakly to higher inflation from increased government spending to reduce debt service costs and stabilize debt, while tax rates are cut to stabilize inflation. In contrast, under the institutional design of policy with an independent central bank that follows an active Taylor rule, the Ramsey policy actually cuts taxes in the face of the same government spending shock, reducing inflation and offsetting the increase in debt service costs that active monetary policy induces. Despite this anti-inflationary policy on the part of the fiscal policymaker, the equilibrium rate of inflation when the central bank was independent is an order of magnitude higher than when monetary and fiscal policy were jointly optimal. Evidently, the nature of the policy interactions in theory is complex and state contingent, as it appears to be in the empirical regime-switching literature.

Empirical evidence and optimal policy argue that regime M is not the only relevant monetary–fiscal policy mix. Interactions between monetary and fiscal policy are both pervasive and changeable. Understanding the nature of the policy dynamics—both the interactions between monetary and fiscal authorities and the political conflict that drives fiscal policy choices—is likely to be critical to identifying and understanding the evolution of observed policy regimes.

6.5 Common Misperceptions

Economists generally agree that historical episodes of high and volatile inflation rates inevitably have fiscal roots. Building on Sargent and Wallace's (1981) unpleasant monetarist arithmetic logic, Sargent (1986) makes a forceful historical case for hyperinflation's fiscal roots. The association between fiscal dominance—exogenous primary surpluses in Sargent and Wallace—and rampant inflation outcomes is so ingrained that many macroeconomists also believe that regime F fiscal behavior—a weak response of surpluses to debt—necessarily produces bad economic performance.[az]

That belief is unfounded. Bad economic policies can produce bad economic outcomes in any policy regime. And regime F is no more susceptible to undesirable

[az] Cochrane (2005) and Leeper and Walker (2013) give detailed descriptions of how the fiscal theory differs from unpleasant monetarist arithmetic.

equilibria than any other monetary–fiscal mix. Both the theoretical and the empirical results we have reviewed underscore this point.

Fiscal dominance can produce explosive inflation, as Loyo (1999) argues happened in Brazil. But explosiveness is the outgrowth of monetary behavior that is incompatible with fiscal dominance. When fiscal policy is active, ever-increasing inflation arises when the central bank aggressively raises the policy interest rate in a misguided effort to combat inflation. The active fiscal behavior transforms higher interest rates into more rapid growth in nominal government debt, higher aggregate demand, and higher inflation.

Perhaps ironically, Cochrane (2011a), Sims (2013), and Del Negro and Sims (2015) argue that many of the monetary anomalies in the theoretical literature arise primarily because money-only analyses trivialize the role that fiscal policy can play in delivering stable price-level behavior. Those anomalies include Obstfeld and Rogoff's (1983) speculative hyperinflations and Benhabib et al.'s (2002) deflationary traps. Fiscal policy can rule out both cases by adopting behavior that deviates in some fashion from typical regime M fiscal behavior. To eliminate hyperinflations, surpluses need to rise proportionately to excess inflation outside inflation's target range.[ba] To ensure that the economy will not get mired in a deflationary trap, fiscal policy must commit to running deficits or shrinking primary surpluses until inflation reaches its target. Both of these policy functions make fiscal choices explicitly contingent on inflation outcomes.

Monetary policy alone is powerless to eliminate these undesirable equilibria. Ruling out those equilibria requires fiscal policy to deviate from purely passive behavior that centers entirely on debt stabilization.

Skeptics who question whether the economic mechanisms in regime F have ever been observed point to instances in which government debt has grown rapidly, while inflation has been low and steady as *prima facie* evidence that inflation is solely a monetary phenomenon. But this criticism is akin to treating the income velocity of money as constant and finding cases where monetary expansions were not followed by higher nominal spending.

Consider the U.S. experience in the aftermath of the financial crisis. Nominal government debt grew from $4.4 trillion to $10.6 trillion from December 2007 and December 2014, a growth rate of 240% that raised the debt–GDP ratio from 30.5% to 61.0%.[bb] Despite this massive growth in debt, U.S. consumer price inflation averaged 1.9% between 2008 and 2014. With the Federal Reserve pegging the federal funds rate near zero from December 2008 onward, monetary policy behavior appears to have been

[ba] Cochrane (2011a) points out that hyperinflations do not violate any equilibrium conditions, so they are perfectly reasonable equilibria. They are also likely to be welfare reducing and undesirable.

[bb] These numbers come from the Federal Reserve Bank of Dallas's privately held gross federal debt and the U.S. Department of Commerce's annual nominal GDP data. Congressional Budget Office (2015) reports that federal debt held by the public rose from 35% to 74% over the same period.

passive, as in regime F. But the theory in this chapter predicts that if the debt expansion is not associated with higher taxes, private-sector wealth increases, raising aggregate demand and inflation. Where is the inflation that the fiscal theory predicts?

Like constant velocity, simple expositions of the fiscal theory serve pedagogical purposes, but severely constrain the theory's empirical predictions. Missing from the simple theory is that debt's value derives from the *present value* of expected surpluses and that the present value also depends on the expected path of real discount rates. Real interest rates have been decidedly negative in the United States. Kiley (2015) estimates that the real federal funds rate was negative from the onset of the recession through the middle of 2015. Even yields on 5-year Treasury inflation-indexed securities were negative or hovering around zero from September 2010 through 2015, reaching a nadir of −1.47% in October 2012. To the extent that these low rates flowed into real discount rates applied to government debt, the expected present value of surpluses was very high indeed over this period, even in the absence of any anticipated increases in primary surpluses. And along with the low real interest rates that the Federal Reserve sought to achieve, the crisis brought a flight to quality in which investors fled from nongovernment-insured asset classes to government securities, which drove down real treasury bond yields.

Any demand stimulus created by the nominal debt expansion would be offset, at least in part, by the increase in the value of debt that low real discount rates induce. It would take a careful quantitative analysis to make this case convincingly, but we see no a priori refutation of regime F from these observations.

If anything, the logic of the fiscal theory may help to explain the anomaly of why inflation did not fall *as much* as conventional money-only models predicted. The lack of persistent deflation during the recent recession caused some prominent economists to question the validity of conventional Phillips curve models where inflation is driven by measures of economic slack.[bc] Del Negro et al. (2015) argue that conventional models with a new Keynesian Phillips curve can account for the lack of deflation despite a large negative output gap provided prices are sufficiently sticky and inflation expectations remain anchored at positive levels. In their model, the anchoring comes from the anticipation that monetary policy will achieve future rates of inflation that are close to target. An alternative hypothesis is that expectations of future inflationary financing of the large increases in government debt are providing the necessary anchor.

A second canonical example thrown up by skeptics is Japan. Since 1993, Japanese government debt has risen from 75% to 230% of GDP, while inflation has averaged a mere 0.21%. For 20 years beginning in 1995, the Bank of Japan's overnight call rate has been below 0.5% and at 0.1% or lower for more than 12 of those years. Evidently, Japanese monetary policy has been passive. Once again, where is the inflation that the fiscal theory predicts?

[bc] For example, Hall (2011) and Ball and Mazumder (2011).

Japan is a complicated case. Real interest rates have been low, just as in the United States recently, but there is more to the story.[bd] Japan is the poster child for inconsistency in macroeconomic policies, as Krugman (1998), Ito (2006), Ito and Mishkin (2006), and Hausman and Wieland (2014) document. Fiscal policies have see-sawed between stimulus and austerity. Even as Prime Minister Abe appeared to announce an end to the inconsistency and Japanese economic activity and inflation were showing signs of life, Japan raised the consumption tax rate from 5% to 8% in April 2014. Consumer price inflation fell from 2.7% in 2014 to below 1% in 2015 (Leeper, 2016).

Japan has been mired in the trade-off between fiscal sustainability and economic reflation. To a fiscal theorist, Japan's obsession with government debt reduction is puzzling. Central to a regime F equilibrium is that agents' expectations are anchored on fiscal policies that do not raise surpluses when debt expands. Unsettled fiscal policies like those in Japan are unlikely to have so anchored expectations, so it is not clear that Japan resides in regime F; there may be no contradiction of the fiscal theory to explain.

7. PRACTICAL IMPLICATIONS

Viewing practical issues through the joint lenses of monetary and fiscal policies sheds fresh light on policy problems. That new light can also lead to sharply different perspectives on these problems.

7.1 Inflation Targeting

Nearly 30 countries with independent central banks have embraced numerical inflation targeting as the operating principle for monetary policy. Very few of these countries sought simultaneously to adopt fiscal policies that are compatible with the chosen inflation targets. This discussion of the policy interactions that are prerequisites for successful inflation targeting does not depend on the prevailing monetary–fiscal regime, so it applies whether policies reside in regime M or regime F.

The derivations rely on a few generic first-order conditions, a government budget identity, and the condition that optimizing households will not want to over- or under-accumulate assets. For this reason, the results have broad implications that extend well beyond the details of particular models. Consider an economy with a geometrically decaying maturity structure of zero-coupon nominal government bonds. The government's budget identity is

[bd] Imakubo et al. (2015) calculate that real yields on zero-coupon bonds at 1-, 2-, and 3-year maturities fluctuated between 0.5% and −0.5% from the middle of 1995 until 2012, when they fell to almost −2.0% in 2014.

$$\frac{P_t^m B_t^m}{P_t} = \frac{(1+\rho P_t^m) B_{t-1}^m}{P_t} - s_t$$

Letting $Q_{t,t+k} \equiv \beta^k \frac{u'(c_{t+k})}{u'(c_t)} \frac{P_t}{P_{t+k}}$, asset-pricing conditions yield

$$\frac{1}{R_t} = E_t Q_{t,t+1}$$
$$P_t^m = E_t Q_{t,t+1}(1+\rho P_{t+1}^m)$$

and the term structure relationship is

$$P_t^m = E_t \sum_{k=0}^{\infty} \rho^k \left(\prod_{j=0}^{k} \frac{1}{R_{t+j}} \right)$$

These conditions deliver the usual bond valuation equation

$$\frac{(1+\rho P_t^m) B_{t-1}^m}{P_t} = E_t \sum_{i=0}^{\infty} \beta^i \frac{u'(c_{t+i})}{u'(c_t)} s_{t+i}$$

Rewrite the valuation equation by replacing $(1+\rho P_t^m)$ using

$$1+\rho P_t^m = 1 + E_t \sum_{k=1}^{\infty} (\beta\rho)^k \frac{u'(c_{t+k})}{u'(c_t)} \frac{P_t}{P_{t+k}}$$

and, for simplicity, assume a constant-endowment economy, so $\frac{u'(c_{t+i})}{u'(c_t)} = 1$, to generate

$$\left[\sum_{k=0}^{\infty} (\beta\rho)^k \left(\prod_{j=1}^{k} \frac{1}{\pi_{t+j}} \right) \right] \frac{B_{t-1}^m}{P_t} = E_t \sum_{k=0}^{\infty} \beta^k s_{t+k} \tag{44}$$

Imagine an economy that takes as given variables dated $t-1$ and earlier, but commits to hitting an inflation target in all subsequent dates, so $\pi_{t+k} \equiv \pi^*$ for $k \geq 0$. Valuation equation (44) becomes

$$\frac{B_{t-1}^m / P_{t-1}}{EPV_t(s)} = \pi^* - \beta\rho \tag{45}$$

where $EPV_t(s) \equiv E_t \sum_{k=0}^{\infty} \beta^k s_{t+k}$.

This expression imposes stringent conditions on the expected present value of primary surpluses, though not on the surplus path, if the inflation target is to be achieved. For given initial real debt, if the economy adopts a policy of "too high" surpluses, then the inflation target that is achievable is lower than the desired target, π^*. Another way of seeing the tension between monetary and fiscal policy in this equation is to note that the condition requires the fiscal policymaker to adopt a debt target, which it passively adjusts surpluses to achieve. This means that any period of austerity that raises surpluses must induce a subsequent relaxation of policy to bring $EPV_t(s)$ in line with the outstanding debt stock and the inflation target. An austerity program that never took its foot off the gas would undermine the inflation target just as surely as would a myopic fiscal policymaker prone to runaway deficits. Are current fiscal frameworks consistent with such targets?

Both before and since the recent crisis, policymakers have been adopting fiscal rules designed to reverse increases in government debt. For example, following its banking crisis of 1992 Sweden adopted two fiscal rules: a net lending target of 1% of GDP over the economic cycle and a nominal expenditure ceiling 3 years ahead. This ceiling is consistent with ensuring that government expenditure falls as a share of GDP. Similarly, the "debt brake" in Switzerland requires that central government expenditure cannot grow faster than average revenue growth, while the German debt brake introduced in 2011 imposes a limit on federal net lending of 0.35% of GDP. In the United Kingdom, the 2015 Charter for Budget Responsibility requires the government to run a primary surplus in "normal" times. All these measures aim not only to stabilize the debt-to-GDP ratio but to ensure that it is falling over time. And to the extent that the rules are maintained, the pace of debt reduction should increase over time as less of any surplus is devoted to servicing the existing stock of debt. Because these rules fail to include provisions to target a long-run debt-to-GDP ratio, which would relax austerity measures as that target was approached, the rules run the risk of chronically undershooting the inflation target.

From a theoretical perspective, the rules do not make surpluses contingent on debt or the price level. This makes fiscal behavior active, placing it in regime F. When the fiscal policymaker adopts an active rule, as Section 2.3 shows, the monetary authority's ability to control inflation depends crucially on the maturity structure of the outstanding debt and on the nature of its policy response. With a pegged nominal interest rate, inflationary expectations remain consistent with the inflation target and surprise deviations from that target provide the revaluation effects needed to stabilize debt. But if the central bank attempts to come as close to active as possible by setting $\alpha_\pi = \beta$, the rate of inflation follows a random walk, permanently deviating from the inflation target in the face of fiscal shocks. If the policy objective is to smooth the inflationary costs of revaluation effects, then the optimal policy exercises suggest that a persistent deviation from the inflation target is desirable, so long as the persistence matches the maturity structure of the government's debt portfolio. With only single-period debt, there is no advantage in having a prolonged increase or decrease in inflation following a fiscal shock because only the initial period's inflation helps to reduce the real value of government liabilities. But when debt is of longer maturity, allowing inflation to rise and then gradually decline as the predetermined debt stock matures reduces the discounted value of inflationary costs associated with the required revaluation effects.

Successful inflation targeting requires more than a resolute central bank that follows "best practice" monetary policy behavior that includes clear objectives, transparency that leads to effective communications, and accountability. Even with all these elements in place, expression (45) implies that the central bank can achieve π^* only if fiscal policy is compatible with that target. If fiscal behavior requires a long-run inflation rate that differs from π^*, even best practice monetary policy cannot succeed in anchoring long-run inflation expectations or inflation outturns on target.

7.2 Returning to "Normal" Monetary Policy

The financial crisis has seen a substantial increase in debt-to-GDP ratios in many advanced economies, although the immediate need for fiscal adjustment may have been muted due to the reduced debt service costs as real interest rates have fallen since the financial crisis. To see this consider a small change to our policy problem in the endowment economy, in Section 4.2, where we allow the households' discount factor, $\tilde{\beta}_t$, to rise temporarily to $\tilde{\beta} > \beta$, capturing the flight to quality observed in the financial crisis. If we assume government spending is held constant, the policy problem becomes

$$
\begin{aligned}
L_t = E_0 \frac{1}{2} \sum_{t=0}^{\infty} \beta^t [& -\frac{1}{2}(\tau_t^2 + \theta(\nu_t - 1)^2) \\
& + \mu_t(\tilde{\beta}_t E_t \frac{(1+\rho P_{t+1}^m)}{P_t^m} \nu_{t+1} - 1) \\
& + \lambda_t(b_t P_t^m - (1+\rho P_t^m) b_{t-1} \nu_t - g_t - z_t + \tau_t)
\end{aligned}
$$

which yields the first-order conditions

$$
\begin{aligned}
&\tau_t : -\tau_t + \lambda_t = 0 \\
&\nu_t : -\theta(\nu_t - 1) + \mu_{t-1} \frac{(1+\rho P_t^m)}{P_{t-1}^m} \beta^{-1} \tilde{\beta}_{t-1} - (1+\rho P_t^m) \lambda_t b_{t-1} = 0 \\
&P_t^m : -\frac{\mu_t}{P_t^m} + \mu_{t-1} \rho \frac{\nu_t}{P_{t-1}^m} \beta^{-1} \tilde{\beta}_{t-1} + \lambda_t (b_t - \rho \nu_t b_{t-1}) = 0 \\
&b_t : \lambda_t P_t^m - \beta E_t (1+\rho P_{t+1}^m) \nu_{t+1} \lambda_{t+1} = 0
\end{aligned}
$$

Under a perfect foresight equilibrium this implies the tax-smoothing result is recast as

$$
\tau_t = \beta \tilde{\beta}_t^{-1} \tau_{t+1}
$$

which means that the tax rate will be rising during the period in which households have an increased preference for holding government bonds over consumption. Intuitively, the original tax-smoothing result balances the short-run costs of raising taxes to reduce debt against the long-run benefit of lower debt. These costs and benefits are finely balanced with the interest rate on the debt being exactly offset by the policymaker's rate of time preference so that steady-state debt follows a random walk in the face of shocks. When the interest on debt is less than the policymaker's rate of time preference, the policymaker prefers to delay the fiscal adjustment and will allow debt to accumulate, stabilizing debt only after the period of increased household preference for debt holdings has passed.

To the extent that a return to "normal" monetary policy is associated with a rise in debt service costs, optimal policy suggests that efforts to stabilize debt are enhanced at this point. But under the Ramsey policy, inflation surprises to revalue debt are effective only if carried out before the predetermined debt stock matures. Therefore the delay in debt stabilization also reduces the efficacy of promising to raise prices in the future placing more of the burden of adjustment on taxation. At the same time, the higher debt stock that emerges at the point of normalization raises the potential time-inconsistency problems inherent in the Ramsey policy; at this point we may start to see increased pressure to inflate away the debt.

More generally, higher central bank interest rates have powerful fiscal consequences when government debt levels are elevated. In the United States, the Congressional Budget Office (2014) estimates that net interest costs will quadruple between 2014 and 2024 to reach 3.3% of GDP.[be] Those interest costs must be financed somehow—by higher taxes and lower spending now or by faster growth in debt and other adjustments in the future. In light of the political dynamics today in the United States, it is not obvious how those costs will be financed.

Central bankers are well aware of the fiscal consequences of their actions. King (1995) refers to "unpleasant fiscal arithmetic"—a process of monetary disinflation raises real interest rates and destabilizes government debt until the credibility of the disinflation is established. But, he argues, the higher debt may actually undermine that credibility and unpleasant monetarist arithmetic may re-emerge. One interpretation is that King worries about the danger that the fiscal consequences of disinflation may force the central bank to reverse a return to "normal" interest rates.

7.3 Why Central Banks Need to Know the Prevailing Regime

Davig and Leeper (2006), Bianchi (2012), Bianchi and Ilut (2014), and Chen et al. (2015) suggest that there have been switches in the conduct of fiscal policy between passive and active rules. And fiscal switches are not always associated with compensating switches in monetary policy that place the economy in either regime M or regime F. If these policy permutations were permanent, they would either result in indeterminacy (passive

[be] The CBO expects a relatively modest interest in treasury interest rates over that period, with the 10-year rate rising from 2.8 to 4.7 percentage points and the average rate on debt held by the public rising from 1.8 to 3.9 percentage points. Cochrane (2014) considers a scenario in which the Fed raises interest rates to 5% and with them, real interest rates. At a 100% debt–GDP ratio, the increased interest costs amount to $900 billion.

monetary and fiscal policy) or nonexistence of equilibrium (active monetary and fiscal policy). But if policy is expected to return to either the M or F regime sufficiently often, then these policy combinations can still deliver determinate equilibria. So there are four possible permutations of monetary and fiscal policy that may coexist, but only two, if permanent, deliver unique bounded equilibria. The prevailing policy configuration can have profound implications for the conduct of monetary policy, as we illustrate in the endowment economy with Section 2's policy rules.[bf]

Regardless of regime, inflationary dynamics are

$$E_t(\nu_{t+1}-\nu^*)=\frac{\alpha_\pi}{\beta}(\nu_t-\nu^*) \tag{46}$$

Under regime M with an active monetary policy ($\alpha_\pi > \beta$), monetary policy can target inflation in each period, $\nu_t = \nu^*$, while the passive fiscal policy stabilizes debt

$$E_t\left(\frac{b_{t+1}}{R_{t+1}}-\frac{b^*}{R^*}\right)=(\beta^{-1}-\gamma)\left(\frac{b_t}{R_t}-\frac{b^*}{R^*}\right)-E_t\varepsilon^F_{t+1}$$

provided $\gamma > \beta^{-1} - 1$.

Suppose we know the economy will enter this regime in period T, at which point inflation will be at its target $\nu_T = \nu^*$ and the fiscal rule will stabilize whatever debt is inherited at time T. In this case, it does not matter whether or not the monetary policy rule is active or passive prior to period T, since T-step-ahead expected inflation is

$$E_t\nu_{t+T}-\nu^*=\left(\frac{\alpha_\pi}{\beta}\right)^{T-t}(\nu_t-\nu^*)$$

which implies that inflation will be on target between today and period T. If fiscal policy is active, debt will be moving off target between today and period T, but the passive fiscal rule will, from that point on, stabilize debt. If fiscal policy is passive before period T, this would facilitate the debt stabilization prior to T and the targeting of inflation would be uninterrupted by any change of regime at time T.

We now assume that at time T agents anticipate the economy will enter regime F where monetary policy is passive ($\alpha_\pi < \beta$), and fiscal policy does not respond to debt ($\gamma = 0$). Now the period T price level needs to adjust to satisfy the bond valuation equation at time T given the level of inherited nominal debt B_{T-1}. When $\gamma = 0$, the fiscal rule is $s_t = s^* + \varepsilon^F_t$ and the solution for real debt is

$$E_t\frac{B_{T-1}}{R_{T-1}P_{T-1}}=\frac{b^*}{R^*}+\sum_{j=1}^{\infty}\beta^j E_t\varepsilon^F_{T-1+j}$$

[bf] See Davig et al. (2010) and Leeper (2011) for related analyses.

The price level does not jump in period T, but it does adjust in period t when the switch to regime F in period T is first anticipated. The implications for inflation beyond period T depend on how passive the monetary policy rule is. With an interest rate peg, $\alpha = 0$, inflationary expectations remain on target, $E_t\nu_{t+1} = \nu^*$, but there will be innovations to inflation to ensure the bond valuation equation holds in the face of additional fiscal shocks occurring from period T onwards. With some monetary policy response to inflation, $0 < \alpha_\pi < \beta$, the initial jump in the price level will result in a temporary, but sustained rise in inflation whose evolution obeys equation (46). As Section 4 shows, sustaining the rise in inflation enhances the revaluation effect, but the longer is debt maturity, the greater is the reduction in distortions caused by higher inflation.

How does anticipating the F regime in period T affect the conduct of policy prior to period T? With fiscal policy following a rule that may or may not be passive, the expected evolution of government debt follows

$$E_t\left(\frac{B_{t+1}}{R_{t+1}P_{t+1}} - \frac{b^*}{R^*}\right) = (\beta^{-1} - \gamma)\left(\frac{B_t}{R_tP_t} - \frac{b^*}{R^*}\right) - E_t\varepsilon^F_{t+1}$$

We can iterate this forward until period T as

$$E_t\left(\frac{B_{T-1}}{R_{T-1}P_{T-1}} - \frac{b^*}{R^*}\right) = (\beta^{-1} - \gamma)^{T-1-t}\left(\frac{B_t}{R_tP_t} - \frac{b^*}{R^*}\right) + \sum_{j=0}^{T-1-t}(\beta^{-1} - \gamma)^j E_t\varepsilon^F_{t+1+j}$$

which defines the initial debt level $\dfrac{B_t}{R_tP_t}$ required to ensure the economy enters regime F in period T with the appropriate level of debt $\dfrac{B_{T-1}}{R_{T-1}P_{T-1}}$ without any discrete jumps in the price level at that time. This depends upon the extent to which fiscal policy prior to period T acts to stabilize debt as determined by the fiscal feedback parameter, γ, and the expected value of fiscal shocks over that period. If the move to the F regime is sufficiently long in the future and fiscal policy is sufficiently aggressive in stabilizing debt, then there will be little need for surprise inflation in the initial period to ensure the appropriate debt level is bequeathed to the future. But if the switch is more imminent or the fiscal stabilization prior to period T is muted, then an initial jump in prices will be required to ensure the bond valuation equation holds. The inflationary implications of this prior to period T depend on the conduct of monetary policy. If monetary policy is active prior to period T, any initial jump in prices will be explosive until the F regime is established in period T. This happens because the period t price-level jump ensures the bond valuation equation holds, while inflation dynamics are determined by equation (46), which is explosive under an active monetary policy. This is a bounded equilibrium because the process for inflation stabilizes when the policy regime changes in period T. But before period T, the active monetary policy actually destabilizes prices. Postponing the switch to the F regime means that the period of explosive inflation dynamics remains in place for longer.

This analysis has the flavor of a game of chicken between the monetary and fiscal policymakers. The monetary authority can stick to an active monetary policy rule and achieve its inflation target, provided everyone is sure that policy will eventually be supported by a passive fiscal policy which stabilizes debt. Debt dynamics will be unstable in such a scenario until the fiscal authorities relent and adopt a passive fiscal policy. But when there is the suspicion that monetary policy will eventually turn passive to support a fiscal policy that does not stabilize debt, then conventional anti-inflation policies today may actually worsen inflation outcomes.

8. CRITICAL ASSESSMENT AND OUTLOOK

We conclude by examining the areas where further theoretical and empirical work is needed.

8.1 Further Theoretical Developments

This section highlights areas in which additional theoretical work on monetary–fiscal interactions would be fruitful.

8.1.1 Default and the Open Economy

This chapter has focused on closed-economy models, abstracting from issues of sovereign default and open-economy dimensions that have come together in the recent sovereign debt crisis in the Euro Area. In the early applications of the fiscal theory to the open economy, a key issue was whether or not individual country government budget identities were consolidated into a single global bond valuation equation.[bg] If so, with multiple passive monetary policies, each country's price level and exchange rate are indeterminate. In this equilibrium, one country accumulates the debt of another, an outcome whose political equilibrium Sims (1997) argues is unstable. If such equilibria are ruled out, then we return to having a bond valuation equation for each country and fiscal policies in one economy carry implications for outcomes in the second economy. For example, a determinate active/passive policy pair can be achieved across countries rather than within countries (Leith and Wren-Lewis, 2008).

Similar issues arise in a monetary union. With a single passive monetary policy, it is possible to ensure determinacy with only one active fiscal policy (Leith and Wren-Lewis, 2006). These analyses have the troubling feature that the tail seems to wag the dog—a small monetary union member that fails to pursue passive fiscal policy can determine the price level for the entire union. This raises questions about whether these early applications of the fiscal theory to the open economy have appropriately captured

[bg] See Sims (1997), Loyo (1997), Woodford (1998b), Dupor (2000), Canzoneri et al. (2001a), and Daniel (2001).

cross-country heterogeneity—including different price-level processes across member states—and the cross-country implications of the interactions between monetary and fiscal policy. More recent work seeks to model the gross asset/liability positions of countries to capture the kinds of revaluation effects generated by price level and exchange rate movements.[bh] That work finds that the gross asset/liability positions can be several multiples of GDP even when net positions are not, implying that the revaluation effects stressed in this chapter are likely to be both quantitatively important and more complex in open-economy settings.

Recent events highlight the need to bring sovereign default into the analysis. In a model similar to our endowment economy, but augmented with an exogenous default risk, Uribe (2006) demonstrates that default can give rise to fiscal theory-type effects, with anticipated, but delayed defaults potentially destabilizing an active inflation targeting policy in much the same way that anticipating a move to regime F can do.

While many analyses of strategic default focus on real economies—for example, D'Erasmo et al. (2016)—when default through inflation is available as an alternative financing option, it is either assumed to be equivalent to outright default, or possibly less costly if it is less damaging to the balance sheets of a country's banking sector than an outright default (Gros, 2011). Given that inflation is costly, it is not obvious that this will always be the case. A useful line of work would consider the nature of the strategic default decision in environments in which debt revaluations through surprise current inflation and bond prices are possible. Kriwoluzky et al. (2014) is an interesting paper that contrasts outright default for a country engaged in a monetary union with the redenomination of debt following exit from the union. They find that the possibility of exit significantly worsens the preexit/default debt dynamics. Similarly, Burnside et al. (2001) argue that the speculative attacks on fixed-currency regimes in the Asian crisis of 1997 sprung from expectations that large revaluations of debt were required to finance the projected deficits that ongoing bank bailouts were expected to engender. In richer models where default is state dependent and the economic costs of default arise through the impact of default on domestic banks' balance sheets the set of monetary and fiscal interactions is widened further (Bi et al., 2015; Bocola, 2016). There is plenty of scope to deepen our understanding of default vs inflation financing in a sovereign debt crisis.

8.1.2 Better Rules

Analyses of optimal monetary and fiscal policy rules in approximated economies is quite clear about the kinds of simple rules that can mimic the Ramsey policy. Fairly aggressive inflation targeting using an inertial Taylor rule, coupled with a passive fiscal policy that

[bh] See Lane and Milesi-Ferretti (2001) for the first issue of a dataset of external portfolios and Devereux and Sutherland (2011) for a numerical method to endogenously embed such positions in open-economy macro models.

very gradually stabilizes debt, comes close to achieving the welfare levels that the Ramsey policy acquires (Schmitt-Grohé and Uribe, 2007; Kirsanova and Wren-Lewis, 2012). The nonlinear solutions to the optimal policy problem that this chapter described reveal that the policy mix depends crucially on both the level of debt and its maturity. With high levels of short-maturity debt, it is optimal to use monetary policy to stabilize debt and adjust distortionary taxation to mitigate the inflationary consequences of such a policy. This suggests that there may be a family of simple implementable rules which could improve welfare by introducing a degree of state-dependence to the policy mix.

Similarly, studies often seek to assess the importance of automatic stabilizers by adding output to the fiscal rules. Kliem and Kriwoluzky (2014) argue, though, that this is not the most data-coherent specification of policy behavior and that rules conditioned on other macroeconomic variables better capture the cyclical properties of fiscal instruments. Those proposed rules also improve welfare in DSGE models. Taken together, this suggests that there is scope for extending the range of simple rules considered in the literature to find alternatives that are both empirically and normatively more appealing.

8.1.3 Strategic Interactions

Estimates of regime-switching policies find that the policy mix is not always aligned with either regime M or regime F. There are also periods in which policies are in conflict—either doubly active or doubly passive. Introducing strategic interactions between policy authorities into optimal policy analysis may help to put theory in better line with data. Literature that looks at such interactions often relies on linear-quadratic approximation or simplifying assumptions to obtain tractable results.[bi] Blake and Kirsanova (2011) consider the desirability of central bank conservatism in a standard new Keynesian economy augmented with fiscal policy and an associated independent fiscal policymaker. They consider three forms of strategic interaction: either monetary or fiscal leadership, where the leader anticipates the response of the follower, or a Nash equilibrium between the two policymakers. The striking result, which echoes Section 5.4 in which the monetary authority followed a Taylor rule while the fiscal authority optimized, is that central bank conservatism always reduces welfare. Blake and Kirsanova also find that the quantitative results depend on the level of debt around which the economy is linearized. This argues that such analyses could usefully be extended to a nonlinear framework to explore the state dependencies in the strategic monetary and fiscal policy interactions. How robust is the institutional policy design to the strategic interactions implied by independent fiscal and monetary policymakers? To what extent can such interactions explain the observed policy switches in empirical analyses based on simple ad hoc rules?

[bi] Adam and Billi (2008) and Dixit and Lambertini (2003) consider the strategic interactions between monetary and fiscal policymakers, although in abstracting from the existence of government debt they rule out the mechanisms that have been the focus of this chapter.

8.1.4 Political Economy

Theoretical work on optimal policy, particularly fiscal policy, often implies policy behavior that bears little resemblance to observed policy. Benigno and Woodford's (2004) and Schmitt-Grohé and Uribe's (2004) analyses of jointly optimal monetary and fiscal policies suggest that when the policymaker can make credible promises about future actions, the steady-state level of debt should follow a random walk—in response to shocks, debt will be allowed to rise permanently because the short-run costs of reducing debt exactly balance the long-run benefits. This policy prescription is clearly at odds with the mounting concerns over rising debt levels in several advanced economies, which have led the IMF to predict that most governments will be involved with consolidation efforts for several years. The expected pace of consolidation is particularly rapid in the economies that are subject to pressures in the financial markets from worries over fiscal sustainability (International Monetary Fund, 2011).

If instead we assume that policymakers cannot make credible promises about how they behave in the future—policy is constrained to be time-consistent—then the implied policy outcomes can be equally unconvincing: instead of implying that debt should permanently rise following negative fiscal shocks, the theory tends to imply that the policymaker will be tempted to aggressively reduce the debt stock, often at rates that far exceed those observed in practice (Leith and Wren-Lewis, 2013). In standard new Keynesian models, time-consistent policy will not only call for a rapid debt correction, but it will make the long-run equilibrium value of debt negative, as the fiscal authority seeks to accumulate a stock of assets to help offset other frictions in the economy. The analysis in this chapter and in Leeper et al. (2015a), by allowing for a realistically calibrated debt maturity structure, can plausibly slow the pace of fiscal adjustments to levels which are not obviously inconsistent with those observed. And by assuming that the fiscal policymaker discounts the future more highly than households, as a crude means of capturing the short-termism that political frictions can engender, Leeper et al. (2015a) find that the time-consistent policy can support reversion to plausible debt–GDP ratios.

Although an inability to commit can go some way toward explaining this discrepancy between actual policy and the normative prescriptions of the theoretical literature, it seems likely that the political dimensions of policymaking are also important. Political economy aspects of actual fiscal policy have recently been laid bare in the abandoning of fiscal rules in Europe during the financial crisis, the brinkmanship over the raising of the debt ceiling in the United States, and the withholding or awarding of bail-out funds to Greece and other Eurozone economies from the Troika composed of the European Commission, the ECB, and the IMF. In this vein the New Political Economy literature seeks to identify mechanisms that can explain the trends in debt–GDP levels in many developed economies in recent decades.

Alesina and Passalacqua (2016) identify several reasons why governments may pursue policies that raise government debt to suboptimally high levels: (1) fiscal illusion—voters misunderstand the budget identity and are enticed to vote for a party that supports unsustainable tax cuts or spending increases; (2) political business cycles—voters are unsure of the competence of potential governments, so fiscal policy can be used by incumbents to signal competence; (3) delayed stabilization—political factions squabble over who bears the costs of fiscal consolidations, thereby delaying debt stabilization; (4) debt as a strategic variable—political parties use debt to tie the hands of their political opponents when they are out of office; (5) bargaining over policy in heterogeneous legislatures; (6) rent seeking by politicians; and (7) intergenerational redistributions. Some of these mechanisms are more naturally located in majoritarian systems—for example, political business cycles and strategic use of debt—while others are more likely to be associated with continuous strategic interactions between political actors outside of election periods—for example, delayed stabilizations and bargaining within legislatures—which are a feature of proportional/multiparty systems or heterogeneity within parties under a two-party system.

This New Political Economy literature typically does not consider monetary and fiscal policy interactions of the type considered in this chapter, so there is a need to integrate the two literatures. Political conflict inherent in the conduct of fiscal policy may explain why it is possible to obtain a data-coherent optimal policy description of monetary policy—albeit with fluctuations in the degree of monetary policy conservatism—while a similar description for fiscal policy is less easily achieved with policy switching between active and passive rules, with only short-lived periods in which policy is optimal (Chen et al., 2015).

Despite the difficulty of allowing for strategic interactions between the monetary and fiscal policymakers, this may not be going far enough if we are to understand the evolution of the monetary–fiscal policy mix. While treating an independent central bank as a single policymaker may be an acceptable approximation, it is less obvious that fiscal policy is best described by the actions of a single benevolent policymaker. A longer term research goal is to tractably integrate the New Political Economy literature into the analysis of monetary and fiscal policy interactions. Can we explain the changing nature of those interactions?

Political frictions vary substantially across countries. For example, in the United States and the United Kingdom debt levels fell fairly consistently following World War II until the early 1980s, before expanding consistently under Republican administrations in the United States, while not having such a clear partisan pattern in the United Kingdom. The current Conservative government in the United Kingdom is promising an aggressive austerity policy which seeks to run a permanent surplus from 2017. Any use of political frictions to explain the dynamics of debt and other macro variables must also explain such cross-country differences, particularly since it is not obvious that U.S. Republicans and U.K. Conservatives have fundamentally different views on the optimal size of the state.

8.1.5 Money

By focusing on cashless economies we have side-stepped the literature that considers the role of inflation as a tool of public finance vs its impact on money as a medium of exchange (Phelps, 1973). More recent research finds that the nature of the time-consistency problem facing a policymaker who issues nominal debt can depend crucially on the effects of inflation on the transactions technology (Martin, 2009, 2011; Niemann et al., 2013). We have also ignored the central bank's balance sheet, which precludes an analysis of fiscal aspects of unconventional monetary policies which have been discussed in Sims (2013), Del Negro and Sims (2015), and Reis (2013, 2015). Analyzing such unconventional monetary policies or technological developments like virtual money within frameworks that allow for interactions between such developments and fiscal policy are obvious areas for further research.

8.2 Further Empirical Work

This section proposes several directions in which to take empirical work on monetary–fiscal interactions.

8.2.1 Data Needs

In the early days of real business cycle research, Prescott (1986) argued that "theory is ahead of measurement," and, in particular, that theory can guide the measurement of key economic time series. This rings especially true for research on how monetary and fiscal policies affect inflation. Empirical applications in which the debt valuation equation plays a central role require observations on objects that are not readily available: the market value of privately held government liabilities—explicit debt and other commitments—the maturity structure of that debt, actual and expected primary surpluses, and actual and expected real discount rates. Compiling such data across countries and across monetary–fiscal regimes is the first step in an empirical agenda on policy interactions.

8.2.2 Identifying Regime

Empirical work surveyed in Section 6 highlights the difficulties in distinguishing whether regime M, regime F, or some other regimes generated observed time series. It remains to thoroughly explore which features of private and policy behavior are critical for breaking the near observational equivalence of regimes. Surprisingly, little work experiments with alternative specifications of policy behavior, particularly in DSGE models. Instead, most researchers—including us—adopt the simple rules that have become "standard." There is ample room for such experimentation.

Closely related is Geweke's (2010) argument that models are inherently incomplete in the sense that they lack "some aspect of a joint distribution over all parameters, latent variables, and models under consideration [p. 3]." For example, central bank money-only models that follow Smets and Wouters (2007) impose a dogmatic prior that places zero probability mass on regime F parameters. This procedure rejects a priori regions of the parameter space that the work reviewed in Section 6.3 finds fit data equally well. As we have seen, monetary policy actions have very different impacts in regimes M and F, so it matters a great deal to a policymaker, who is using model output to reach decisions, whether regime F is even possible. It would be valuable to apply existing tools for confronting model uncertainty to issues of monetary–fiscal regime (Hansen and Sargent, 2007; Geweke, 2010).

A different angle on model fit pursues DeJong and Whiteman's (1991) idea to ask: what type of prior over policy parameters is needed to support the inference that regime M (or regime F) generated the data? This exercise elicits the strength of a researcher's beliefs about regime when the researcher chooses to focus solely on one possible monetary–fiscal mix.

8.2.3 Generalizing Regime Switching

Existing work that estimates DSGE models with recurring policy regime switching tends to make simplifying assumptions about the nature of both private behavior and the policy process. Those assumptions can be systematically relaxed to arrive at more general models usable for policy analysis. And the fit of the models needs to be scrutinized in the manner that, for example, Smets and Wouter's (2007) specification has been. Until the fit of switching models is carefully evaluated, fixed-regime DSGE models will continue to dominate in policy institutions.[bj]

Recent econometric innovations permit estimation of endogenous regime change (Chang et al., 2015a). That technique treats policy regime as a latent process akin to time-varying probabilities of regime change. Generalizations of those methods to multivariate settings with multiple regimes that switch nonsynchronously could be integrated with DSGE models in which agents learn about the prevailing regime. Setups like that could shed empirical light on endogenous interactions among monetary and fiscal regimes, such as those that arise from the strategic interactions and political economy dynamics that Sections 8.1.3 and 8.1.4 mention.[bk]

8.2.4 Historical Analyses

Friedman and Schwartz (1963a) set the standard for historical analyses of monetary policy. But fiscal policy plays almost no role in their narrative. Stein (1996) is an excellent account of the evolution of fiscal policy in the United States, but his goals are different,

[bj] Sims and Zha (2006) is an exception, though they consider only monetary switching.

[bk] Chang et al. (2015b) estimate single-equation models of U.S. monetary and fiscal behavior to infer how an endogenous switch in one policy's regime predicts and switch in the other policy's regime. Empirical work along these lines connects more clearly to theory than do estimates in which regimes change exogenously.

so he does not connect the fiscal actions on which he reports to macroeconomic activity. A thorough analysis of the monetary–fiscal history of a country that brings to bear modern macroeconomic theory is a bit ambitious, though sorely needed. Short of a "Monetary *and Fiscal* History" that parallels Friedman and Schwartz, there are a great many historical episodes that can be reinterpreted in light of monetary–fiscal interactions.

Across countries there have been many short- and long-lived periods in which central banks have pegged interest rates, yet inflation has remained stable, as Cochrane (2015) points out. This observation seems to contradict Friedman's (1968) warning that pegged rates produce ever-increasing inflation. Has fiscal behavior played a role in delivering stable prices during interest rate pegs?

It would be instructive to bring fiscal behavior explicitly into a reexamination of the gold standard. What are the fiscal requirements of maintaining a fixed parity under the classical gold standard? Or of resuming convertibility after a suspension? Bordo and Hautcoeur (2007) contrast the French and British experiences after they suspended during World War I. Bordo (2011) suggests that France adopted a passive monetary/active fiscal policy mix that lead to substantially larger price-level increases in France than in Britain, which pursued active monetary and passive fiscal policies.

What role has fiscal policy played in accommodating or ending deflationary episodes? These have been well documented—Temin and Wigmore (1990), Bernanke and James (1991), Bordo and Filardo (2005), and Velde (2009) for example—but in the absence of an analytical understanding of how fiscal policy behaves under a gold standard, discussions of policy interactions remain informal (Eggertsson, 2008; Jalil and Rua, 2015).

How have large runups of government debt been financed historically? Hall and Sargent (2011, 2014) have made substantial progress on this important question in recent years.[bl] Although historically most large debt expansions were associated with wars, advanced economies since the financial crisis—and quite possibly going forward—are experiencing nonwar-related debt growth. What does history teach about how policy can best respond to high levels of government debt?

8.3 A Final Word

Macroeconomists have an unfortunate history of arguing over whether monetary or fiscal policy in *the* primary force behind inflation.[bm] If a reader leaves this chapter with a single message, that message should be: the fiscal theory and the quantity theory—or its recent manifestation, the Wicksellian theory—are parts of a more general theory of price-level determination in which monetary and fiscal policies always interact with private-sector behavior to produce the equilibrium aggregate level of prices. Within a certain

[bl] But see also Bordo and White (1991) on the Napoleonic wars and Sargent and Velde (1995) on the French revolution.

[bm] See, for example, Andersen and Jordan (1968) or Friedman and Heller (1969).

parametric family of monetary and fiscal rules, the two seemingly distinct perspectives arise from different regions of the policy parameter space, but there is no sense in which one view is "right" and the other is "wrong." Ultimately, it is an empirical question whether we can discern whether and under what circumstances one view is the dominant factor in inflation dynamics.

We would also encourage macroeconomists to entertain the possibility that both views are "right" most of the time and that the process of price-level determination is more complex than benchmark theories have so far described.

ACKNOWLEDGMENTS

This chapter has benefited from collaborations and discussions with many coauthors and colleagues and we thank them. We also thank Jon Faust, Ding Liu, Jim Nason, Charles Nolan, Fei Tan, and Todd Walker for conversations and Bob Barsky, John Cochrane, John Taylor, and Harald Uhlig for comments.

REFERENCES

Adam, K., Billi, R.M., 2008. Monetary conservatism and fiscal policy. J. Monetary Econ. 55 (8), 1376–1388.

Aiyagari, S.R., Marcet, A., Sargent, T.J., Seppälä, J., 2002. Optimal taxation without state-contingent debt. J. Polit. Econ. 110 (6), 1220–1254.

Akitoby, B., Komatsuzaki, T., Binder, A., 2014. Inflation and public debt reversals in the G7 countries. IMF Working Paper No. 14/96, June.

Alesina, A., Passalacqua, A., 2016. The political economy of government debt. In: Taylor, J.B., Uhlig, H. (Eds.), Handbook of Macroeconomics, vol. 2B. Elsevier, Amsterdam, Netherlands, pp. 2605–2657.

Alvarez, F., Kehoe, P.J., Neumeyer, P.A., 2004. The time consistency of optimal monetary and fiscal policies. Econometrica 72 (2), 541–567.

Andersen, L.C., Jordan, J.L., 1968. Monetary and fiscal actions: a test of their relative importance in economic stabilization. Fed. Reserve Bank St. Louis Rev. November, 11–24.

Angeletos, G.M., 2002. Fiscal policy with non-contingent debt and the optimal maturity structure. Q. J. Econ. 117 (3), 1105–1131.

Auernheimer, L., Contreras, B., 1990, February. Control of the Interest Rate with a Government Budget Constraint: Determinacy of the Price Level and Other Results. Texas A&M University, College Station, TX.

Ball, L., Mazumder, S., 2011. Inflation dynamics and the great recession. Brookings Papers Econ. Act. Spring, 337–402.

Banco Central do Brasil, 2015. Inflation report. 17(4), December.

Barro, R.J., 1979. On the determination of the public debt. J. Polit. Econ. 87 (5), 940–971.

Barro, R.J., Gordon, D.B., 1983. A positive theory of monetary policy in a natural-rate model. J. Polit. Econ. 91 (4), 589–610.

Bassetto, M., 2002. A game-theoretic view of the fiscal theory of the price level. Econometrica 70 (6), 2167–2195.

Begg, D.K.H., Haque, B., 1984. A nominal interest rate rule and price level indeterminacy reconsidered. Greek Econ. Rev. 6 (1), 31–46.

Benhabib, J., Schmitt-Grohé, S., Uribe, M., 2001. The perils of Taylor rules. J. Econ. Theor. 96 (1–2), 40–69.

Benhabib, J., Schmitt-Grohé, S., Uribe, M., 2002. Avoiding liquidity traps. J. Polit. Econ. 110 (3), 535–563.

Benigno, P., Woodford, M., 2004. Optimal monetary and fiscal policy: a linear-quadratic approach. In: Gertler, M., Rogoff, K. (Eds.), NBER Macroeconomics Annual 2003. MIT Press, Cambridge, MA, pp. 271–333.

Bernanke, B., James, H., 1991. The gold standard, deflation, and financial crisis in the great depression: an international comparison. In: Hubbard, R.G. (Ed.), Financial Markets and Financial Crises. University of Chicago Press, Chicago, pp. 33–68.

Bi, H., Leeper, E.M., Leith, C., 2013. Uncertain fiscal consolidations. Econ. J. 123 (566), F31–F63.

Bi, H., Leeper, E.M., Leith, C., 2015. Financial Intermediation and Government Debt Default. University of Glasgow, Glasgow, Scotland.

Bianchi, F., 2012. Evolving monetary/fiscal policy mix in the United States. Am. Econ. Rev. Papers Proc. 101 (3), 167–172.

Bianchi, F., Ilut, C., 2014. Monetary/Fiscal Policy Mix and Agents' Beliefs. Duke University, Durham, NC.

Blake, A.P., Kirsanova, T., 2011. Inflation conservatism and monetary-fiscal interactions. Int. J. Central Bank. 7 (2), 41–83.

Bocola, L., 2016. The pass through of sovereign risk. J. Polit. Econ. forthcoming.

Bohn, H., 1988. Why do we have nominal government debt? J. Monetary Econ. 21 (1), 127–140.

Bohn, H., 1990. Tax smoothing with financial instruments. Am. Econ. Rev. 80, 1217–1230.

Bohn, H., 1998. The behavior of U.S. public debt and deficits. Q. J. Econ. 113 (3), 949–963.

Bordo, M., 2011. Comments on 'Perceptions and misperceptions of fiscal inflation'. Slides, Rutgers University, June.

Bordo, M., Filardo, A., 2005. Deflation and monetary policy in a historical perspective: remembering the past or being condemned to repeat it. Econ. Policy (October), 799–844.

Bordo, M.D., Hautcoeur, P.C., 2007. Why didn't France follow the British stabilisation after World War I? Eur. Rev. Econ. Hist. 11 (1), 3–37.

Bordo, M., White, E.N., 1991. A tale of two currencies: British and French finance during the napoleonic wars. J. Econ. Hist. 51 (2), 303–316.

Brunner, K., Meltzer, A.H., 1972. Money, debt, and economic activity. J. Polit. Econ. 80 (5), 951–977.

Brunnermeier, M.K., Sannikov, Y., 2013. Redistributive monetary policy. In: The Changing Policy Landscape. Federal Reserve Bank of Kansas City Economic Conference Proceedings, 2012 Jackson Hole Symposium, pp. 331–384.

Buera, F., Nicolini, J.P., 2004. Optimal maturity structure of government debt without state contingent bonds. J. Monetary Econ. 51 (3), 531–554.

Buiter, W.H., 2002. The fiscal theory of the price level: a critique. Econ. J. 112 (481), 459–480.

Burnside, C., Eichenbaum, M., Rebelo, S., 2001. Prospective deficits and the Asian currency crisis. J. Polit. Econ. 109 (6), 1155–1197.

Calvo, G.A., 1983. Staggered prices in a utility maximizing model. J. Monetary Econ. 12 (3), 383–398.

Calvo, G.A., Guidotti, P., 1992. Optimal maturity of nominal government debt. Int. Econ. Rev. 33 (4), 895–919.

Canzoneri, M.B., Cumby, R.E., Diba, B.T., 2001a. Fiscal discipline and exchange rate systems. Econ. J. 111 (474), 667–690.

Canzoneri, M.B., Cumby, R.E., Diba, B.T., 2001b. Is the price level determined by the needs of fiscal solvency? Am. Econ. Rev. 91 (5), 1221–1238.

Carvalho, C., Ferrero, A., 2014. What Explains Japan's Persistent Deflation? University of Oxford, Oxford, UK.

Chang, Y., Choi, Y., Park, J.Y., 2015a. Regime Switching Model with Endogenous Autoregressive Latent Factor. Indiana University, Bloomington, IN.

Chang, Y., Kwak, B., Leeper, E.M., 2015b. Monetary-Fiscal Interactions with Endogenous Regime Change. Indiana University, Bloomington, IN.

Chari, V.V., Christiano, L.J., Kehoe, P.J., 1994. Optimal fiscal policy in a business cycle model. J. Polit. Econ. 102 (4), 617–652.

Chen, X., Leeper, E.M., Leith, C., 2015. U.S. Monetary and Fiscal Policy: Conflict or Cooperation? University of Glasgow, Glasgow, Scotland.

Chung, H., Leeper, E.M., 2007. What has financed government debt? NBER Working Paper No. 13425, September.

Cochrane, J.H., 1999. A frictionless view of U.S. inflation. In: Bernanke, B.S., Rotemberg, J.J. (Eds.), NBER Macroeconomics Annual 1998, vol. 13. MIT Press, Cambridge, MA, pp. 323–384.

Cochrane, J.H., 2001. Long term debt and optimal policy in the fiscal theory of the price level. Econometrica 69 (1), 69–116.

Cochrane, J.H., 2005. Money as stock. J. Monetary Econ. 52 (3), 501–528.

Cochrane, J.H., 2011a. Determinacy and identification with Taylor rules. J. Polit. Econ. 119 (3), 565–615.

Cochrane, J.H., 2011b. Understanding policy in the great recession: some unpleasant fiscal arithmetic. Eur. Econ. Rev. 55 (1), 2–30.

Cochrane, J.H., 2014. Monetary policy with interest on reserves. J. Econ. Dyn. Control 49 (December), 74–108.

Cochrane, J.H., 2015. Do Higher Interest Rates Raise or Lower Inflation? Hoover Institution, Stanford, CA.

Congressional Budget Office, 2014. CBO's projection of federal interest payments. http://www.cbo.gov/publication/45684. September 3.

Congressional Budget Office, 2015. The Long-Term Budget Outlook. U.S. Congress, Washington, DC.

Daniel, B.C., 2001. The fiscal theory of the price level in an open economy. J. Monetary Econ. 48 (2), 293–308.

Davig, T., Leeper, E.M., 2006. Fluctuating macro policies and the fiscal theory. In: Acemoglu, D., Rogoff, K., Woodford, M. (Eds.), NBER Macroeconomics Annual, vol. 21. MIT Press, Cambridge, pp. 247–298.

Davig, T., Leeper, E.M., Walker, T.B., 2010. 'Unfunded liabilities' and uncertain fiscal financing. J. Monetary Econ. 57 (5), 600–619.

Davig, T., Leeper, E.M., Walker, T.B., 2011. Inflation and the fiscal limit. Eur. Econ. Rev. 55 (1), 31–47.

Debortoli, D., Nunes, R.C., Yared, P., 2014. Optimal Government Debt Maturity. Columbia University, New York, NY.

DeJong, D.N., Whiteman, C.H., 1991. Reconsidering 'trends and random walks in macroeconomic time series'. J. Monetary Econ. 28 (2), 221–254.

DeJong, D.N., Ingram, B.F., Whiteman, C.H., 1996. A Bayesian approach to calibration. J. Business Econ. Stat. 14 (1), 1–9.

Del Negro, M., Schorfheide, F., 2004. Priors from general equilibrium models for VARs. Int. Econ. Rev. 45 (2), 643–673.

Del Negro, M., Sims, C.A., 2015. When does a central bank's balance sheet require fiscal support? In: Goodfriend, M., Zin, S.E. (Eds.), Monetary Policy: An Unprecedented Predicament, Carnegie-Rochester-NYU Conference Series on Public Policy, vol. 73. Amsterdam, pp. 1–19.

Del Negro, M., Giannoni, M.P., Schorfheide, F., 2015. Inflation in the great recession and new Keynesian models. Am. Econ. J. Macroecon. 7 (1), 168–196.

D'Erasmo, P., Mendoza, E.G., Zhang, J., 2016. What is a sustainable public debt? In: Taylor, J.B., Uhlig, H. (Eds.), Handbook of Macroeconomics, vol. 2B. Elsevier, Amsterdam, Netherlands, pp. 2499–2603.

Devereux, M.B., Sutherland, A., 2011. Country portfolios in open economy macro models. J. Eur. Econ. Assoc. 9 (2), 337–369.

Dixit, A., Lambertini, L., 2003. Interactions of commitment and discretion in monetary and fiscal policies. Am. Econ. Rev. 93 (5), 1522–1542.

Dupor, B., 2000. Exchange rates and the fiscal theory of the price level. J. Monetary Econ. 45 (3), 613–630.

Eggertsson, G.B., 2008. Great expectations and the end of the depression. Am. Econ. Rev. 98 (4), 1476–1516.

Eichenbaum, M., 1992. Comment on 'interpreting the macroeconomic time series facts: the effects of monetary policy'. Eur. Econ. Rev. 36, 1001–1011.

Eusepi, S., Preston, B., 2013. Fiscal Foundations of Inflation: Imperfect Knowledge. Monash University, Melbourne, Australia.

Faraglia, E., Marcet, A., Scott, A., 2008. Fiscal insurance and debt management in OECD economies. Econ. J. 118 (527), 363–386.

Friedman, M., 1948. A monetary and fiscal framework for economic stability. Am. Econ. Rev. 38 (2), 245–264.

Friedman, M., 1968. The role of monetary policy. Am. Econ. Rev. 58 (1), 1–17.

Friedman, M., 1970. The Counter-Revolution in Monetary Theory. Institute of Economic Affairs, London.

Friedman, M., Heller, W.W., 1969. Monetary vs. Fiscal Policy—A Dialogue. W.W. Norton & Company, New York.

Friedman, M., Schwartz, A.J., 1963a. A Monetary History of the United States, 1867–1960. Princeton University Press, Princeton, NJ.

Friedman, M., Schwartz, A.J., 1963b. Money and business cycles. Rev. Econ. Stat. 45 (1 Pt. 2, Suppl.), 32–64.

Galí, J., 2008. Monetary Policy, Inflation, and the Business Cycle. Princeton University Press, Princeton, NJ.

Geweke, J., 2010. Complete and Incomplete Econometric Models. Princeton University Press, Princeton, NJ.

Ghosh, A., Kim, J.I., Mendoza, E.G., Ostry, J.D., Qureshi, M.S., 2012. Fiscal fatigue, fiscal space and debt sustainability in advanced economies. Econ. J. 123 (566), F4–F30.

Gonzalez-Astudillo, M., 2013. Monetary-fiscal policy interaction: interdependent policy rule coefficients. Finance and Economics Discussion Series No. 2013-58, Federal Reserve Board, July.

Gros, D., 2011. Speculative attacks within or outside a monetary union: default versus inflation. CEPS Policy Briefs, No. 257, November.

Hall, G.J., Sargent, T.J., 2011. Interest rate risk and other determinants of post-WWII U.S. government debt/GDP dynamics. Am. Econ. J. Macroecon. 3 (3), 1–27.

Hall, G.J., Sargent, T.J., 2014. Fiscal discriminations in three wars. In: Goodfriend, M., Zin, S.E. (Eds.), Fiscal Policy in the Presence of Debt Crises. Carnegie-Rochester-NYU Conference Series on Public Policy. J. Mon. Econ., vol. 61. Amsterdam, pp. 148–166.

Hall, R.E., 2011. The long slump. Am. Econ. Rev. 101 (2), 431–469.

Hansen, L.P., Sargent, T.J., 2007. Robustness. Princeton University Press, Princeton.

Hausman, J.K., Wieland, J.F., 2014. Abenomics: preliminary analysis and outlook. Brookings Papers Econ. Act. Spring, 1–63.

Hilscher, J., Raviv, A., Reis, R., 2014. Inflating away the debt? An empirical assessment. NBER Working Paper No. 20339, July.

Hur, J., 2013. Fiscal Financing and the Effects of Government Spending: A VAR Approach. California State University, Northridge.

Imakubo, K., Kojima, H., Nakajima, J., 2015. The natural yield curve: its concept and measurement. Bank of Japan Working Paper Series No. 15-E-5, June.

International Monetary Fund, 2011. Fiscal Monitor–Shifting Gears: Tacking Challenges on the Road to Fiscal Adjustment. IMF, Washington, DC.

Ito, T., 2006. Japanese monetary policy: 1998–2005 and beyond. In: Monetary Policy in Asia: Approaches and Implementation. Bank for International Settlements, pp. 105–132.

Ito, T., Mishkin, F.S., 2006. Two decades of Japanese monetary policy and the deflation problem. In: Rose, A.K., Ito, T. (Eds.), Monetary Policy Under Very Low Inflation in the Pacific Rim, NBER-EASE, vol. 15. University of Chicago Press, Chicago, pp. 131–193.

Jalil, A., Rua, G., 2015. Inflation Expectations and Recovery from the Depression in 1933: Evidence from the Narrative Record. Occidental College, Los Angeles, CA.

Kiley, M.T., 2015. What can the data tell us about the equilibrium real interest rate? Finance and Economics Discussion Series No. 2015-077, Federal Reserve Board, August.

Kim, S., 2003. Structural shocks and the fiscal theory of the price level in the sticky price model. Macroecon. Dyn. 7 (5), 759–782.

King, M., 1995. Commentary: monetary policy implications of greater fiscal discipline. In: Budget Deficits and Debt: Issues and OptionsFederal Reserve Bank of Kansas City Economic Conference Proceedings, 1995 Jackson Hole Symposium, pp. 171–183.

Kirsanova, T., Wren-Lewis, S., 2012. Optimal feedback on debt in an economy with nominal rigidities. Econ. J. 122 (559), 238–264.

Kliem, M., Kriwoluzky, A., 2014. Toward a Taylor rule for fiscal policy. Rev. Econ. Dyn. 17 (2), 294–302.

Kliem, M., Kriwoluzky, A., Sarferaz, S., 2015. Monetary-fiscal policy interaction and fiscal inflation: a tale of three countries. Eur. Econ. Rev. forthcoming.

Kliem, M., Kriwoluzky, A., Sarferaz, S., 2016. On the low-frequency relationship between public deficits and inflation. J. Appl. Econ. 31 (3), 566–583.

Kocherlakota, N., Phelan, C., 1999. Explaining the fiscal theory of the price level. Fed. Reserve Bank Minneapolis Q. Rev. 23, 14–23.

Kriwoluzky, A., Müller, G.J., Wolf, M., 2014. Exit Expectations in Currency Unions. University of Bonn, Bonn, Germany.

Krugman, P.R., 1998. It's Baaack: Japan's slump and the return of the liquidity trap. Brookings Papers Econ. Act. 2, 137–187.

Lagos, R., Wright, R., 2005. A unified framework for monetary theory and policy analysis. J. Polit. Econ. 113 (3), 463–484.

Lane, P.R., Milesi-Ferretti, G.M., 2001. The external wealth of nations: measures of foreign assets and liabilities for industrial and developing countries. J. Int. Econ. 55 (2), 263–294.

Leeper, E.M., 1989. Policy rules, information, and fiscal effects in a 'Ricardian' model. Federal Reserve Board, International Finance Discussion Paper No. 360, August.

Leeper, E.M., 1991. Equilibria under 'active' and 'passive' monetary and fiscal policies. J. Monetary Econ. 27 (1), 129–147.

Leeper, E.M., 2011. Anchors aweigh: how fiscal policy can undermine 'good' monetary policy. In: Céspedes, L.F., Chang, R., Saravia, D. (Eds.), Monetary Policy Under Financial Turbulence. Banco Central de Chile, Santiago, pp. 411–453.

Leeper, E.M., 2016. Fiscal analysis is darned hard. In: Ódor, ´L. (Ed.), Rethinking Fiscal Policy After the Crisis. Cambridge University Press, Cambridge, UK.

Leeper, E.M., Li, B., 2015. On the Bias in Estimates of Fiscal Policy Behavior. Indiana University, Bloomington, IN.

Leeper, E.M., Nason, J.M., 2014. Bringing financial stability into monetary policy. Center for Applied Economics and Policy Research Working Paper No. 2014-003, Indiana University, November.

Leeper, E.M., Sims, C.A., 1994. Toward a modern macroeconomic model usable for policy analysis. In: Fischer, S., Rotemberg, J.J. (Eds.), NBER Macroeconomics Annual. MIT Press, Cambridge, MA, pp. 81–118.

Leeper, E.M., Walker, T.B., 2013. Perceptions and misperceptions of fiscal inflation. In: Alesina, A., Giavazzi, F. (Eds.), Fiscal Policy After the Financial Crisis. University of Chicago Press, Chicago, pp. 255–299.

Leeper, E.M., Zhou, X., 2013. Inflation's role in optimal monetary-fiscal policy. NBER Working Paper No. 19686, November.

Leeper, E.M., Leith, C., Liu, D., 2015a. Optimal Time-Consistent Monetary, Fiscal and Debt Maturity Policy. University of Glasgow, Glasgow, Scotland.

Leeper, E.M., Traum, N., Walker, T.B., 2015b. Clearing up the fiscal multiplier morass. NBER Working Paper No. 21433, July.

Leith, C., Liu, D., 2014. The inflation bias under Calvo and Rotemberg pricing. University of Glasgow Working Paper No. 2014-6.

Leith, C., Wren-Lewis, S., 2006. Compatibility between monetary and fiscal policy under emu. Eur. Econ. Rev. 50 (6), 1529–1556.

Leith, C., Wren-Lewis, S., 2008. Interactions between monetary and fiscal policy under flexible exchange rates. J. Econ. Dyn. Control 32 (9), 2854–2882.

Leith, C., Wren-Lewis, S., 2013. Fiscal sustainability in a new Keynesian model. J Money Credit Bank. 45 (8), 1477–1516.

Ljungqvist, L., Sargent, T.J., 2004. Recursive Macroeconomic Theory, second ed. MIT Press, Cambridge, MA.

Loyo, E., 1997. Going International with the Fiscal Theory of the Price Level. Princeton University, Princeton, NJ.

Loyo, E., 1999. Tight Money Paradox on the Loose: A Fiscalist Hyperinflation. Harvard University, Cambridge, MA.

Lucas Jr., R.E., Stokey, N.L., 1983. Optimal fiscal and monetary policy in an economy without capital. J. Monetary Econ. 12 (1), 55–93.

Marcet, A., Scott, A., 2009. Debt and deficit fluctuations and the structure of bond markets. J. Econ. Theory 21 (1), 473–501.

Martin, F.M., 2009. A positive theory of government debt. Rev. Econ. Dyn. 12 (4), 608–631.

Martin, F.M., 2011. On the joint determination of fiscal and monetary policy. J. Monetary Econ. 58 (2), 132–145.

McCallum, B.T., 1984. Are bond-financed deficits inflationary? J. Polit. Econ. 92 (February), 123–135.

McCallum, B.T., 2001. Indeterminacy, bubbles, and the fiscal theory of price level determination. J. Monetary Econ. 47 (1), 19–30.

Mendoza, E.G., Ostry, J.D., 2008. International evidence on fiscal solvency: is fiscal policy 'responsible'? J. Monetary Econ. 55 (6), 1081–1093.

Missale, A., 1999. Public Debt Management. Oxford University Press, Oxford.

Niemann, S., Pichler, P., Sorger, G., 2013. Public debt, discretionary policy, and inflation persistence. J. Econ. Dyn. Control 37 (6), 1097–1109.

Obstfeld, M., Rogoff, K., 1983. Speculative hyperinflations in maximizing models: can we rule them out? J. Polit. Econ. 91 (4), 675–687.

Persson, M., Persson, T., Svensson, L.E.O., 1987. Time consistency of fiscal and monetary policy. Econometrica 55 (6), 1419–1431.

Persson, M., Persson, T., Svensson, L.E.O., 2006. Time consistency of fiscal and monetary policy: a solution. Econometrica 74 (1), 193–212.

Phelps, E.S., 1973. Inflation in the theory of public finance. Swedish J. Econ. 75 (1), 67–82.

Prescott, E.C., 1986. Theory ahead of business cycle measurement. Carnegie-Rochester Conference Series on Public Policy, North-Holland, pp. 11–44.

Reis, R., 2013. The mystique surrounding the central bank's balance sheet, applied to the European crisis. Am. Econ. Rev. Papers Proc. 103 (3), 135–140.

Reis, R., 2015. QE in the Future: The Central Bank's Balance Sheet in a Fiscal Crisis. Columbia University, New York, NY.

Rotemberg, J.J., 1982. Sticky prices in the United States. J. Polit. Econ. 90 (December), 1187–1211.

Rotemberg, J.J., 1996. Prices, output, and hours: an empirical analysis based on a sticky price model. J. Monetary Econ. 37 (June), 505–533.

Sargent, T.J., 1986. The ends of four big inflations. In: Sargent, T.J. (Ed.), Rational Expectations and Inflation. Harper & Row, New York.

Sargent, T.J., Velde, F.R., 1995. Macroeconomic features of the French revolution. J. Polit. Econ. 103 (3), 474–518.

Sargent, T.J., Wallace, N., 1981. Some unpleasant monetarist arithmetic. Fed. Reserve Bank Minneapolis Q. Rev. 5 (Fall), 1–17.

Schmitt-Grohé, S., Uribe, M., 2004. Optimal fiscal and monetary policy under sticky prices. J. Econ. Theor. 114 (2), 198–230.

Schmitt-Grohé, S., Uribe, M., 2007. Optimal simple and implementable monetary and fiscal rules. J. Monetary Econ. 54 (6), 1702–1725.

Shim, S.D., 1984. Inflation and the Government Budget Constraint: International Evidence. Department of Economics, University of Minnesota. Unpublished Ph.D. Dissertation, August.

Sims, C.A., 1972. Money, income, and causality. Am. Econ. Rev. 62 (4), 540–552.

Sims, C.A., 1992. Interpreting the macroeconomic time series facts: the effects of monetary policy. Eur. Econ. Rev. 36, 975–1000.

Sims, C.A., 1994. A simple model for study of the determination of the price level and the interaction of monetary and fiscal policy. Econ. Theor. 4 (3), 381–399.

Sims, C.A., 1997, September. Fiscal Foundations of Price Stability in Open Economies. Yale University, New Haven, CT.

Sims, C.A., 1998. Econometric implications of the government budget constraint. J. Econ. 83 (1–2), 9–19.

Sims, C.A., 1999a. Domestic currency denominated government debt as equity in the primary surplus. Presented at the August 1999 Meetings of the Latin American region of the Econometric Society.

Sims, C.A., 1999b. The precarious fiscal foundations of EMU. De Economist 147 (4), 415–436.

Sims, C.A., 2001. Fiscal consequences for mexico of adopting the dollar. Journal of Money, Credit and Banking 33 (2, Part 2), 597–616.

Sims, C.A., 2011. Stepping on a rake: the role of fiscal policy in the inflation of the 1970s. Eur. Econ. Rev. 55 (1), 48–56.

Sims, C.A., 2013. Paper money. Am. Econ. Rev. 103 (2), 563–584.

Sims, C.A., Zha, T., 2006. Were there regime switches in US monetary policy? Am. Econ. Rev. 96 (1), 54–81.

Smets, F., Wouters, R., 2007. Shocks and frictions in US business cycles: a Bayesian DSGE approach. Am. Econ. Rev. 97 (3), 586–606.

Stein, H., 1996. The Fiscal Revolution in America, second ed. revised AEI Press, Washington, DC.

Tan, F., 2014. Two Econometric Interpretations of U.S. Fiscal and Monetary Policy Interactions. Indiana University, Bloomington, IN.

Tan, F., 2015. An Analytical Approach to New Keynesian Models Under the Fiscal Theory. Indiana University, Bloomington, IN.

Tan, F., Walker, T.B., 2014. Solving Generalized Multivariate Linear Rational Expectations Models. Indiana University, Bloomington, IN.

Taylor, J.B., 1993. Discretion versus policy rules in practice. Carnegie-Rochester Conf. Series Publ. Policy 39, 195–214.

Taylor, J.B., 1995. Monetary policy implications of greater fiscal discipline. In: Budget Deficits and Debt: Issues and Options. Federal Reserve Bank of Kansas City Economic Conference Proceedings, 1995 Jackson Hole Symposium, pp. 151–170.

Temin, P., Wigmore, B.A., 1990. The end of one big deflation. Explorations Econ. Hist. 27 (4), 483–502.

The Economist, 2016. Irredeemable? A former star of the emerging world faces a lost decade. http://www.economist.com/news/briefing/21684778-formerstar-emerging-world-faces-lost-decade-irredeemable, January 2.

Tobin, J., 1963. An essay on the principles of debt management. In: Commission on Money and Credit (Ed.), Fiscal and Debt Management Policies. Prentice-Hall, Englewood Cliffs, NJ, pp. 143–218.

Tobin, J., 1980. Asset Accumulation and Economic Activity. University of Chicago Press, Chicago.

Traum, N., Yang, S.C.S., 2011. Monetary and fiscal policy interactions in the post-war U.S. Eur. Econ. Rev. 55 (1), 140–164.

Uribe, M., 2006. A fiscal theory of sovereign risk. J. Monetary Econ. 53 (8), 1857–1875.

Velde, F.R., 2009. Chronicle of a deflation unforetold. J. Polit. Econ. 117 (4), 591–634.

Wallace, N., 1981. A Modigliani-Miller theorem for open-market operations. Am. Econ. Rev. 71 (3), 267–274.

Woodford, M., 1995. Price-level determinacy without control of a monetary aggregate. Carnegie-Rochester Conf. Series Publ. Policy 43, 1–46.

Woodford, M., 1998a. Control of the public debt: a requirement for price stability? In: Calvo, G., King, M. (Eds.), The Debt Burden and Its Consequences for Monetary Policy. St. Martin's Press, New York, pp. 117–154.

Woodford, M., 1998b. Public Debt and the Price Level. Princeton University, Princeton, NJ.

Woodford, M., 1999. Comment on Cochrane's 'a frictionless view of U.S. inflation'. In: Bernanke, B.S., Rotemberg, J.J. (Eds.), NBER Macroeconomics Annual 1998, vol. 13. MIT Press, Cambridge, MA, pp. 390–419.

Woodford, M., 2001. Fiscal requirements for price stability. J. Money Credit Bank. 33 (3), 669–728.

Woodford, M., 2003. Interest and Prices: Foundations of a Theory of Monetary Policy. Princeton University Press, Princeton, NJ.

CHAPTER 31

Fiscal Multipliers: Liquidity Traps and Currency Unions[☆]

E. Farhi[*], I. Werning[†]
[*]Harvard University, Cambridge, MA, United States
[†]MIT, Cambridge, MA, United States

Contents

[☆] We thank the editors John Taylor and Harald Uhlig for detailed comments, as well as suggestions and comments by Gabriel Chodorow-Reich, Jon Steinsson, and Michael Weber.

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.06.006

Abstract

We provide explicit solutions for government spending multipliers during a liquidity trap and within a fixed exchange regime using standard closed and open-economy New Keynesian models. We confirm the potential for large multipliers during liquidity traps. For a currency union, we show that self-financed multipliers are small, always below unity, unless the accompanying tax adjustments involve substantial static redistribution from low to high marginal propensity to consume agents, or dynamic redistribution from future to present non-Ricardian agents. But outside-financed multipliers which require no domestic tax adjustment can be large, especially when the average marginal propensity to consume on domestic goods is high or when government spending shocks are very persistent. Our solutions are relevant for local and national multipliers, providing insight into the economic mechanisms at work as well as the testable implications of these models.

Keywords

Currency unions, Non-Ricardian effects, Open economy model, Liquidity traps, New Keynesian effects

JEL Classification Code

E62

1. INTRODUCTION

Economists generally agree that macroeconomic stabilization should be handled first and foremost by monetary policy. Yet monetary policy can run into constraints that impair its effectiveness. For example, the economy may find itself in a liquidity trap, where interest rates hit zero, preventing further reductions in the interest rate. Similarly, countries that belong to currency unions, or states within a country, do not have the option of an independent monetary policy. Some economists advocate for fiscal policy to fill this void, increasing government spending to stimulate the economy. Others disagree, and the issue remains deeply controversial, as evidenced by vigorous debates on the magnitude of fiscal multipliers. No doubt, this situation stems partly from the lack of definitive empirical evidence, but, in our view, the absence of clear theoretical benchmarks also plays an important role. Although various recent contributions have substantially furthered our understanding, to date, the implications of standard macroeconomic models have not been fully worked out. This is the goal of this chapter. By clarifying the theoretical mechanisms in a unified way, we hope that it will help stimulate more research to validate or invalidate different aspects of the models.

We solve for the response of the economy to changes in the path for government spending during liquidity traps or within currency unions using standard New Keynesian closed and open-economy monetary models. A number of features distinguish our approach and contribution. First, our approach departs from the existing literature by focusing on *fiscal multipliers* that encapsulate the effects of spending for any path for government spending, instead of solving for a particular multiplier associated with the expansion of a single benchmark path for spending (eg, an autoregressive shock process to spending). Second, we obtain simple closed-form solutions for these multipliers. The more explicit and detailed expressions help us uncover the precise mechanisms underlying the effects of fiscal policy and allow us to deliver several new results.

Third, our analysis confirms that constraints on monetary policy are crucial, but also highlights that the nature of the constraint is also important. In particular, we draw a sharp contrast between a liquidity trap, with a binding zero-lower bound, and a currency union, with a fixed exchange rate.

Finally, in addition to nominal rigidities and constraints on monetary policy, we stress the importance of incorporating financial frictions for the analysis of fiscal policy. We do so by extending the benchmark models to include both incomplete markets and non-Ricardian borrowing constrained consumers, allowing for high and heterogeneous marginal propensities to consume out of current income. These financial market imperfections may be especially relevant in the aftermath of a financial crisis, situations where fiscal stimulus is often considered.

Our analysis has obvious implications for the interpretation of recent empirical studies on national and local multipliers. The empirical literature adopts different definitions of summary fiscal multipliers. For example, one popular notion used in many empirical studies consists in computing the ratio of some (discounted or not) average of the impulse responses of output and government spending in response to an innovation in government spending, up to some horizon (in practice 2 or 3 years). We show how our results can be used to compute such numbers analytically, and also discuss alternative definitions of summary fiscal multipliers.

Our results confirm that, in these standard models, fiscal policy can be especially potent during a liquidity trap. In the standard Ricardian model, the multiplier for output is always greater than one. We explicit the way in which the mechanism works through inflation. Higher government spending during a liquidity trap stimulates inflation. With fixed nominal interest rates, this reduces real interest rates which increases current private consumption. The increase in consumption in turn leads to more inflation, creating a feedback loop. The fiscal multiplier is increasing in the degree of price flexibility, which is intuitive given that the mechanism relies on the response of inflation. We show that in the model, backloading spending leads to larger effects; the rationale is that inflation then has more time to affect spending decisions.

For a country or region in a currency union, by contrast, government spending is less effective at increasing output. In particular, in the standard Ricardian model, we show that private consumption is crowded out by government spending, so that the multiplier is less than one. Moreover, price flexibility diminishes the effectiveness of spending, instead of increasing it. We explain this result using a simple argument that illustrates its robustness. Government spending leads to inflation in domestically produced goods and this loss in competitiveness depresses private spending.

It may seem surprising that fiscal multipliers are less than one when the exchange rate is fixed, contrasting with multipliers above one in liquidity traps. We show that even though in both cases the nominal interest rate is fixed, there is a crucial difference: a fixed exchange rate implies a fixed nominal interest rate, but the reverse is not true. Indeed, we prove that the liquidity trap analysis implicitly combines a shock to government spending with a one-off devaluation. The positive response of consumption relies entirely on this devaluation. A currency union rules out such a devaluation, explaining the difference in the response of consumption.

In the context of a country in a currency union, our results uncover the importance of transfers from outside—from other countries or regions. In the short run, when prices have not fully adjusted, positive transfers from outside increase the demand for home goods, stimulating output. We compute "transfer multipliers" that capture the response of the economy to such transfers. We show that these multipliers may be large when there is a high degree of home bias (ie, low degree of openness).

Note that the analysis of outside transfers requires some form of market incompleteness. Otherwise, with complete financial markets, any outside transfer would be completely undone by private insurance arrangements with outsiders. Such an extreme offset is unlikely to be realistic. Thus, we modify the standard open-economy model, which assumes complete markets, to consider the case with incomplete markets.

Understanding the effect of outside transfers is important because such transfers are often tied to government spending. This is relevant for the literature estimating local multipliers, which exploits cross-sectional variation, examining the effects of government spending across regions, states, or municipalities, within a country. In the US federal military spending allocated to a particular state is financed by the country as a whole. The same is true for exogenous differences, due to idiosyncratic provisions in the law, in the distribution of a federal stimulus package. Likewise, idiosyncratic portfolio returns accruing to a particular state's coffers represent a windfall for this state against the rest.

When changes in spending are financed by such outside transfers, the associated multipliers are a combination of self-financed multipliers and transfer multipliers. As a result, multipliers may be substantially larger than one even in a currency union. This difference is more significant when the degree of home bias is large, since this increases the marginal propensity to spend on home produced goods.

The degree of persistence in government spending is also important. Because agents seek to smooth consumption over time, the more temporary the government spending shock, the more the per-period transfer that accompanies the increase in spending is saved in anticipation of lower per-period transfers in the future. As a result, the difference in the effects on current output between outside-financed and self-financed government spending can be large for relatively persistent shocks, but may be small if shocks are relatively temporary. However, as we shall see, this distinction is blurred in the presence of liquidity constraints.

We explore non-Ricardian effects from fiscal policy by introducing hand-to-mouth consumers in addition to permanent income consumers. We think of this as a tractable way of modeling liquidity constraints. Both in a liquidity trap and in a currency union, government spending now has additional effects because of the differences in marginal propensities to consume of both groups of agents.

First, the incidence of taxes across these two groups matters, and redistribution from low marginal propensity to consume permanent-income agents to high marginal propensity to consume hand-to-mouth agents increases output. Second, since the model is non-Ricardian, the timing of taxes matters.

Both these effects can play a role independently of government spending. Indeed, one may consider tax changes without any change in government spending. However, changes in government spending must be accompanied by changes in taxes. As a result, whether government spending is, at the margin, debt-financed or tax-financed matters. Likewise, the distributional makeup of tax changes, across marginal propensities to consume, also matters. These effects can potentially substantially increase fiscal multipliers, both in liquidity traps and for countries or regions in a currency union. In particular, they may raise the multipliers above one for a region within a currency union.

Most importantly, liquidity constraints significantly magnify the difference between self-financed and outside-financed fiscal multipliers for temporary government spending shocks. Intuitively, a higher marginal propensity to consume implies that a greater part of the outside transfer is spent in the short run, contributing towards an increase in fiscal multipliers.

Overall, this discussion brings back the old Keynesian emphasis on the marginal propensity to consume. In particular, for temporary government spending shocks, the difference between self-financed and outside-financed fiscal multipliers is large when the average marginal propensity to consume on domestic goods is large—either due to a large number of liquidity constrained agents or due to a high degree of home bias in spending.

Finally, we show how to bridge our results for small open economies in a currency union and closed economies in a liquidity trap by simultaneously considering the effects government spending in all the countries within a currency union, depending on whether the currency union is in a liquidity trap or whether the central bank of the union can target inflation by adjusting interest rates.

Related Literature

Our chapter is related to several strands of theoretical and empirical literatures. We will discuss those that are most closely related.

We contribute to the literature that studies fiscal policy in the New Keynesian model in liquidity traps. Eggertsson (2011), Woodford (2011), and Christiano et al. (2011) show that fiscal multipliers can be large at the zero lower bound, while Werning (2012) studies optimal government spending with and without commitment to monetary policy. Gali and Monacelli (2008) study optimal fiscal policy in a currency union, but they conduct an exclusively normative analysis and do not compute fiscal multipliers. The results and simulations reported in Corsetti et al. (2011), Nakamura and Steinsson (2011), and Erceg and Linde (2012) show that fiscal multipliers are generally below one under fixed exchange rates yet higher than under flexible exchange rates (away from the zero bound), somewhat validating the conventional Mundell–Flemming view that fiscal policy is more effective with fixed exchange rates (see, eg, Dornbusch, 1980). Our solutions extend these results and help sharpen the intuition for them, by discussing the role of implicit devaluations and transfers. Gali et al. (2007) introduce hand-to-mouth consumers and study the effects of government spending under a Taylor rule in a closed economy. Our setup extends such an analysis to liquidity traps and currency unions in an open economy. Cook and Devereux (2011) study the spillover effects of fiscal policy in open economy models of the liquidity trap. We also examine this question but focus on a different context, that of a currency union, depending on whether it is or not in a liquidity trap.

Our chapter is also related to a large empirical literature on fiscal multipliers. Estimating national fiscal multipliers poses serious empirical challenges. The main difficulties arise from the endogeneity of government spending, the formation of expectations about future tax and spending policies, and the reaction of monetary policy. Most of the literature tries to resolve these difficulties by resorting to Structural VARs. Some papers use military spending as an instrument for government spending. The relevant empirical literature is very large, so we refer the reader to Ramey (2011) for a recent survey. Estimating fiscal multipliers in liquidity traps is nearly impossible because liquidity traps are rare. The closest substitute is provided by estimates that condition of the level of economic activity. Some authors (see, eg, Gordon and Krenn, 2010; Auerbach and Gorodnichenko, 2012) estimate substantially larger national multipliers during deep recessions, but the magnitude of these differential effects remains debated (see, eg, Barro and Redlick, 2009).

States or regions within a country offer an attractive alternative with plausible exogenous variations in spending. Indeed the literature on local multipliers has recently been very active, with contributions by Clemens and Miran (2010), Cohen et al. (2010), Serrato and Wingender (2010), Shoag (2010), Acconcia et al. (2011), Chodorow-Reich et al. (2011), Fishback and Kachanovskaya (2010), and Nakamura and Steinsson (2011). These papers tend to find large multipliers. Our chapter helps interpret these findings. Government spending at the local level in these experiments is generally tied to transfers from outside. It follows that these estimates may be interpreted as combining spending and transfer multipliers, as we define them here.

2. MULTIPLIERS AND SUMMARY MULTIPLIERS

We first set the stage by taking a purely statistical perspective and use it discuss the connection between theory and empirical work.

Suppose one has isolated a relationship between output and government spending encoded in the dynamic response of both variables to a particular structural shock of interest. One may then summarize this relationship into a single "fiscal multiplier" number in a number of ways. Of course, the entire impulse response contains strictly more information, but the multiplier may be a convenient way to summarize it. In the rest of this chapter, we derive the response of output to *any* spending shock for a set of standard macroeconomic models. The implications of each model are encoded in a set of coefficients or loadings, which can be mapped into dynamic responses to output for any impulse from spending.

2.1 Responses and Shocks

2.1.1 Impulse Responses

Suppose we have two time series $\{\hat{g}_t, \hat{y}_t\}$ for government spending and output respectively and that these series (after detrending) are stationary. Assume we can write these two series as a linear function of current and past shocks

$$\hat{g}_t = \hat{A}^g(L)\hat{\varepsilon}_t = \sum_{j=1}^{J} A^{gj}(L)\varepsilon_t^j = \sum_{j=1}^{J}\sum_{k=0}^{\infty}\psi_k^{gj}\varepsilon_{t-k}^j$$

$$\hat{y}_t = \hat{A}^y(L)\hat{\varepsilon}_t = \sum_{j=1}^{J} A^{yj}(L)\varepsilon_t^j = \sum_{j=1}^{J}\sum_{k=0}^{\infty}\psi_k^{yj}\varepsilon_{t-k}^j$$

where the vector of shocks $\hat{\varepsilon}_t = (\varepsilon_t^1, \varepsilon_t^2, \ldots, \varepsilon_t^J)'$ have zero mean and are uncorrelated over time, $\mathbb{E}[\varepsilon_t] = 0$ and $\mathbb{E}[\hat{\varepsilon}_t \hat{\varepsilon}_s'] = 0$ for $t \neq s$. Let us next isolate the effect of one particular shock $j \in J$ and define the components $\{g_t, y_t\}$ explained by this shock. Dropping the j subscript we write this as

$$g_t = A^g(L)\varepsilon_t = \sum_{k=0}^{\infty} \psi_k^g \varepsilon_{t-k} \tag{1a}$$

$$y_t = A^y(L)\varepsilon_t = \sum_{k=0}^{\infty} \psi_k^y \varepsilon_{t-k} \tag{1b}$$

where ε_t is a scalar shock with zero mean and is uncorrelated over time, $\mathbb{E}[\varepsilon_t] = 0$ and $\mathbb{E}[\varepsilon_t \varepsilon_s] = 0$ for $t \neq s$. The natural interpretation is that this particular shock, ε_t, is an exogenous structural shock to government spending. The coefficients $\{\psi_k^i\}$ are the impulse response functions (IRFs) to this shock. The responses can then be interpreted as encompassing a causal relationship. Strictly speaking, however, most of the discussion below does not require this interpretation.

2.1.2 VARs and Instruments

One way to obtain the decomposition of the series described above is using a structural VAR approach. To see this, suppose the original variables $\hat{g}_t$ and $\hat{y}_t$ are part of a VAR, which may include $J-2$ other variables (eg, inflation and interest rates). Suppose ε_t is one of the shocks. By definition, this shock is white noise and is orthogonal to the remaining $J-1$ shocks in the VAR at all leads and lags. In practice, the shock ε_t may be identified using structural assumptions, such as short-run or long-run restrictions. Under appropriate conditions, the shock may then acquire the economic interpretation of a fiscal shock and the response to output can be interpreted as an estimate of the causal relationship between spending and output.

Alternatively, the decomposition may result from an external instrumental variable. Suppose we have a scalar time series $\{z_t\}$ and let the Wold representation of z_t be[a]

$$z_t = A^z(L)\varepsilon_t = \sum_{k=0}^{\infty} \psi_k^z \varepsilon_{t-k}.$$

Thus, the shock ε_t is defined and identified as the innovation from the Wold representation of the instrument z_t. Now project $(\hat{g}_t, \hat{y}_t)$ linearly onto contemporaneous and lagged values of z_t, obtaining the predictors g_t and y_t (with residuals $\tilde{g}_t$ and $\tilde{y}_t$). These can then be represented as in (1). Once again, if the instrument is deemed exogenous to other economic fundamental shocks, then this shock may acquire economic interpretation as a fiscal shock and the response of output and spending can be interpreted as an estimate of the causal relationship between these variables.

2.2 Summary Multipliers

The sequences $\{\psi_k^g, \psi_k^y\}$ provide a full characterization of the joint behavior of $\{y_t\}$ and $\{g_t\}$, with respect to the shock $\{\varepsilon_t\}$. Suppose one insists on summarizing this

[a] Abstracting from the deterministic component.

relationship by a single number, called a "fiscal multiplier." First define the contemporaneous multiplier

$$m_k = \frac{\psi_k^y}{\psi_k^g}$$

indexed by $k = 0, 1, \ldots$ A general summary multiplier may take a ratio of the form

$$M^y = \frac{\sum\limits_{k=0}^{\infty} \lambda_k^y \psi_k^y}{\sum\limits_{k=0}^{\infty} \lambda_k^g \psi_k^g} = \frac{\sum\limits_{k=0}^{\infty} \lambda_k^y \psi_k^g}{\sum\limits_{k=0}^{\infty} \lambda_k^g \psi_k^g} \sum_{k=0}^{\infty} m_k \omega_k$$

where $\omega_k = \lambda_k^y \psi_k^g / \sum_{k=0}^{\infty} \lambda_k^y \psi_k^g$ is a weight that adds up to unity. A simple case is to add up the unweighted the reaction over the first N periods,

$$M^y = \frac{\sum\limits_{k=0}^{N} \psi_k^y}{\sum\limits_{k=0}^{N} \psi_k^g} = \sum_{k=0}^{N} m_k \omega_k,$$

where $\omega_k = \psi_k^g / \sum_{k=0}^{N} \psi_k^g$.

2.2.1 Regression Based Summary Multipliers: OLS and IV

Another popular way to proceed in obtaining a summary fiscal multiplier is regress output on spending and to take the coefficient on spending as a summary multiplier. Consider the relationship

$$\hat{y}_t = \beta^{OLS} \hat{g}_t + u_t^{OLS},$$

where $\mathbb{E}[\hat{g}_t u_t^{OLS}] = 0$ and

$$\beta^{OLS} \equiv \frac{\mathbb{E}[\hat{g}_t \hat{y}_t]}{\mathbb{E}[\hat{g}_t^2]} = \frac{\sum\limits_{j=1}^{J} \sum\limits_{k=0}^{\infty} \psi_k^{yj} \psi_k^{gj}}{\sum\limits_{j=1}^{J} \sum\limits_{k=0}^{\infty} \left(\psi_k^{gj}\right)^2} = \sum_{j=1}^{J} \sum_{k=0}^{\infty} m_k^j \omega_k^j.$$

where

$$m_k^j = \frac{\psi_k^{yj}}{\psi_k^{gj}}, \qquad \omega_k^j \equiv \frac{(\psi_k^{gj})^2}{\sum\limits_{l=0}^{\infty} (\psi_l^{gj})^2}.$$

Thus, the population regression recovers a weighted average of the k-multipliers associated with each shock j.

Consider next an instrumental variable regression

$$\hat{y}_t = \beta^{IV}\hat{g}_t + u_t^{IV}$$

where $\mathbb{E}[z_t u_t^{IV}] = 0$ and

$$\beta^{IV} \equiv \beta^{OLS} \equiv \frac{\mathbb{E}[y_t z_t]}{\mathbb{E}[g_t z_t]} = \frac{\sum\limits_{k=0}^{\infty} \psi_k^y \psi_k^z}{\sum\limits_{k=0}^{\infty} \psi_k^g \psi_k^z} = \sum_{k=0}^{\infty} m_k \omega_k,$$

with weights

$$\omega_k \equiv \frac{\psi_k^g \psi_k^z}{\sum\limits_{l=0}^{\infty} \psi_l^g \psi_k^z}.$$

These weights are positive if ψ_k^g and ψ_k^z take the same sign.[b]

2.3 Connection to Models

As we will show, the implications of a model for fiscal spending can be encoded in a sequence of theoretical multipliers $\{\alpha_{t,k}\}$, where the element $\alpha_{t,k}$ represents the predicted response of output in period t to government spending in period k. This response is calculated as the first-order effect by linearizing the model.

What is the connection between $\{\alpha_{t,k}\}$ and the impulse responses $\{\psi_k^g\}$ and $\{\psi_k^g\}$ discussed above? Suppose we can interpret ε_t as an exogenous shock to the path for spending as summarized by $\{\psi_k^g\}$ and we can interpret the change in spending as a having causal endogenous response in output summarized by $\{\psi_t^y\}$. In the model both responses would be related by

$$\psi_k^y = \sum_{k'=0}^{\infty} \psi_{k'}^g \alpha_{k,k'},$$

for all $t = 0, 1, \ldots$ Given the theoretical multipliers, this relationship give us the output response $\{\psi_k^y\}$ for any given government spending response $\{\psi_t^g\}$.

Under what conditions can we invert this relationship and identify the theoretical multipliers $\{\alpha_{t,k}\}$ from the responses $\{\psi_k^g\}$ and $\{\psi_k^y\}$? For a single pair of $\{\psi_k^g\}$ and $\{\psi_k^y\}$ the answer is generally negative. For any given k the $\alpha_{k,\cdot}$ sequence is not identified: we can only identify the value of the sum $\sum_{k'=0}^{\infty} \psi_{k'}^g \{\alpha_{k,k'}\}$.

[b] In some cases, for example, Nakamura and Steinsson (2011), the IV regressions are run in differences. It is straightforward to adjust the calculations above in this case.

Without further information identification would only be possible if we had multiple responses, $\{\psi_k^g\}$ restrictions, $\{\psi_k^y\}$, that is, multiple spending shocks.

A special case obtains if the response is purely forward looking, as is the case in some of the simplest macroeconomic models. To see this, assume that $\alpha_{t,k} = \alpha_{0,k-t}$ for $k=t,t+1,\ldots$ and $\alpha_{t,k} = 0$ for $k=1,2,\ldots,t-1$. Then we have

$$\psi_t^y = \sum_{k=t}^{\infty} \psi_k^g \alpha_{0,\,k-t}.$$

Then we can identify the entire sequence $\{\alpha_{0,k-t}\}$ from the pair of sequences $\{\psi_k^g\}$ and $\{\psi_k^y\}$, provided we satisfy a standard rank condition (so that the set of sequences $\{\psi_{k-t}^g\}$ for $t \in \{0,1,\ldots\}$ are linearly independent).

3. A CLOSED ECONOMY

We consider a one-time shock to the current and future path of spending that is realized at the beginning of time $t=0$ that upsets the steady state. To simplify and focus on the impulse response to this shock, we abstract from ongoing uncertainty at other dates.[c] We adopt a continuous time framework. This is convenient for some calculations but is completely inessential to any of our results.

The remainder of this section specifies a standard New Keynesian model environment; readers familiar with this setting may wish to skip directly to Section 4.

Households

There is a representative household with preferences represented by the utility function

$$\int_0^{\infty} e^{-\rho t}\left[\frac{C_t^{1-\sigma}}{1-\sigma} + \chi\frac{G_t^{1-\sigma}}{1-\sigma} - \frac{N_t^{1+\phi}}{1+\phi}\right] dt,$$

where N_t is labor, and C_t is a consumption index defined by

$$C_t = \left(\int_0^1 C_t(j)^{\frac{\epsilon-1}{\epsilon}} dj\right)^{\frac{\epsilon}{\epsilon-1}},$$

where $j \in [0, 1]$ denotes an individual good variety. Thus, ϵ is the elasticity between varieties produced within a given country. We denote by $P_t(j)$ is the price of variety j, and by

$$P_t = \left(\int_0^1 P_t(j)^{1-\epsilon} dj\right)^{\frac{1}{1-\epsilon}}$$

the corresponding price index.

[c] Since we are interested in a first order approximation of the equilibrium response to shocks, which can be solved by studying the log-linearized model, the presence of ongoing uncertainty would not affect any of our calculation or conclusions (we have certainty equivalence).

Households seek to maximize their utility subject to the budget constraints

$$\dot{D}_t = i_t D_t - \int_0^1 P_t(j) C_t(j) dj + W_t N_t + \Pi_t + T_t$$

for $t \geq 0$ together with a no-Ponzi condition. In this equation, W_t is the nominal wage, Π_t represents nominal profits and T_t is a nominal lump sum transfer. The bond holdings of home agents are denoted by D_t and the nominal interest rate for the currency union is denoted by i_t.

Government

Government consumption G_t is an aggregate of varieties just as private consumption,

$$G_t = \left(\int_0^1 G_t(j)^{\frac{\epsilon-1}{\epsilon}} dj \right)^{\frac{\epsilon}{\epsilon-1}}.$$

For any level of expenditure $\int_0^1 P_t(j) G_t(j) dj$, the government splits its expenditure across these varieties to maximize G_t. Spending is financed by lump-sum taxes. Ricardian equivalence holds, so that the timing of these taxes is irrelevant.

Firms

A typical firm produces a differentiated good with a linear technology

$$Y_t(j) = A_t N_t(j),$$

where A_t is productivity in the home country.

We allow for a constant employment tax $1 + \tau^L$, so that real marginal cost is given by $\frac{1+\tau^L}{A_t}\frac{W_t}{P_t}$. We take this employment tax to be constant in our model, as in standard in the literature. The tax rate is set to offset the monopoly distortion so that $\tau^L = -\frac{1}{\varepsilon}$. However, none of our results hinge on this particular value.

We adopt the standard Calvo price-setting framework. In every moment a randomly flow ρ_δ of firms can reset their prices. Those firms that reset choose a reset price P_t^r to solve

$$\max_{P_t^r} \int_0^\infty e^{-\rho_\delta s - \int_0^s i_{t+z} dz} \left(P_t^r Y_{t+s|t} - (1+\tau^L) W_t \frac{Y_{t+s|t}}{A_t} \right),$$

where $Y_{t+k|t} = \left(\frac{P_t^r}{P_{t+k}} \right)^{-\epsilon} Y_{t+k}$, taking the sequences for W_t, Y_t and P_t as given.

3.1 Equilibrium Conditions

We now summarize equilibrium conditions for the home country. Market clearing in the goods and labor market requires that:

$$Y_t = C_t + G_t,$$

$$N_t = \frac{Y_t}{A_t}\Delta_t,$$

where Δ_t is an index of price dispersion $\Delta_t = \int_0^1 \left(\frac{P_{H,t}(j)}{P_{H,t}}\right)^{-\epsilon}$. The Euler equation

$$\sigma\frac{\dot{C}_t}{C_t} = i_t - \pi_t - \rho$$

ensures the agents' intertemporal optimization, where $\pi_t = \dot{P}_t/P_t$ is inflation.

The natural allocation is a reference allocation that prevails if prices are flexible and government consumption is held constant at its steady state value G. We denote the natural allocation with a bar over variables.

We omit the first-order conditions for the price-setting problem faced by firms here. We shall only analyze a log-linearized version of the model which collapses these equilibrium conditions into the New Keynesian Phillips curve presented below.

4. NATIONAL MULTIPLIERS IN A LIQUIDITY TRAP

To obtain multipliers, we study the log-linearized equilibrium conditions around the natural allocation with constant government spending. Define

$$c_t = (1-\mathcal{G})(\log(C_t) - \log(\bar{C}_t)) \approx \frac{C_t - \bar{C}_t}{Y},$$

$$y_t = \log Y_t - \log \bar{Y}_t \approx \frac{Y_t - \bar{Y}_t}{Y} \qquad g_t = \mathcal{G}(\log G_t - \log G) \approx \frac{G_t - G}{Y},$$

where $\mathcal{G} = \frac{G}{Y}$. So that we have, up to a first order approximation,

$$y_t = c_t + g_t.$$

The log linearized system is then

$$\dot{c}_t = \hat{\sigma}^{-1}(i_t - \pi_t - \bar{r}_t), \tag{2}$$

$$\dot{\pi}_t = \rho\pi_t - \kappa(c_t + (1-\xi)g_t), \tag{3}$$

where $\hat{\sigma} = \frac{\sigma}{1-\mathcal{G}}$, $\lambda = \rho_\delta(\rho + \rho_\delta)$, $\kappa = \lambda(\hat{\sigma} + \phi)$ and $\xi = \frac{\hat{\sigma}}{\hat{\sigma} + \phi}$. Eq. (2) is the Euler equation and Eq. (3) is the New Keynesian Philips curve. Here, $\bar{r}_t$ is the natural rate of interest, defined as the real interest rate that prevail at the natural allocation, ie, Eq. (2) with $c_t = 0$ for all $t \geq 0$ implies $i_t - \pi_t = \bar{r}_t$ for all $t \geq 0$.

It will prove useful to define the following two numbers ν and $\bar{\nu}$ (the eigenvalues of the system):

$$\nu = \frac{\rho - \sqrt{\rho^2 + 4\kappa\hat{\sigma}^{-1}}}{2} \qquad \bar{\nu} = \frac{\rho + \sqrt{\rho^2 + 4\kappa\hat{\sigma}^{-1}}}{2}.$$

If prices were completely flexible, then consumption and labor are determined in every period by two static conditions: the labor consumption condition and the resource constraint. Spending affects the solution and gives rise to the neoclassical multiplier $1 - \xi$, which is positive but less than 1 and entirely due to a wealth effect on labor supply.

From now on, we take as given a path for the interest rate $\{i_t\}$ summarizing monetary policy. To resolve or sidestep issues of multiplicity one can assume that there is a date T such that $c_t = g_t = \pi_t = 0$ and $i_t = \bar{r}_t$ for $t \geq T$.[d] A leading example is a liquidity trap scenario where $i_t = 0$ and $\bar{r}_t < 0$ for $t < T$. However, although this is a useful interpretation but is not required for the analysis below.

Remark 1 Suppose $c_T = 0$ for some date T, then

$$c_t = \int_t^T \left(i_{t+s} - \pi_{t+s} - \bar{r}_{t+s}\right) ds,$$

so that given the inflation path $\{\pi_t\}$ the consumption path $\{c_t\}$ is independent of the spending path $\{g_t\}$.

This remark highlights that the mechanism by which government spending affects consumption, in the New Keynesian model, is inflation which affects the real interest rate. One can draw two implications from this. First, other policy instruments that affect inflation, such as taxes, may have similarly policy effects. Second, empirical work on fiscal multipliers has not focused on the role inflation plays and it may be interesting to test the predicted connection between output and inflation present in New Keynesian models.

4.1 Fiscal Multipliers Solved

Since the system is linear it admits a closed form solution. We can express any solution with government spending as

$$c_t = \tilde{c}_t + \int_0^\infty \alpha_s^c g_{t+s} ds, \tag{4a}$$

[d] Note that T may be arbitrarily large and will have no impact on the solution provided below. Indeed, the characterization of the equilibrium is valid even without selecting an equilibrium this way: one just interprets c^* and π^* below any equilibrium in the set of equilibrium attained when $g_t = 0$ for all t. The solution then describes the entire set of equilibria for other spending paths $\{g_t\}$.

$$\pi_t = \tilde{\pi}_t + \int_0^\infty \alpha_s^\pi g_{t+s} ds, \tag{4b}$$

where $\{\tilde{c}_t, \tilde{\pi}_t\}$ are equilibria with $g_t = 0$ for all t. We focus on the integral term $\int_0^\infty \alpha_s^i g_{t+s} ds$ for $i = c$, π as a measure of the effects of fiscal policy $g \neq 0$. We assume the integrals are well defined, although we allow and discuss the case where it is $+\infty$ or $-\infty$ below.

Focusing on consumption, we call the sequence of coefficients $\{\alpha_s^c\}$ *fiscal multipliers*. It is crucial to note that these are *total private consumption multipliers* and not *output multipliers*. Indeed, output is given by

$$y_t = \tilde{y}_t + g_t + \int_0^\infty \alpha_s^c g_{t+s} ds.$$

Whereas the natural benchmark for consumption multipliers is 0, that for output multipliers is 1.

The coefficients α_s^c do not depend on calendar time t, nor do they depend on the interest rate paths $\{i_t\}$ and $\{r_t\}$. Thus, the impact on consumption or output, given by the term $\int_0^\infty \alpha_s^c g_{t+s} ds$, depends only on the future path for spending summarized weighted by $\{\alpha_s^c\}$.

There are two motivations for adopting $\int_0^\infty \alpha_s^c g_{t+s} ds$ as a measure of the impact of fiscal policy, one more practical, the other more conceptual.

1. The more practical motivation applies if the economy finds itself in a liquidity trap with nominal interest rates immobilized at zero, at least for some time. Fiscal multipliers $\{\alpha_s^c\}$ can then be used to predict the effects of fiscal policy. To see this, suppose the zero lower bound is binding until T so that $i_t = 0$ for $t < T$; suppose that after T monetary policy delivers an equilibrium with zero inflation, so that $\pi_t = 0$ for $t \geq T$. As is well known, the resulting equilibrium without government spending ($g_t = 0$ for all t) features a negative consumption gap and deflation: $\tilde{c}_t$, $\tilde{\pi}_t < 0$ for $t < T$ (see, eg, Werning, 2012).

   Now, consider a stimulus plan that attempts to improve this outcome by setting $g_t > 0$ for $t < T$ and $g_t = 0$ for $t \geq T$. Then $\int_0^\infty \alpha_s^c g_{t+s} ds = \int_0^{T-t} \alpha_s^c g_{t+s} ds$ is precisely the effect of the fiscal expansion on consumption c_t, relative to the outcome without the stimulus plan $\tilde{c}_t$.

   More generally, suppose that after the trap spending may be nonzero and that monetary may or may not be described as securing zero inflation. Even in this case, we may still use fiscal multipliers to measure the impact of fiscal policy during the liquidity trap: one can write $c_t = c_T + \int_0^{T-t} \alpha_s^c g_{t+s} ds$ for $t < T$, where the c_T encapsulates the combined effects of fiscal and monetary policy after the trap $t \geq T$.

2. More conceptually, our fiscal multipliers provide a natural decomposition of the effects of the fiscal policy, over what is attainable by monetary policy alone.

   Eqs. (4a) and (4b) characterize the entire set of equilibria for $g \neq 0$ by providing a one-to-one mapping between equilibria with $g = 0$. Both $\tilde{c}_t$ and $\tilde{\pi}_t$ are equilibria with

$g = 0$ and are affected by monetary policy, as summarized, among other things, by the interest rate path $\{i_t\}$.

We can represent these facts as a relationship between the set of equilibria with and without government spending,

$$\mathcal{E}_g = \mathcal{E}_0 + \alpha \cdot g,$$

where $\mathcal{E}_0$ represents the set of equilibria when $g_t = 0$ for all t, while $\mathcal{E}_g$ is the set of equilibria for a given path for spending $g = \{g_t\}$. Here $\alpha = \{\alpha_s^c, \alpha_s^\pi\}$ collects the fiscal multipliers and the cross product $\alpha \cdot g$ represents the integrals $\int_0^\infty \alpha_s^i g_{t+s} ds$ for $i = c, \pi$. The set $\mathcal{E}_g$ is a displaced version of $\mathcal{E}_0$ in the direction $\alpha \cdot g$. Each equilibrium point in $\mathcal{E}_0$ is shifted in parallel by $\alpha \cdot g$ to another equilibrium point in $\mathcal{E}_g$ and it shares the same nominal interest rate path $\{i_t\}$. This last fact is unimportant for this second conceptual motivation, since the focus is on comparing the two sets, not equilibrium points. Instead, the important issue is that $\alpha \cdot g$ measures the influence of government spending on the set of equilibria. This provides a conceptual motivation for studying the multipliers α, since they summarize this influence. In other words, without spending one can view monetary policy as selecting from the set $\mathcal{E}_0$, while with government spending monetary policy can choose from $\mathcal{E}_g$. The effects of fiscal policy on the new options is then precisely determined by the shift $\alpha \cdot g$. Fig. 1 represents this idea pictorially.[e]

Our first result delivers a closed-form solution for fiscal multipliers. Using this closed form one can characterize the multiplier quite tightly.

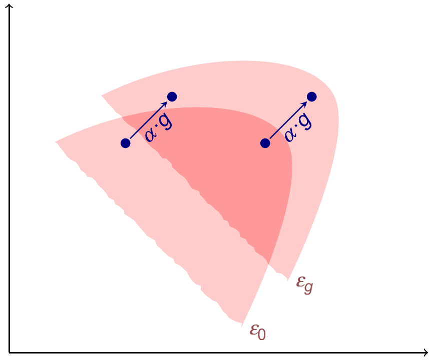


Fig. 1 A schematic depiction of the set of equilibria without government spending and the set of equilibria for a given spending path $\{g_t\}$.

[e] The figure is purposefully abstract and meant to convey the notion of a parallel shift only, so we have not labeled either axis and the shape of the sets is purely for illustrative purposes.

Proposition 1 (Closed Economy Multipliers) *The fiscal multipliers are given by*

$$\alpha_s^c = \hat{\sigma}^{-1}\kappa(1-\xi)e^{-\bar{\nu}s}\left(\frac{e^{(\bar{\nu}-\nu)s}-1}{\bar{\nu}-\nu}\right).$$

The instantaneous fiscal multiplier is zero $\alpha_0^c = 0$, *but the fiscal multipliers are positive, increasing and convex for large* s *so that* $\lim_{s\to\infty}\alpha_s^c = \infty$.

The left panel of Fig. 2 displays these consumption multipliers α_s^c as a function of s for a standard calibration. The proposition states that current spending has no effect on consumption: $\alpha_0^c = 0$. By implication, changes in spending that are very temporary are expected to have negligible effects on consumption and have an output multiplier that is near unity. As stated earlier, the effects of government spending on consumption work through inflation. Current spending does affect the current inflation rate and thus affects the growth rate of consumption. However, since this higher inflation is so short lived the lower growth rate for consumption has no significant stretch of time to impact the level of consumption.

In contrast, spending that takes place in the far future can have a very large impact. The further out into the future, the larger the impact, since α_s^c is increasing in s. Indeed, in the limit the effect becomes unbounded, since $\lim_{s\to\infty}\alpha_s^c = \infty$. The logic behind these results is that spending at $s > 0$ increases inflation over the entire interval of time $[0, s]$. This then lowers the real interest over this same time interval and lowers the growth rate of consumption. Since the long-run consumption level is fixed, the lower growth rate raises the level of consumption. This rise in consumption in turn leads to higher inflation, creating a feedback cycle. The larger the interval $[0, s]$ over which these effect have time to act, the larger is the effect on consumption.

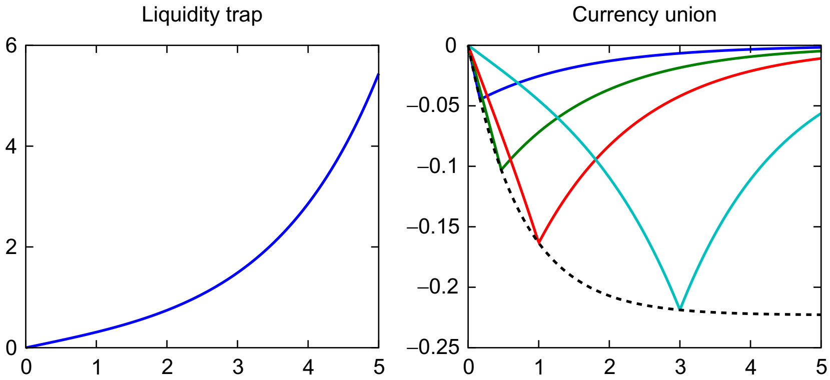


Fig. 2 Liquidity trap and currency union consumption multipliers α_s^c and $\alpha_{s-t}^{c,t,CM}$ as a function of s. Each curve for $\alpha_{s-t}^{c,t,CM}$ is plotted for different values of t ∈{0.25, 0.5, 1, 3}. The black dashed line shows the lower envelope. Parameters are $\sigma = 1$, $\eta = \gamma = 1$, $\varepsilon = 6$, $\phi = 3$, $\lambda = 0.14$, and $\alpha = 0.4$.

The fact that fiscal multipliers are unbounded as $s \to \infty$ stands in strong contrast to the zero multiplier at $s = 0$. It also has important implications. For example, a positive path for spending $\{g_t\}$ that is very backloaded can create a very large response for consumption. This is the case if the shock to spending is very persistent.

Example 1 (AR(1) Spending) Suppose $g_t = g e^{-\rho_g t}$, then if $\rho_g > -\nu > 0$ the response of consumption c_t is finite and given by

$$\int \alpha_s^c g e^{-\rho_g (t+s)} ds = \frac{\hat{\sigma}^{-1} \kappa (1-\xi)}{(\rho_g + \nu)(\rho_g + \bar{\nu})} g e^{-\rho_g t}.$$

The condition $\rho_g > -\nu > 0$ requires spending to revert to zero fast enough to prevent the integral from being infinite.

Some paths for spending imply an infinite value for $\int_0^\infty \alpha_s^c g_s ds$. For instance, this is the case in the example above when $\rho_g < -\nu$. How should one interpret such cases? Technically, this may invalidate our approximation. However, we think the correct economic conclusion to draw is that spending will have an explosive positive effect on consumption. One way to see this is to truncate the path of spending $\{g_t\}$, by setting $g_t = 0$ for all $t \geq T$ for some large T. This ensures that $\int_0^T \alpha_s^c g_s ds$ is finite but the response is guaranteed to be very large if the cutoff is large.

Next, we ask how fiscal multipliers are affected by the degree of price stickiness. Departures from the neoclassical benchmark, where the consumption multiplier is negative, require some stickiness in prices. Perhaps surprisingly, the resulting Keynesian effects turn out to be decreasing in the degree of price stickiness.

Proposition 2 (Price Stickiness) *The fiscal multipliers* $\{\alpha_s^c\}$

1. *are zero when prices are rigid* $\kappa = 0$*;*
2. *are increasing in price flexibility* κ*;*
3. *converge to infinity,* $\alpha_s^c \to \infty$*, in the limit as prices become fully flexible so that* $\kappa \to \infty$.

The logic for these results relies on the fact that spending acts on consumption through inflation. At one extreme, if prices were perfectly rigid then inflation would be fixed at zero and spending has no effect on consumption. As prices become more flexible spending has a greater impact on inflation and, hence, on consumption. Indeed, in the limit as prices become perfectly flexible, inflation becomes so responsive that the effects on consumption explode.

Recall that our fiscal multipliers are calculated under the assumption that the path for interest rates remains unchanged when spending rises. These results seem less counterintuitive when one realizes that such a monetary policy, insisting on keeping interest rates unchanged, may be deemed to be looser when prices are more flexible and inflation reacts more. Of course, this is precisely the relevant calculation when the economy finds itself in a liquidity trap, so that interest rates are up against the zero lower bound.

We capture backloading by a first order dominant shift in the cumulative distribution of spending for a given net present value of output. Backloading leads to a higher path of consumption at every point in time. This is simply because backloading gives more time to the feedback loop between output and inflation to play out.

When applied in a liquidity trap setting it is important to keep in mind the correct interpretation of this result. Our calculations compare spending paths at constant interest rates. In a liquidity trap, this translates to changes in spending before the end of the liquidity trap. If spending is delayed past the liquidity trap this affects consumption differently. For example, if after the end of the trap T monetary policy targets zero inflation, then government spending lowers consumption at T. This feeds back to consumption at $t = 0$, according to $c_t = c_T + \int_0^{T-t} \alpha_s^c g_{t+s} ds$ for $t < T$, lowering the impact on consumption and potentially reversing it. We conclude that backloading spending within the trap increases summary multipliers, but delaying spending past the trap reduce it.

4.2 Summary Fiscal Multipliers Again

Up to now we have discussed properties of fiscal multipliers $\{\alpha_s^c\}$. Usually, fiscal multipliers are portrayed as a single number that summarizes the impact of some change in spending on output or consumption, perhaps conditional on the state of the economy or monetary policy. This requires collapsing the entire sequence of fiscal multipliers $\{\alpha_s^c\}$ into a single number $\bar{\alpha}$, which we shall call a *summary fiscal multiplier,* such as

$$M^c = \frac{\int_0^\infty \lambda_t^c \int_0^\infty \alpha_s^c g_{t+s} ds\, dt}{\int_0^\infty \lambda_t^g g_t dt},$$

where $\{\lambda_t^c\}$ and $\{\lambda_t^g\}$ are weights. It is most natural to consider symmetric weights, with $\lambda_t^g = \lambda_t^c = \lambda_t$, which assume from now on. The simplest weight sets $\lambda_t = 1$ for $t \leq \tau$ and $\lambda_t = 0$ for $t > \tau$, which then computes the ratio of the total responses over the interval $[0, \tau]$. Another possibility is to set $\lambda_t = e^{-\rho t}$, to compute the ratio of the present value responses over the entire horizon.[f]

Note that since $y_t = c_t + g_t$ we have that the output multiplier (defined analogously to the consumption multiplier) is simply[g]

$$M^y = M^c + 1.$$

As this discussion makes clear there are many possibilities for summary multipliers and no universal criteria to select them. Instead, one can adapt the summary multiplier to the

[f] The empirical counterpart of such an infinite-horizon calculation is, however, impractical.

[g] That is, we define

$$M^y = \frac{\int_0^\infty \lambda_t \left(\int_0^\infty \alpha_s^c g_{t+s} ds + g_t\right) dt}{\int_0^\infty \lambda_t g_t dt}.$$

application and relevant policy at hand. The characterizations provided in the previous section have implications for any of these measures. Namely,

i if spending $\{g_t\}$ converges to being concentrated at $t = 0$ then $M^c \to 0$;

ii the more backloaded is government spending for a given net present value, the higher is M^c;

iii the multiplier M^c is increasing in flexibility, it is zero with rigid prices $\kappa = 0$ and goes to infinity in the limit of flexible prices $\kappa \to \infty$.

Example 2 Suppose we have an autoregressive spending path $g_t = ge^{-\rho_g t}$ for $\rho_g > 0$. The summary multiplier is independent of g_0 and given by

$$M^c = \frac{\int_0^\infty \lambda_t \int_0^\infty \alpha_s^c g_{t+s} ds\, dt}{\int_0^\infty \lambda_t g_t dt} = \frac{\int_0^\infty \lambda_t \int_0^\infty \alpha_s^c e^{-\rho_g(t+s)} ds\, dt}{\int_0^\infty \lambda_t e^{-\rho_g t} dt} = \int_0^\infty \alpha_s^c e^{-\rho_g s} ds.$$

Higher values of ρ_g shift weight towards the future. More persistence leads to higher summary multipliers.

4.3 Endogenous Spending: Policy Shocks vs Policy Rules

Up to now we have considered exogenous changes in government spending and their impact on output—a fiscal policy shock. Many stimulus policies, however, are best thought of as responding endogenously to the state of the economy—a fiscal policy rule.

Since the state of the economy depends on the model parameters, this implies that model parameters may play a double role when evaluating fiscal policy rules, as opposed to evaluating fiscal policy shocks.

In this short section we briefly touch on this issue using two examples. Formally, a change in parameters may affect both the structural fiscal multipliers $\{\alpha_t^c\}$, as we have discussed, and the path for government spending $\{g_t\}$. Both may have effects on output and summary fiscal multipliers.

Example 3 Christiano et al. (2011) compute summary fiscal multipliers in a liquidity trap. They assume a policy for government spending that increases spending by a constant amount as long the economy remains in the liquidity trap. They vary the degree of price flexibility and the duration of the liquidity trap and compute the fiscal multiplier (see Fig. 2).

Their summary multiplier is equivalent to computing the initial output response divided by the initial spending increase. Their results suggest that parameter values that make the recession worse also lead to larger multipliers. In some cases, this follows because the parameters affect the fiscal multipliers $\{\alpha_s^c\}$ directly. For example, this is the case for the degree of price flexibility κ. Higher price flexibility makes the recession worse and leads to higher fiscal multipliers, as shown in Proposition 2.

However, in other cases their conclusion rely on the indirect effects that these parameters have on the policy experiment $\{g_t\}$ itself. Indeed, this may affect summary multipliers even when our multipliers $\{\alpha_s^c\}$ are unchanged. Their setup features Poisson uncertainty regarding the length of the trap, but the same logic applies in a deterministic setting, when the liquidity trap has a known duration T.[h]

Suppose the economy is in a liquidity trap with zero interest rates for $t \leq T$ and returns to the natural allocation $c_t = g_t = 0$ for $t \geq T$. Consider fiscal policy interventions that increase spending during the trap, $g_t = g$ for $t \leq T$ and $g_t = 0$ for $t > T$. Higher T then leads to a deeper recession (see Werning, 2012) but has no effect on fiscal multipliers $\{\alpha_t^c\}$. However, the summary impact multiplier computed as

$$\frac{\int_0^T \alpha_s^c g \, ds}{g} = \int_0^T \alpha_s^c \, ds,$$

is increasing and convex in T. A longer liquidity trap increases this summary multiplier even though spending at any point in time is equally effective (α_s^c unchanged). It would be wrong to conclude that a stimulus plan with a fixed duration $\tau \leq T$ (a policy *shock*), such as a year or two, becomes more powerful when T increases. Rather, if $g_t = g$ for all $t \leq T$ (a policy rule) when T increases, then the effect on output is larger simply because the increase in T extends the time frame over which a fixed increase in spending g takes place, leading to an increase in the cumulative change in spending, Tg. Since cumulative spending increases, the impact effect would be larger even if, counter to the model, α_s^c were constant. Moreover, this effect is amplified because the extension backloads spending, and Proposition 1 shows that this is particularly effective since α_s^c is increasing in s.

Example 4 Another perspective is provided when g_t is set as a linear function of current consumption

$$g_t = -\Psi c_t,$$

for some $\Psi > 0$. Then the Phillips curve becomes

$$\dot{\pi}_t = \rho \pi_t - \kappa(c_t + (1-\xi)g_t) = \rho \pi_t - \kappa(1-(1-\xi)\Psi)c_t.$$

Suppose further that $\Psi = (1-\xi)^{-1}$, so that spending "fills the gap" and $c_t + (1-\xi)g_t = 0$. We maintain the assumption that $c_t = g_t = 0$ for $t \geq T$. Inflation is then zero for all $t \geq 0$ and the outcome for consumption is *as if* prices were completely rigid. Now, with this fiscal policy in place, consider different values for price flexibility κ. Neither the outcome for consumption $\{c_t\}$ nor the spending path $\{g_t\}$ depend on κ. Thus, in this special case, for given T, the fiscal rule can be interpreted as a fiscal shock, since it is independent of κ.

[h] Their parameter p, which represents the probability of remaining in the trap, has an effect similar to T in our deterministic setting.

However, the benchmark equilibrium outcome without spending, ie, $g_t = 0$, is decreasing in price flexibility κ (see Werning, 2012). Thus, fiscal policy has a greater effect on consumption when prices are more flexible. This is consistent with Proposition 2 regarding the effects of price flexibility on $\{\alpha_s^c\}$.

5. AN OPEN ECONOMY MODEL OF A CURRENCY UNION

We now turn to open economy models similar to Farhi and Werning (2012a,b) which in turn build on Gali and Monacelli (2005, 2008).

The model focuses on a continuum of regions or countries that share a common currency. One interpretation is that these regions are states or provinces within a country. Our analysis is then directly relevant to the literature estimating "local" multipliers, exploiting cross-sectional variation in spending behavior across states in the United States to estimate the effects on income and employment. Another interpretation is to member countries within a currency union, such as the European Monetary Union (EMU). Our analysis then sheds light on the debates over fiscal policy, stimulus vs austerity, for periphery countries.

For concreteness, from now on we will refer to these economic units (regions or countries) simply as countries. We focus on the effects around a symmetric steady state after a fiscal policy is realized in every country. A crucial ingredient is how private agents share risk internationally. We consider the two polar cases: (i) incomplete markets, where agents can only trade a risk-free bond; and (ii) complete markets with perfect risk sharing. These two market structures have different implications for fiscal multipliers.

5.1 Households

There is a continuum measure one of countries $i \in [0, 1]$. We focus attention on a single country, which we call "home" and can be thought of as a particular value $H \in [0, 1]$. We will focus on a one time shock, so that all uncertainty is realized at $t = 0$. Thus, we can describe the economy after the realization of the shock as a deterministic function of time.

In every country, there is a representative household with preferences represented by the utility function

$$\int_0^\infty e^{-\rho t}\left[\frac{C_t^{1-\sigma}}{1-\sigma} + \chi\frac{G_t^{1-\sigma}}{1-\sigma} - \frac{N_t^{1+\phi}}{1+\phi}\right]dt,$$

where N_t is labor, and C_t is a consumption index defined by

$$C_t = \left[(1-\alpha)^{\frac{1}{\eta}} C_{H,t}^{\frac{\eta-1}{\eta}} + \alpha^{\frac{1}{\eta}} C_{F,t}^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}},$$

where $C_{H,t}$ is an index of consumption of domestic goods given by

$$C_{H,t} = \left(\int_0^1 C_{H,t}(j)^{\frac{\epsilon-1}{\epsilon}} dj \right)^{\frac{\epsilon}{\epsilon-1}},$$

where $j \in [0, 1]$ denotes an individual good variety. Similarly, $C_{F,t}$ is a consumption index of imported goods given by

$$C_{F,t} = \left(\int_0^1 C_{i,t}^{\frac{\gamma-1}{\gamma}} di \right)^{\frac{\gamma}{\gamma-1}},$$

where $C_{i,t}$ is, in turn, an index of the consumption of varieties of goods imported from country i, given by

$$C_{i,t} = \left(\int_0^1 C_{i,t}(j)^{\frac{\epsilon-1}{\epsilon}} dj \right)^{\frac{\epsilon}{\epsilon-1}}.$$

Thus, ϵ is the elasticity between varieties produced within a given country, η the elasticity between domestic and foreign goods, and γ the elasticity between goods produced in different foreign countries. An important special case obtains when $\sigma = \eta = \gamma = 1$. We call this the Cole–Obstfeld case, in reference to Cole and Obstfeld (1991).

The parameter α indexes the degree of home bias, and can be interpreted as a measure of openness. Consider both extremes: as $\alpha \to 0$ the share of foreign goods vanishes; as $\alpha \to 1$ the share of home goods vanishes. Since the country is infinitesimal, the latter captures a very open economy without home bias; the former a closed economy barely trading with the outside world.

Households seek to maximize their utility subject to the budget constraints

$$\dot{D}_t = i_t D_t - \int_0^1 P_{H,t}(j) C_{H,t}(j) dj - \int_0^1 \int_0^1 P_{i,t}(j) C_{i,t}(j) dj di + W_t N_t + \Pi_t + T_t$$

for $t \geq 0$. In this equation, $P_{H,t}(j)$ is the price of domestic variety j, $P_{i,t}$ is the price of variety j imported from country i, W_t is the nominal wage, Π_t represents nominal profits and T_t is a nominal lump-sum transfer. All these variables are expressed in the common currency. The bond holdings of home agents is denoted by D_t and the common nominal interest rate within the union is denoted by i_t.

We sometimes allow for transfers across countries that are contingent on shocks. These transfers may be due to private arrangements in complete financial markets. or due to government arrangements. These transfers can accrue to the government or directly to the agents. This is irrelevant since lump-sum taxes are available. For example, we sometimes consider the assumption of complete markets where agents in different countries can perfectly share risks in a complete set of financial markets. Agents form international portfolios, the returns of which result in international transfers that are contingent on the realization of the shock. A different example is in

Section 8 where we consider government spending in the home country paid for by a transfer from the rest of the world. In this case, we have in mind a direct transfer to the government of the home country, or simply spending paid for and made by the rest of the world.

5.2 Government

Government consumption G_t is an aggregate of different varieties. Importantly, we assume that government spending is concentrated exclusively on domestic varieties

$$G_t = \left(\int_0^1 G_t(j)^{\frac{\epsilon-1}{\epsilon}} dj \right)^{\frac{\epsilon}{\epsilon-1}}.$$

For any level of expenditure $\int_0^1 P_{H,t}(j) G_t(j) dj$, the government splits its expenditure across varieties to maximize government consumption G_t. Spending is financed by lump-sum taxes. The timing of these taxes is irrelevant since Ricardian equivalence holds in our basic model. We only examine a potentially non-Ricardian setting in Section 7 where we introduce hand-to-mouth consumers into the model.

5.3 Firms

5.3.1 Technology

A typical firm in the home economy produces a differentiated good using a linear technology

$$Y_t(j) = A_{H,t} N_t(j),$$

where $A_{H,t}$ is productivity in the home country. We denote productivity in country i by $A_{i,t}$.

We allow for a constant employment tax $1 + \tau^L$, so that real marginal cost deflated by Home PPI is $\frac{1+\tau^L}{A_{H,t}} \frac{W_t}{P_{H,t}}$. We take this employment tax to be constant and set to offset the monopoly distortion so that $\tau^L = -\frac{1}{\varepsilon}$, as is standard in the literature. However, none of our results hinge on this particular value.

5.3.2 Price-Setting Assumptions

We assume that the Law of One Price holds so that at all times, the price of a given variety in different countries is identical once expressed in the same currency.

We adopt the Calvo price setting framework, where in every period, a randomly flow ρ_δ of firms can reset their prices. Those firms that get to reset their price choose a reset price P_t^r to solve

$$\max_{P_t^r} \int_0^\infty e^{-\rho_\delta s - \int_0^s i_{t+z} dz} \left(P_t^r Y_{t+s|t} - (1+\tau^L) W_t \frac{Y_{t+s|t}}{A_{H,t}} \right),$$

where $Y_{t+k|t} = \left(\frac{P_t^r}{P_{H,t+k}} \right)^{-\epsilon} Y_{t+k}$, taking the sequences for W_t, Y_t, and $P_{H,t}$ as given.

5.4 Terms of Trade and Real Exchange Rate

It is useful to define the following price indices: the home Consumer Price Index (CPI) is

$$P_t = [(1-\alpha) P_{H,t}^{1-\eta} + \alpha P_t^{*1-\eta}]^{\frac{1}{1-\eta}},$$

the home Producer Price Index (PPI)

$$P_{H,t} = \left[\int_0^1 P_{H,t}(j)^{1-\epsilon} dj \right]^{\frac{1}{1-\epsilon}},$$

and P_t^* is the price index for imported goods. The terms of trade are defined by

$$S_t = \frac{P_t^*}{P_{H,t}}.$$

Similarly let the real exchange rate be

$$\mathcal{Q}_t = \frac{P_t^*}{P_t}.$$

5.5 Equilibrium Conditions

We now summarize the equilibrium conditions. For simplicity of exposition, we focus on the case where all foreign countries are identical. Because agents face the same sequence of interest rates optimal consumption satisfies

$$C_t = \Theta C_t^* \mathcal{Q}_t^{\frac{1}{\sigma}},$$

where Θ is a relative Pareto weight which might depend on the realization of the shocks, and C_t^* is union-wide consumption. The goods market clearing condition is

$$Y_t = (1-\alpha) C_t \left(\frac{\mathcal{Q}_t}{S_t} \right)^{-\eta} + \alpha S_t^\gamma C_t^* + G_t.$$

We also have the labor market clearing condition

$$N_t = \frac{Y_t}{A_{H,t}} \Delta_t,$$

where Δ_t is an index of price dispersion $\Delta_t = \int_0^1 \left(\frac{P_{H,t}(j)}{P_{H,t}} \right)^{-\epsilon}$ and the Euler equation

$$\sigma\frac{\dot{C}_t}{C_t} = i_t - \pi_t - \rho,$$

where $\pi_t = \dot{P}_t/P_t$ is CPI inflation. Finally, we must include the country-wide budget constraint

$$N\dot{F}A_t = (P_{H,t}Y_t - P_tC_t) + i_tNFA_t,$$

where NFA_t is the country's net foreign assets at t, which for convenience, we measure in home numeraire. We impose a standard no-Ponzi condition, $e^{-\int_0^t i_s ds}NFA_t \rightarrow 0$ as $t \rightarrow \infty$.

Absent transfers or insurance across countries NFA_0 must be equal to zero. Instead, when markets are complete we require that $\Theta = 1$. We then solve for the initial value of NFA_0 that is needed, for each shock realization. This value can be interpreted as an insurance transfer from the rest of the world.

Finally with Calvo price setting we have the equations summarizing the first-order condition for optimal price setting. We omit these conditions since we will only analyze a log-linearized version of the model.

6. NATIONAL AND LOCAL FISCAL MULTIPLIERS IN CURRENCY UNIONS

To compute local multipliers, we study the log-linearized equilibrium conditions around a symmetric steady state with zero inflation. We denote the deviations of total private consumption (by domestic and foreigners), output, and public consumption on domestic goods relative to steady state output by

$$c_t = (1-\mathcal{G})(\log(Y_t - G_t) - \log(Y-G)) \approx \frac{Y_t - G_t - (Y-G)}{Y},$$

$$y_t = \log(Y_t) - \log(Y) \approx \frac{Y_t - Y}{Y} \qquad g_t = \mathcal{G}(\log G_t - \log G) \approx \frac{G_t - G}{Y},$$

where $\mathcal{G} = \frac{G}{Y}$ denotes the steady state share of government spending in output. Then we have, up to a first order approximation,

$$y_t = c_t + g_t,$$

Note that c_t does not represent private domestic total consumption (of home and foreign goods); instead it is private consumption (domestic and foreign) of domestic goods. In a closed economy the two coincide, but in an open economy, for our purposes, the latter is more relevant and convenient.

The log linearized system can then be written as a set of differential equations

$$\dot{\pi}_{H,t} = \rho\pi_{H,t} - \kappa(c_t + (1-\xi)g_t) - \lambda\hat{\sigma}\alpha(\omega-1)c_t^* - (1-\mathcal{G})\lambda\hat{\sigma}\alpha\omega\theta, \tag{5}$$

$$\dot{c}_t = \hat{\sigma}^{-1}(i_t^* - \pi_{H,t} - \rho) - \alpha(\omega - 1)\dot{c}_t^*, \tag{6}$$

with an initial condition and the definition of the variable θ,

$$c_0 = (1 - \mathcal{G})(1 - \alpha)\theta + c_0^*, \tag{7}$$

$$\theta = (1 - \mathcal{G})\int_0^{+\infty} e^{-\rho s}\rho \frac{(\omega - \sigma)}{\omega + (1 - \alpha)(1 - \sigma)} c_s ds + (1 - \mathcal{G}) \frac{1 - \alpha + \alpha\omega}{\omega + (1 - \alpha)(1 - \sigma)} \frac{\rho}{\alpha} \mathrm{nfa}_0, \tag{8}$$

and either

$$\mathrm{nfa}_0 = 0 \tag{9}$$

if markets are incomplete or

$$\theta = 0 \tag{10}$$

if markets are complete, where $\mathrm{nfa}_0 = \frac{NFA_0}{Y}$ is the normalized deviation of the initial net foreign asset position from ($\mathrm{nfa}_0 = 0$ at the symmetric steady state) and $\theta = \log\Theta$ is the wedge in the log-linearized Backus–Smith equation ($\theta = 0$ at the symmetric steady state). In these equations, we have used the following definitions: $\lambda = \rho_\delta(\rho + \rho_\delta)$, $\kappa = \lambda(\hat{\sigma} + \phi)$, $\xi = \frac{\hat{\sigma}}{\hat{\sigma} + \phi}$,

$$\omega = \sigma\gamma + (1 - \alpha)(\sigma\eta - 1),$$
$$\hat{\sigma} = \frac{\sigma}{1 - \alpha + \alpha\omega} \frac{1}{1 - \mathcal{G}}.$$

Eq. (5) is the New Keynesian Philips Curve. Eq. (6) is the Euler equation. Eq. (7) is derived from the requirement that the terms of trade are predetermined at $t = 0$ because prices are sticky and the exchange rate is fixed. Finally Eq. (8) together with either (9) or (10) depending on whether markets are incomplete or complete, represents the country budget constraint. In the Cole–Obstfeld case $\sigma = \eta = \gamma = \Omega = 1$, so that the complete and incomplete markets solutions coincide. Away from the Cole–Obstfeld case, the complete and incomplete markets solutions differ. The incomplete markets solution imposes that the country budget constraint (8) with $\mathrm{nfa}_0 = 0$, while the complete markets solution solves for the endogenous value of nfa_0 that ensures that the country budget constraint (8) holds with $\theta = 0$. This can be interpreted as an insurance payment from the rest of the world.

These equations form a linear differential system with forcing variables $\{g_t, g_t^*, i_t^*\}$. It will prove useful to define the following two numbers ν and $\bar{\nu}$ (the eigenvalues of the system):

$$\nu = \frac{\rho - \sqrt{\rho^2 + 4\kappa\hat{\sigma}^{-1}}}{2} \qquad \bar{\nu} = \frac{\rho + \sqrt{\rho^2 + 4\kappa\hat{\sigma}^{-1}}}{2}.$$

6.1 Domestic Government Spending

We first consider the experiment where the only shock is domestic government spending, so that $i_t^* = \rho$, $g_t^* = y_t^* = c_t^* = 0$. Note that if $g_t = 0$ throughout then $\theta = 0$ and $y_t = c_t = 0$. We shall compute the deviations from this steady state when $g_t \neq 0$.

The assumptions one makes about financial markets can affect the results. We consider, in turn, both the cases of complete markets and incomplete markets.

6.1.1 Complete Markets

We start by studying the case where markets are complete. This assumption is representative of most of the literature, and is often adopted as a benchmark due to its tractability. The key implication is that consumption is insured against spending shocks. In equilibrium, private agents make arrangements with the rest of the world to receive transfers when spending shoots up and, conversely, to make transfers when spending shoots down. As a result, government sending shocks to not affect consumption on impact. Formally, we have $\theta = 0$, so the system becomes

$$\begin{aligned} \dot{\pi}_{H,t} &= \rho \pi_{H,t} - \kappa(c_t + (1-\xi)g_t), \\ \dot{c}_t &= -\hat{\sigma}^{-1} \pi_{H,t}, \end{aligned}$$

with initial condition

$$c_0 = 0.$$

Because the system is linear, we can write

$$\begin{aligned} c_t &= \int_{-t}^{\infty} \alpha_s^{c,t,CM} g_{t+s} ds, \\ \pi_{H,t} &= \int_{-t}^{\infty} \alpha_s^{\pi,t,CM} g_{t+s} ds, \end{aligned}$$

where the superscript *CM* stands for complete markets. Note two important differences with the closed economy case. First, there are both forward- and backward-looking effects from government spending; the lower bound in these integrals is now given by $-t$ instead of 0. At every point in time, consumption is pinned down by the terms of trade which depend on past inflation. Second, the multipliers depend on calendar time t.

It is important to remind the reader that the sequence of coefficients $\{\alpha_s^{c,t,CM}\}$ represents a notion of fiscal multiplier for total private consumption of domestic goods (by domestic and foreigners) and not for domestic output, which is given by

$$y_t = g_t + \int_{-t}^{\infty} \alpha_s^{c,t,CM} g_{t+s} ds.$$

Whereas the natural benchmark for consumption multipliers is 0, that for output multipliers is 1.

Proposition 3 (Open Economy Multipliers, Complete Markets) *Suppose that markets are complete, then the fiscal multipliers are given by*

$$\alpha_s^{c,t,CM} = \begin{cases} -\hat{\sigma}^{-1}\kappa(1-\xi)e^{-\nu s}\dfrac{1-e^{(\nu-\bar{\nu})(t+s)}}{\bar{\nu}-\nu} & s<0, \\ -\hat{\sigma}^{-1}\kappa(1-\xi)e^{-\bar{\nu}s}\dfrac{1-e^{-(\bar{\nu}-\nu)t}}{\bar{\nu}-\nu} & s\geq 0. \end{cases}$$

It follows that

1. *for $t=0$ we have $\alpha_s^{c,t,CM}=0$ for all s;*

2. *for $t>0$ we have $\alpha_s^{c,t,CM}<0$ for all s;*

3. *for $t\to\infty$ we have $\alpha_{s-t}^{c,t,CM}\to 0$ for all s;*

4. *spending at zero and infinity have no impact: $\alpha_{-t}^{c,t,CM}=\lim_{s\to\infty}\alpha_s^{c,t,CM}=0$.*

The right panel of Fig. 2 displays consumption multipliers for a standard calibration. Consumption multipliers are very different in an open economy with a fixed exchange rate. For starters, part (1) says that the initial response of consumption is always zero, simply restating the initial condition above that $c_0=0$. This follows from the fact that the terms of trade are predetermined and complete markets insure consumption.

Part (2) proves that the consumption response at any other date is actually negative. Note that the Euler equation and the initial condition together imply that

$$c_t = -\hat{\sigma}^{-1}\log\frac{P_{H,t}}{P_H}.$$

Government spending increases demand, leading to inflation, a rise in $P_{H,t}$. In other words, it leads to an appreciation in the terms of trade and this loss in competitiveness depresses private demand, from both domestic and foreign consumers. Although we have derived this result in a specific setting, we expect it to be robust. The key ingredients are that consumption depends negatively on the terms of trade and that government spending creates inflation.

It may seem surprising that the output multiplier is necessarily less than one whenever the exchange rate is fixed, because this contrasts sharply with our conclusions in a closed economy with a fixed interest rate. They key here is that a fixed exchange rate implies a fixed interest rate, but the reverse is not true. We expand on this idea in the next section.

Part (3) says that the impact of government spending at any date on private consumption vanishes in the long run. This exact long run neutrality relies on the assumption of complete markets; otherwise, there are potential long-run neoclassical wealth effects from accumulation of foreign assets.

Part (4) says that spending near zero and spending in the very far future have negligible impacts on consumption at any date. Spending near zero affects inflation for a trivial amount of time and thus have has insignificant effects on the level of home prices. Similarly, spending in the far future has vanishing effects on inflation at any date.

Example 5 (AR(1) Spending) Suppose that $g_t = ge^{-\rho_g t}$ and that markets are complete. Then

$$c_t = -ge^{\nu t}\frac{1-e^{-(\nu+\rho_g)t}}{\nu+\rho_g}\frac{\hat{\sigma}^{-1}\kappa(1-\xi)}{\bar{\nu}+\rho_g}.$$

For $g > 0$, this example shows that c_t is always negative. In other words, in the open economy model with complete markets, output always expands less than the increase in government spending. The intuition is simple. Because the terms of trade are predetermined, private spending on home goods is also predetermined so that $c_0 = 0$. Government spending initially leads to inflation because the total (public and private) demand for home goods is increased in the short run. With fixed nominal interest rates, inflation depresses real interest rates, leading to a decreasing path of private consumption of domestic goods, so that c_t becomes negative. The inflationary pressures are greatest at $t = 0$ and they then recede over time as public and private demand decrease. Indeed at some point in time, inflation becomes negative and in the long run, the terms of trade return to their steady state value. At that point, private consumption of domestic goods $\hat{c}_t$ reaches its minimum and starts increasing, returning to 0 in the long run. The crucial role of inflation in generating $c_t < 0$ is most powerfully illustrated in the rigid price case. When prices are entirely rigid, we have $\kappa = 0$ so that $c_t = 0$ throughout.[i]

An interesting observation is that the openness parameter α enters Proposition 3 or Example 5 only through its effect on $\hat{\sigma}$.[j] As a result, in the Cole–Obstfeld case $\sigma = \eta = \gamma = 1$ and the private consumption multipliers $\alpha_s^{c,t,CM}$ are completely independent of openness α. Away from the Cole–Obstfeld case, $\alpha_s^{c,t,CM}$ depends on α, but its dependence can be positive or negative depending on the parameters.[k]

Next, we ask how fiscal multipliers are affected by the degree of price stickiness.

Proposition 4 (Price Stickiness) *The fiscal multipliers $\{\alpha_s^{c,t,CM}\}$ depend on price flexibility as follows:*

1. *when prices are rigid so that $\kappa = 0$, we have $\alpha_s^{c,t,CM} = 0$ for all s and t;*

[i] Note that the above calculation is valid even if $\rho_g < 0$, as long as $\bar{\nu} + \rho_g > 0$. If this condition is violated, then c_t is $-\infty$ for $g > 0$ and $+\infty$ for $g < 0$.

[j] Recall that $\hat{\sigma} = \frac{\sigma}{1+\alpha[(\sigma\gamma-1)+(\sigma\eta-1)-\alpha(\sigma\eta-1)]}\frac{1}{1-\mathcal{G}}$.

[k] For example, when $\sigma\eta > 1$ and $\sigma\gamma > 1$, $\alpha_s^{c,t,CM}$ is increasing in α for $\alpha \in [0, \min\{\frac{(\sigma\gamma-1)+(\sigma\eta-1)}{2(\sigma\eta-1)}, 1\}]$ and decreasing in α for $\alpha \in [\min\{\frac{(\sigma\gamma-1)+(\sigma\eta-1)}{2(\sigma\eta-1)}, 1\}, 1]$.

2. *when prices become perfectly flexible $\kappa \to \infty$, then for all t, the function $s \to \alpha_s^{c,t,CM}$ converges in distributions to $-(1-\xi)$ times a Dirac distribution concentrated at $s=0$, implying that $\int_{-t}^{\infty} \alpha_s^{c,t,CM} g_{t+s} ds = -(1-\xi) g_t$ for all (continuous and bounded) paths of government spending $\{g_t\}$.*

Unlike in the liquidity trap, fiscal multipliers do not explode when prices become more flexible. In a liquidity trap, government spending sets into motion a feedback loop between consumption and inflation: government spending increases inflation, which lower real interest rates, increases private consumption, further increasing inflation, etc. ad infinitum. This feedback loop is nonexistent in a currency union: government spending increases inflation, appreciates the terms of trade, reduces private consumption, reducing the inflationary pressure. Instead, the allocation converges to the flexible price allocation $c_t = -(1-\xi)g_t$ when prices become very flexible. At the flexible price allocation, private consumption is entirely determined by contemporaneous government spending. Hence the function $\alpha_s^{c,t,CM}$ of s converges in distributions to $-(1-\xi)$ times a Dirac function at $s=0$. This implies that fact that for $s=0$, $\lim_{\kappa\to\infty}\alpha_s^{c,t,CM} = -\infty$ and for $s \neq 0$, $\lim_{\kappa\to\infty}\alpha_s^{c,t,CM} = 0$.

One can reinterpret the neoclassical outcome with flexible prices as applying to the case with rigid prices and a flexible exchange rate that is adjusted to replicate the flexible price allocation. The output multiplier is then less than one. The first result says that with rigid prices but fixed exchange rates, output multipliers are equal to one. In this sense, the comparison between fixed with flexible exchange rates confirms the conventional view from the Mundell–Flemming model that fiscal policy is more effective with fixed exchange rates (see, eg, Dornbusch, 1980). This is consistent with the simulation findings in Corsetti et al. (2011).

6.1.2 Incomplete Markets

We now turn our attention to the case where markets are incomplete. Although the complete market assumption is often adopted for tractability, we believe incomplete markets may be a better approximation to reality in most cases of interest.

A shock to spending may create income effects that affect consumption and labor responses. The complete markets solution secures transfers from the rest of the world that effectively cancel these income effects. As a result, the incomplete markets solution is in general different from the complete market case. One exception is the Cole–Obstfeld case, where $\sigma = \eta = \gamma = 1$.

With incomplete markets, the system becomes

$$\begin{aligned}\dot{\pi}_{H,t} &= \rho\pi_{H,t} - \kappa(c_t + (1-\xi)g_t) - (1-\mathcal{G})\lambda\hat{\sigma}\alpha\omega\theta,\\ \dot{c}_t &= -\hat{\sigma}^{-1}\pi_{H,t},\end{aligned}$$

with initial condition

$$c_0 = (1-\mathcal{G})(1-\alpha)\theta,$$
$$\theta = (1-\mathcal{G})\int_0^{+\infty} e^{-\rho s}\rho\frac{(\omega-\sigma)}{\omega+(1-\alpha)(1-\sigma)}c_s ds.$$

We denote the consumption multipliers with a superscript *IM*, which stands for incomplete markets. We denote by $\hat{t}$ the time such that

$$\frac{e^{\nu\hat{t}}}{1-e^{\nu\hat{t}}} = \omega\frac{\hat{\sigma}}{\hat{\sigma}+\phi}\frac{\alpha}{1-\alpha}.$$

We also define

$$\hat{\Sigma} = (1-\mathcal{G})(1-\alpha)\frac{1}{\bar{\nu}} + (1-\mathcal{G})\frac{\hat{\sigma}}{\hat{\sigma}+\phi}\alpha\omega\frac{1}{\rho}\frac{\nu}{\bar{\nu}}.$$

Note that $\bar{\Omega}=0$ in the Cole–Obstfeld case.

Proposition 5 (Open Economy Multipliers, Incomplete Markets) *Suppose that markets are incomplete, then fiscal multipliers are given by*

$$\alpha_s^{c,t,IM} = \alpha_s^{c,t,CM} + \delta_s^{c,t,IM},$$

where $\alpha_s^{c,t,CM}$ *is the complete markets consumption multiplier characterized in Proposition 3 and*

$$\delta_s^{c,t,IM} = \rho\left[\frac{1-\alpha}{\alpha}e^{\nu t} - \lambda\hat{\sigma}\omega\kappa^{-1}(1-e^{\nu t})\right] \times \frac{\alpha\frac{\omega-\sigma}{\omega+(1-\alpha)(1-\sigma)}}{1-\hat{\Sigma}\frac{1}{1-\mathcal{G}}\rho\frac{\omega-\sigma}{\omega+(1-\alpha)(1-\sigma)}}(1-\xi)e^{-\rho(t+s)}(1-e^{\nu(t+s)}).$$

The difference $\delta_s^{c,t,IM}$ *is* 0 *in the Cole–Obstfeld case* $\sigma=\eta=\gamma=1$. *Away from the Cole–Obstfeld case, the sign of* $\delta_s^{c,t,IM}$ *is the same as the sign of* $\left(\frac{\omega}{\sigma}-1\right)(t-\hat{t})$*; moreover,* $\delta_{-t}^{c,t,IM}=0$ *and* $\lim_{s\to\infty}\delta_s^{c,t,IM}=0$.

The difference between the complete and incomplete market solution vanishes in the Cole–Obstfeld case. Although, away from the this case $\delta_s^{c,t,IM}$ is generally nonzero, it necessarily changes signs (both as a function of *s* for a given *t*, and as a function of *t*, for a given *s*). In this sense, incomplete markets cannot robustly overturn the conclusion of Proposition 3 and guarantee positive multipliers for consumption.

With complete markets

$$\theta = 0,$$

while with incomplete markets

$$\theta = \int_0^{+\infty} e^{-\rho s}(1-\mathcal{G})\rho\frac{\omega-\sigma}{\omega+(1-\alpha)(1-\sigma)}c_t ds.$$

This means that with complete markets, home receives an endogenous transfer nfa_0 from the rest of the world following a government spending shock. In the Cole–Obstfeld case, this transfer is zero, but away from this case, this transfer is nonzero. The difference between these two solutions can then be obtained as the effect of this endogenous transfer.

6.2 Understanding Closed vs Open Economy Multipliers

Fig. 2 provides a sharp illustration of the difference between a liquidity trap and a currency union. In a liquidity trap, consumption multipliers are positive, increase with the date of spending, and become arbitrarily large for long-dated spending. By contrast, in a currency union, consumption multipliers are negative, V-shaped and bounded as a function of the date of spending, and asymptote to zero for long-dated spending.

Before continuing it is useful to pause to develop a deeper understanding of the key difference between the closed and open economy results. The two models are somewhat different—the open economy features trade in goods and the closed economy does not—yet they are quite comparable. Indeed, we will highlight that the crucial difference lies in monetary policy, not model primitives. Although a fixed exchange rate implies a fixed nominal interest rate, the converse is not true.

To make the closed and open economies more comparable, we consider the limit of the latter as $\alpha \to 0$. This limit represents a closed economy in the sense that preferences display an extreme home bias and trade is zero. To simplify, we focus on the case of complete markets so that $\theta = 0$. Even in this limit case, the closed and open economy multipliers differ. This might seems surprising since, after all, both experiments consider the effects of government spending for a fixed nominal interest rate. To understand the difference, we allow for an initial devaluation.

Consider then the open economy model in the closed-economy limit $\alpha \to 0$ and let e_0 denote the new value for the exchange rate after the shock in log deviations relative to its steady-state value (so that $e_0 = 0$ represents no devaluation). The only difference introduced in the system by such one-time devaluation is a change the initial condition to[1]

$$c_0 = \hat{\sigma}^{-1} e_0.$$

[1] The full system allowing for a flexible exchange rate and an independent monetary policy i_t is (with $\theta = 0$ and $c_t^* = 0$)

$$\begin{aligned} \dot{\pi}_{H,t} &= \rho \pi_{H,t} - \kappa(c_t + (1-\xi) g_t), \\ \dot{c}_t &= \hat{\sigma}^{-1}(i_t - \pi_{H,t} - \rho), \\ \dot{e}_t &= i_t - i_t^*, \end{aligned}$$

with initial condition

$$c_0 = \hat{\sigma}^{-1} e_0.$$

If we set $i_t = i_t^*$ then $\dot{e}_t = 0$ so that $e_t = e_0$, which amounts to a one-time devaluation.

The exchange rate devaluation e_0 depreciates the initial terms of trade one for one and increases the demand for home goods through an expenditure switching effect. Of course, this stimulative effect is present in the short run, but vanishes in the long run once prices have adjusted. A similar intuition for the effect of fiscal policy on the exchange rate in a liquidity trap is also discussed in Cook and Devereux (2011).

Now if in the closed economy limit of the open economy model, we set the devaluation e_0 so that $\hat{\sigma}^{-1}e_0$ exactly equals the initial consumption response $\int_0^\infty \alpha_s^c g_{t+s} ds$ of the closed economy model, ie,

$$e_0 = \int_0^\infty \kappa(1-\xi)e^{-\bar{\nu}s}\left(\frac{e^{(\bar{\nu}-\nu)s}-1}{\bar{\nu}-\nu}\right)g_s ds, \tag{11}$$

then we find exactly the same response for consumption and inflation as in the closed economy model. This means that if we combined the government spending shock with an initial devaluation given by (11), then the multipliers of the closed economy limit of the open economy model would coincide with those of the closed economy model.[m]

This analysis shows that the policy analysis conducted for our closed economy model implicitly combines a shock to government spending with a devaluation.[n] In contrast, our open economy analysis assumes fixed exchange rates, ruling out such devaluations. The positive response of consumption in the closed economy model relies entirely on this one-time devaluation. Thus, the key difference between the two models is in monetary policy, not whether the economy is modeled as open or closed. Indeed, we have taken the closed-economy limit $\alpha \to 0$, but the results hold more generally: the degree of openness α matters only indirectly through its impact on $\hat{\sigma}$, ν and $\bar{\nu}$ and in the Cole–Obstfeld case, α actually does not even affect these parameters.

7. LIQUIDITY CONSTRAINTS AND NON-RICARDIAN EFFECTS

In this section, we explore non-Ricardian effects of fiscal policy in a closed and open economy setting. To do so, we follow Campbell and Mankiw (1989), Mankiw (2000), and Gali et al. (2007) and introduce hand-to-mouth consumers, a tractable

[m] Note that the size of this devaluation is endogenous and grows without bound as prices become more flexible, ie, as κ increases. This explains why large multipliers are possible with high values of κ in the closed economy model: they are associated with large devaluations.

[n] To see what this implies, suppose the spending shock has a finite life so that $g_t = 0$ for $t \geq T$ for some T and that monetary policy targets inflation for $t \geq T$. In the closed economy model, inflation is always positive and the price level does not return to its previous level. In contrast, in the open economy model with a fixed exchange rate (no devaluation) inflation is initially positive but eventually negative and the price level returns to its initial steady state value. Indeed, if $g_t > 0$ for $t < T$ and $g_t = 0$ for $t \geq T$ for some T, then inflation is strictly negative for $t \geq T$ and the price level falls towards its long run value asymptotically.

way of modeling liquidity constraints. The latter paper studied the effects of government spending under a Taylor rule in a closed economy. Instead, our focus here is on liquidity traps and currency unions.

7.1 Hand-to-Mouth in a Liquidity Trap

The model is modified as follows. A fraction $1-\chi$ of agents are optimizers, and a fraction χ are hand-to-mouth. Optimizers are exactly as before. Hand-to-mouth agents cannot save or borrow, and instead simply consume their labor income in every period, net of lump-sum taxes. These lump-sum taxes are allowed to differ between optimizers (T_t^o) and hand-to-mouth agents (T_t^r). We define

$$t_t^o = \frac{T_t^o - T^o}{Y} \qquad t_t^r = \frac{T_t^r - T^r}{Y},$$

where T^o and T^r are the per-capita steady state values of T_t^o and T_t^r.

We log-linearize around a steady state where optimizers and hand-to-mouth consumers have the same consumption and supply the same labor. In the appendix, we show that the model can be summarized by the following two equations

$$\begin{aligned}\dot{c}_t &= \tilde{\sigma}^{-1}(i_t - \bar{r}_t - \pi_t) + \tilde{\Theta}_n \dot{g}_t - \tilde{\Theta}_\tau \dot{t}_t^r,\\ \dot{\pi}_t &= \rho\pi_t - \kappa[c_t + (1-\xi)g_t],\end{aligned}$$

where $\tilde{\sigma}$, $\tilde{\Theta}_n$ and $\tilde{\Theta}_\tau$ are positive constants defined in the appendix, which are increasing in χ and satisfy $\tilde{\Theta}_n = \tilde{\Theta}_\tau = 0$ and $\tilde{\sigma} = \hat{\sigma}$ when $\chi = 0$. The presence of hand-to-mouth consumers introduces two new terms in the Euler equation, one involving government spending and the other one involving taxes—both direct determinants of the consumption of hand-to-mouth agents. These terms drop out without hand-to-mouth consumers, since $\chi = 0$ implies $\tilde{\Theta}_n = \tilde{\Theta}_\tau = 0$ and $\tilde{\sigma} = \hat{\sigma}$.

As before we define

$$\tilde{\nu} = \frac{\rho - \sqrt{\rho^2 + 4\kappa\tilde{\sigma}^{-1}}}{2} \qquad \tilde{\tilde{\nu}} = \frac{\rho + \sqrt{\rho^2 + 4\kappa\tilde{\sigma}^{-1}}}{2}.$$

We write the corresponding multipliers with a HM superscript to denote "hand-to-mouth."

Proposition 6 (Closed Economy Multipliers, Hand-to-Mouth) *With hand-to-mouth consumers, we have*

$$c_t = \tilde{c}_t + \tilde{\Theta}_n g_t - \tilde{\Theta}_\tau t_t^r + \int_0^\infty \alpha_s^{c,HM} g_{t+s}\, ds - \int_0^\infty \gamma_s^{c,HM} t_{t+s}^r\, ds,$$

where

$$\alpha_s^{c,HM} = \left(1 + \frac{\widetilde{\Theta}_n}{1-\xi}\right)\widetilde{\alpha}_s^{c,HM} \qquad \gamma_s^{c,HM} = \frac{\widetilde{\Theta}_\tau}{1-\xi}\widetilde{\alpha}_s^{c,HM}.$$

$$\widetilde{\alpha}_s^{c,HM} = \widetilde{\sigma}^{-1}\kappa(1-\xi)e^{-\widetilde{\bar{\nu}}s}\left(\frac{e^{(\widetilde{\bar{\nu}}-\widetilde{\nu})s}-1}{\widetilde{\bar{\nu}}-\widetilde{\nu}}\right).$$

In these expressions, g_t and t_t^r can be set independently of each other because the government can always raise the necessary taxes on optimizing agents by adjusting t_t^o, so that total taxes $t_t = \chi t_t^r + (1-\chi)t_t^o$ are sufficient to balance the government budget over time

$$0 = \int_0^\infty (t_t - g_t)e^{-\rho t}dt.$$

If there are additional constraints on the tax system, then g_t and t_t^r become linked. For example, imagine that tax changes on optimizing and hand-to-mouth have to be identical so that $t_t^o = t_t^r = t_t$. In this case, taxes on hand-to-mouth agents satisfy

$$0 = \int_0^\infty (t_t^r - g_t)e^{-\rho t}dt.$$

Imagine in addition that the government must run a balanced budget, then we must have $t_t^o = t_t^r = t_t = g_t$. In this case, taxes on hand-to-mouth agents satisfy

$$t_t^r = g_t.$$

The presence of hand-to-mouth consumers affects the closed-form solution by modifying the coefficients on spending and adding new terms. The terms fall under two categories: the terms $\widetilde{\Theta}_n g_t - \widetilde{\Theta}_\tau t_t^r$ capturing the concurrent effects of spending and the integral terms $\int_0^\infty \alpha_s^{c,HM} g_{t+s}ds - \int_0^\infty \gamma_s^{c,HM} t_{t+s}^r ds$ capturing the effects of future government spending and future taxes.

The concurrent terms appear because, with hand-to-mouth consumers, current fiscal policy has a direct and contemporaneous impact on spending. They represent traditional Keynesian effects, which are independent of the degree of price flexibility κ. The integral terms capture the effects of future fiscal policy through inflation. They represent New Keynesian terms, which scale with the degree of price flexibility κ, and disappear when prices are perfectly rigid $\kappa = 0$.

Let us start by discussing the concurrent terms $\widetilde{\Theta}_n g_t - \widetilde{\Theta}_\tau t_t^r$. First, the term $-\widetilde{\Theta}_\tau t_t^r$ captures the fact that a reduction in current taxes on hand-to-mouth consumers increases their total consumption directly by redistributing income towards them, away from either unconstrained consumers, who have a lower marginal propensity to consume, or from future hand-to-mouth consumers. Second, the term $\tilde{\Theta}_n g_t$ captures the fact that higher current government spending increases labor income and hence consumption of hand-to-mouth consumers, who have a higher marginal propensity to consume than optimizers. Even when government spending is balanced so that $g_t = \chi t_t^o + (1-\chi)t_t^r$ and taxes are levied equally on optimizers and hand-to-mouth agents so that $t_t^r = g_t$, the sum of the concurrent terms is not exactly zero because of the different effects of government spending and taxes on real wages.

In this case, since $\tilde{\Theta}_\tau = \tilde{\Theta}_n \frac{\mu}{1+\phi}$, the sum of the concurrent terms $\tilde{\Theta}_n g_t - \tilde{\Theta}_\tau t_t^r = \left(1 - \frac{\mu}{1+\phi}\right)\tilde{\Theta}_n g_t$ is likely to be positive in typical calibrations where steady state markups $\mu - 1$ are small compared to ϕ. This is because with sticky prices and flexible wages, real wages increase following increases in government spending, which reduces profit. With heterogeneous marginal propensities to consume, the incidence of this loss across agents matters for private spending, and hence for multipliers, and as we shall see below, these effects can be very large. We refer the reader to the appendix for a complete characterization of fiscal multipliers when these profit effects are taken out (profit offset).

We now turn to the integral terms $\int_0^\infty \alpha_s^{c,HM} g_{t+s} ds - \int_0^\infty \gamma_s^{c,HM} t_{t+s}^r ds$, lower taxes on hand-to-mouth consumers in the future, or higher government spending in the future, stimulates total future consumption.[o] This increases inflation, reducing the real interest rate which increases the current consumption of optimizing agents. This, in turn, stimulates spending by hand-to-mouth consumers. These indirect effects all work through inflation.

Going back to the example where tax changes on hand-to-mouth agents and optimizers discussed above $t_t^o = t_t^r = t_t$, our formulas reveal that the timing of deficits matters. Front-loading fiscal surpluses reduces multipliers through the New Keynesian effects, but increases multipliers early on (and lowers them eventually) through the Keynesian effects.

It is important to understand how these results depend on fixed interest rates, due, say, to a binding zero lower bound. Away from this bound, monetary policy could be chosen

[o] Note that there are conflicting effects of the fraction of hand-to-mouth consumers χ on $\alpha_s^{c,HM} = \left(1 + \frac{\tilde{\Theta}_n}{1-\xi}\right)\tilde{\alpha}_s^{c,HM}$ with $\tilde{\alpha}_s^{c,HM} = \tilde{\sigma}^{-1}\kappa(1-\xi)e^{-\tilde{\tilde{\nu}}s}\left(\frac{e^{(\tilde{\tilde{\nu}}-\tilde{\nu})s}-1}{\tilde{\tilde{\nu}}-\tilde{\nu}}\right)$. On the one hand, future spending increases future output and hence current inflation more when χ is higher, as captured by the multiplicative term $1 + \frac{\tilde{\Theta}_n}{1-\xi}$ which increases with χ. On the other hand, a given amount of inflation leads to less intertemporal substitution when χ is higher, because hand-to-mouth consumers do not substitute intertemporally, as captured by the term $\tilde{\sigma}^{-1}$ which decreases with χ. Overall, for plausible simulations, we find that the former effect tends to be stronger, and potentially much stronger, than the latter. Similar comments apply to the term $\gamma_s^{c,HM}$, which is always positive for $\chi > 0$ but is zero for $\chi = 0$.

to replicate the flexible price allocation with zero inflation. The required nominal interest rate is impacted by the presence of hand-to-mouth consumer

$$i_t = \tilde{\sigma}\left[(1-\xi) + \tilde{\Theta}_n\right]\dot{g}_t + \tilde{\sigma}\,\tilde{\Theta}_\tau \dot{t}^r_t,$$

but consumption is not

$$c_t = -(1-\xi)g_t.$$

Hence away from the zero bound, we get the neoclassical multiplier, which is determined completely statically and does not depend on the presence of hand-to-mouth consumers.[p] In contrast, whenever monetary policy does not or cannot replicate the flexible price allocation, then hand-to-mouth consumers do make a difference for fiscal multipliers. Gali et al. (2007) consider a Taylor rule which falls short of replicating the flexible price allocation. Here, we have focused on fixed interest rates, motivated by liquidity traps.

7.2 Hand-to-Mouth in a Currency Union

We now turn to the open economy version with hand-to-mouth agents.

7.2.1 Complete Markets

We start with the case of complete markets for optimizers. In the appendix, we show that the system becomes

$$\begin{aligned}\dot{\pi}_{H,t} &= \rho\pi_{H,t} - \tilde{\kappa}(c_t + (1-\tilde{\xi})g_t) - (1-\mathcal{G})\lambda\,\tilde{\sigma}\tilde{\alpha}\tilde{\omega}\theta - \tilde{\kappa}\tilde{\tilde{\Theta}}_\tau t^r_t,\\ \dot{c}_t &= -\tilde{\sigma}^{-1}\pi_{H,t} + \tilde{\Theta}_n\dot{g}_t - \tilde{\Theta}_\tau \dot{t}^r_t,\end{aligned}$$

with initial condition

$$c_0 = \tilde{\Theta}_n g_0 - \tilde{\Theta}_\tau t^r_0,$$

for some constants $\tilde{\kappa}$, $\tilde{\alpha}$, $\tilde{\omega}$,$\tilde{\sigma}$, $\tilde{\Theta}_n$, $\tilde{\Theta}_\tau$ and $\tilde{\tilde{\Theta}}_\tau$ defined in the appendix. Importantly $\tilde{\sigma}$, $\tilde{\Theta}_n$, $\tilde{\Theta}_\tau$ are increasing in χ and $\tilde{\Theta}_n$ and $\tilde{\Theta}_\tau$ are decreasing in α. When $\chi = 0$ we have $\tilde{\kappa} = \kappa$, $\tilde{\alpha} = \alpha$, $\tilde{\omega} = \omega$,$\tilde{\sigma} = \hat{\sigma}$, $\tilde{\Theta}_n = 0$, $\tilde{\Theta}_\tau = 0$ and $\tilde{\tilde{\Theta}}_\tau = 0$. As usual, we define

$$\tilde{\nu} = \frac{\rho - \sqrt{\rho^2 + 4\tilde{\kappa}\,\tilde{\sigma}^{-1}}}{2} \qquad \tilde{\bar{\nu}} = \frac{\rho + \sqrt{\rho^2 + 4\tilde{\kappa}\,\tilde{\sigma}^{-1}}}{2}.$$

[p] Note, however, that hand-to-mouth agents might change the associated allocation of optimizers. They just don't matter for the aggregate allocation.

Proposition 7 (Open Economy Multipliers, Hand-to-Mouth, Complete Markets) *With hand-to-mouth agents and complete markets for optimizers, we have*

$$c_t = \tilde{\Theta}_n g_t - \tilde{\Theta}_\tau t_t^r + \int_{-t}^{\infty} \alpha_s^{c,t,HM,CM} g_{t+s} ds - \int_{-t}^{\infty} \gamma_s^{c,t,HM,CM} t_{t+s}^r ds,$$

where

$$\alpha_s^{c,t,HM,CM} = \left(1 + \frac{\tilde{\Theta}_n}{1 - \tilde{\xi}}\right) \tilde{\alpha}_s^{c,t,HM,CM}, \qquad \gamma_s^{c,t,HM,CM} = \frac{\tilde{\Theta}_\tau - \tilde{\tilde{\Theta}}_\tau}{1 - \tilde{\xi}} \tilde{\alpha}_s^{c,t,HM,CM},$$

$$\tilde{\alpha}_s^{c,t,HM,CM} = \begin{cases} -\tilde{\sigma}^{-1} \tilde{\kappa}(1 - \tilde{\xi}) e^{-\tilde{\nu} s} \dfrac{1 - e^{(\tilde{\nu} - \tilde{\bar{\nu}})(t+s)}}{\tilde{\bar{\nu}} - \tilde{\nu}} & s < 0, \\[2ex] -\tilde{\sigma}^{-1} \tilde{\kappa}(1 - \tilde{\xi}) e^{-\tilde{\bar{\nu}} s} \dfrac{1 - e^{-(\tilde{\bar{\nu}} - \tilde{\nu})t}}{\tilde{\bar{\nu}} - \tilde{\nu}} & s \geq 0. \end{cases}$$

Just as in the closed economy case, hand-to-mouth consumers introduce additional Keynesian effects and New Keynesian effects through cumulated inflation, where the former are independent of price flexibility κ while the latter scale with price flexibility κ and disappear when prices are perfectly rigid so that $\kappa = 0$. Just as in the closed economy case, the Keynesian effects increase consumption in response to contemporaneous positive government spending shocks and decrease consumption in response to contemporaneous increases in taxes on hand-to-mouth agents. The difference with the closed economy case is that the New Keynesian effects tend to depress consumption in response to positive government spending shocks. A pure illustration of the Keynesian effect is initial consumption c_0 (for which New Keynesian effects are 0), which is not 0 anymore, but instead $c_0 = \tilde{\Theta}_n g_0 - \tilde{\Theta}_\tau t_0^r$. Importantly $\tilde{\Theta}_n$ and $\tilde{\Theta}_\tau$ are decreasing with the degree of openness α, simply because higher values of α reduce the marginal propensity to consume on domestic goods of hand-to-mouth agents, capturing the "leakage abroad" of fiscal policy.

7.2.2 Incomplete Markets

We now treat the case of incomplete markets for optimizers. We refer the reader to the appendix for the definitions of the constants $\tilde{\Omega}_n$, $\tilde{\Omega}_c$, Σ.

Proposition 8 (Open Economy Multipliers, Hand-to-Mouth, Incomplete Markets) *With hand-to-mouth agents and incomplete markets for optimizers, we have*

$$c_t = \tilde{\Theta}_n g_t - \tilde{\Theta}_\tau t_t^r + \int_{-t}^{\infty} \alpha_s^{c,t,HM,IM} g_{t+s} ds - \int_{-t}^{\infty} \gamma_s^{c,t,HM,IM} t_{t+s}^r ds,$$

where

$$\alpha_s^{c,t,HM,IM} = \alpha_s^{c,t,HM,CM} + \delta_s^{c,t,HM,IM},$$
$$\gamma_s^{c,t,HM,IM} = \gamma_s^{c,t,HM,CM} + \epsilon_s^{c,t,HM,IM},$$

with

$$\begin{aligned}\delta_s^{c,t,HM,IM} = \rho\left[\frac{1-\widetilde{\alpha}}{\widetilde{\alpha}}e^{\widetilde{\nu}t} - (1-\mathcal{G})\lambda\widetilde{\sigma}\widetilde{\kappa}^{-1}\widetilde{\omega}\left(1-e^{\tilde{\nu}t}\right)\right] \\ \times\frac{\widetilde{\alpha}}{1-\Sigma\widetilde{\Omega}_c}\left[e^{-\rho(t+s)}\frac{(1-\mathcal{G})\widetilde{\Omega}_n}{\rho} + e^{-\rho(t+s)}\frac{(1-\mathcal{G})\widetilde{\Omega}_c}{\rho}\widetilde{\Theta}_n\right. \\ \left.+\frac{(1-\mathcal{G})\widetilde{\Omega}_c}{\rho}\left(1-\widetilde{\xi}\right)\left(1+\frac{\widetilde{\Theta}_n}{1-\widetilde{\xi}}\right)e^{-\rho(t+s)}\left(1-e^{\tilde{\nu}(t+s)}\right)\right],\end{aligned}$$

$$\begin{aligned}\epsilon_s^{c,t,HM,IM} = -\rho\left[\frac{1-\widetilde{\alpha}}{\widetilde{\alpha}}e^{\widetilde{\nu}t} - \lambda\widetilde{\sigma}\widetilde{\kappa}^{-1}\widetilde{\omega}\left(1-e^{\tilde{\nu}t}\right)\right] \\ \times\frac{\widetilde{\alpha}}{1-\Sigma\widetilde{\Omega}_c}\left[e^{-\rho(t+s)}\frac{(1-\mathcal{G})\widetilde{\Omega}_\tau}{\rho} - e^{-\rho(t+s)}\frac{(1-\mathcal{G})\widetilde{\Omega}_c}{\rho}\widetilde{\Theta}_\tau\right. \\ \left.+\frac{(1-\mathcal{G})\widetilde{\Omega}_c}{\rho}\left(1-\widetilde{\xi}\right)\frac{\widetilde{\widetilde{\Theta}}_\tau-\widetilde{\Theta}_\tau}{1-\widetilde{\xi}}e^{-\rho(t+s)}\left(1-e^{\tilde{\nu}(t+s)}\right)\right].\end{aligned}$$

The difference between the complete and incomplete market solution $\delta_s^{c,t,HM,IM}$ and $\epsilon_s^{c,t,HM,IM}$ are generally nonzero, can be understood along the same lines as in Section 6 in the absence of hand-to-mouth agents, generally switch signs with t and s, but do not substantively overturn the forces identified in the case of complete markets.

8. OUTSIDE-FINANCED FISCAL MULTIPLIERS

Up to this point, in our open economy analysis of currency unions, we have assumed that each country pays for its own government spending. Actually, with complete markets it does not matter who is described as paying for the government spending, since regions will insure against this expense. In effect, any transfers across regions arranged by governments are undone by the market. With incomplete markets, however, who pays matters. Transfers between regions cannot be undone and affect the equilibrium. Thus, for the rest of this section we assume incomplete markets.

We first examine what happens when the domestic country doesn't pay for the increase in domestic government spending. We show that this can make an important difference and lead to larger multipliers. This is likely to be important in practice: indeed, a large part of the "local multiplier" literature considers experiments where government spending is not paid by the economic region under consideration.

8.1 Outside-Financed Fiscal Multipliers with No Hand-to-Mouth

We first start with the case where there are no hand-to-mouth agents. The only difference with the results with incomplete markets from Section 6.1 is that we now have

$$\theta = (1-\mathcal{G})\int_0^{+\infty} e^{-\rho s}\rho\frac{(\omega-\sigma)}{\omega+(1-\alpha)(1-\sigma)}c_s ds + (1-\mathcal{G})\frac{1-\alpha+\alpha\omega}{\omega+(1-\alpha)(1-\sigma)}\frac{\rho}{\alpha}\mathrm{nfa}_0,$$

where

$$\mathrm{nfa}_0 = \int_0^{\infty} e^{-\rho t} g_t dt$$

is the transfer from foreign to home that pays for the increase in government spending. In the Cole–Obstfeld case $\sigma = \eta = \gamma = \Omega = 1$.

We denote the consumption multipliers with a superscript *PF*, which stands for "paid for" by foreigners.

Proposition 9 (Outside-Financed Open Economy Multipliers) *When domestic government spending is outside-financed, the fiscal multipliers are given by the same expressions as in Proposition 5 with the difference that*

$$\alpha_s^{c,t,PF} = \alpha_s^{c,t,IM} + \delta_s^{c,t,PF},$$

where $\alpha_s^{c,t,IM}$ is the incomplete markets consumption multiplier characterized in Proposition 5 and

$$\delta_s^{c,t,PF} = \rho\left[\frac{1-\alpha}{\alpha}e^{\nu t} - \lambda\hat{\sigma}\omega\kappa^{-1}(1-e^{\nu t})\right] \times \frac{1}{1-\hat{\Sigma}\frac{1}{1-\mathcal{G}}\rho\frac{\omega-\sigma}{\omega+(1-\alpha)(1-\sigma)}}\frac{1-\alpha+\alpha\omega}{\omega+(1-\alpha)(1-\sigma)}e^{-\rho(t+s)}.$$

The sign of $\delta_s^{c,t,PF}$ is the same as that of $(\hat{t}-t)$ and $\lim_{s\to\infty}\delta_s^{c,t,PF} = 0$.

In the Cole–Obstfeld case $\sigma = \eta = \gamma = 1$, the expression simplifies to

$$\delta_s^{c,t,PF} = \left[e^{\nu t}\frac{1-\alpha}{\alpha} - (1-e^{\nu t})\frac{1}{1-\mathcal{G}}\frac{1}{\frac{1}{1-\mathcal{G}}+\phi}\right]\rho e^{-\rho(t+s)}.$$

The intuition is most easily grasped by considering the Cole–Obstfeld case, which we focus on for now. When government spending is outside-financed, there is an associated transfer to domestic agents. Because agents are permanent-income consumers, only the net present value of the per-period transfer matters, which in turn depends on the persistence of the shock to government spending. The effects of this transfer is captured by the term $\delta_s^{c,t,PF}$, which is higher, the higher the degree of home bias (the lower α). Indeed, more generally, we can compute net-present-value transfer multipliers for pure transfers nfa_0 unrelated to government spending[q]:

$$c_t = \beta^{c,t}\text{nfa}_0$$

with

$$\beta^{c,t} = \left[e^{\nu t}\frac{1-\alpha}{\alpha} - \left(1-e^{\nu t}\right)\frac{1}{1-\mathcal{G}}\frac{1}{\frac{1}{1-\mathcal{G}}+\phi}\right]\rho.$$

We can also compute the effects of net-present-value transfers on inflation $\beta^{\pi,t} = -\nu e^{\nu t}\left[\rho\frac{1-\alpha}{\alpha} + \rho\frac{1}{\frac{1}{1-\mathcal{G}}+\phi}\right]$ and on the terms of trade $\beta^{s,t} = -\left[1-e^{\nu t}\right]\left[\rho\frac{1-\alpha}{\alpha} + \rho\frac{1}{\frac{1}{1-\mathcal{G}}+\phi}\right]$ (note that the terms of trade gap equals accumulated inflation $s_t = -\int_0^t \pi_{H,s}ds$). The presence of the discount factor ρ in all these expressions is natural because what matters is the annuity value ρnfa_0 of the transfer.

Net-present-value transfers have opposite effects on output in the short and long run. In the short run, when prices are rigid, there is a Keynesian effect due to the fact that transfers stimulate the demand for home goods: $\beta^{c,0} = \rho\frac{1-\alpha}{\alpha}$. In the long run, when prices adjust, the neoclassical wealth effect on labor supply lowers output: $\lim_{t\to\infty}\beta^{c,t} = -\rho\frac{1}{\frac{1}{1-\mathcal{G}}+\phi}$. In the medium run, the speed of adjustment, from the Keynesian short-run response to the neoclassical long-run response, is controlled by the degree of price flexibility κ, which affects ν.[r]

Note that the determinants of the Keynesian and neoclassical wealth effects are very different. The strength of the Keynesian effect hinges on the relative expenditure share of home goods $\frac{1-\alpha}{\alpha}$: the more closed the economy, the larger the Keynesian effect. The strength of the neoclassical wealth effect depends on the elasticity of labor supply $\frac{1}{\phi}$: the more elastic labor supply, the larger the neoclassical wealth effect.

Positive net-present-value transfers also increase home inflation. The long-run cumulated response in the price of home produced goods equals $\rho\frac{1-\alpha}{\alpha} + \rho\frac{1}{\frac{1}{1-\mathcal{G}}+\phi}$.

[q] In the particular case that we study here, transfers occur concurrently with an increase in government spending and exactly pay for the increase in government spending $\text{nfa}_0 = \int_0^\infty e^{-\rho t}g_t dt$.

[r] Note that ν is decreasing in κ, with $\nu = 0$ when prices are rigid ($\kappa = 0$), and $\nu = -\infty$ when prices are flexible ($\kappa = \infty$).

The first term $\rho\frac{1-\alpha}{\alpha}$ comes from the fact that transfers increase the demand for home goods, due to home bias. The second term $\rho\frac{1}{\frac{1}{1-\mathcal{G}}+\phi}$ is due to a neoclassical wealth effect that reduces labor supply, raising the wage. How fast this increase in the price of home goods occurs depends positively on the flexibility of prices through its effect on ν.[s]

These effects echo the celebrated Transfer Problem controversy of Keynes (1929) and Ohlin (1929). With home bias, a transfer generates a boom when prices are sticky, and a real appreciation of the terms of trade when prices are flexible. The neoclassical wealth effect associated with a transfer comes into play when prices are flexible, and generates an output contraction and a further real appreciation.

In the closed economy limit we have $\lim_{\alpha\to 0}\beta^{c,t}=\infty$. In the fully open economy limit we have $\lim_{\alpha\to 0}\beta^{c,t}=0$. The intuition is that the Keynesian effect of transfers is commensurate with the relative expenditure share on home goods $\frac{1-\alpha}{\alpha}$. This proposition underscores that transfers are much more stimulative than government spending, the more so, the more closed the economy. This robust negative dependence of transfer multipliers $\beta^{c,t}$ on openness α should be contrasted with the lack of clear dependence on openness of government spending multipliers $\alpha_s^{c,t,CM}$ noted above (indeed in the Cole–Obstfeld case, $\alpha_s^{c,t,CM}$ is independent of α).

Example 6 (Outside-Financed Spending, Cole–Obstfeld, AR(1)) Suppose that $g_t=ge^{-\rho_g t}$ and that domestic government spending is outside-financed. In the Cole–Obstfeld case $\sigma=\eta=\gamma=1$, we have

$$c_t=g\left[e^{\nu t}\frac{1-\alpha}{\alpha}-(1-e^{\nu t})\frac{1}{1-\mathcal{G}}\frac{1}{\frac{1}{1-\mathcal{G}}+\phi}\right]\frac{\rho}{\rho+\rho_g}$$
$$-ge^{\nu t}\left(\frac{1-e^{-(\nu+\rho_g)t}}{\nu+\rho_g}\right)\kappa(1-\xi)\frac{1-\mathcal{G}}{\bar{\nu}+\rho_g}.$$

Moreover we have $c_0=g\frac{1-\alpha}{\alpha}\frac{\rho}{\rho+\rho_g}$ and $\lim_{t\to\infty}c_t=-g\frac{1}{1-\mathcal{G}}\frac{1}{\frac{1}{1-\mathcal{G}}+\phi}\frac{\rho}{\rho+\rho_g}$.

Note that the second term on the right-hand side of the expression for c_t in Example 6 is simply the term identified in Example 5 in the complete markets case. The first term arises precisely because government spending is now paid for by foreign.

It is particularly useful to look at the predictions of this proposition for $t=0$ and $t\to\infty$. In the case of a stimulus $g>0$, we have $c_0>0>\lim_{t\to\infty}c_t$. Following a positive stimulus shock, we can get $c_0>0$ and actually $c_t>0$ for some time (because $\theta>0$) and eventually $c_t<0$. The conclusion would be that an unpaid for fiscal stimulus at

[s] Recall that ν is decreasing in the degree of price flexibility κ.

home has a larger consumption multiplier in the short run and smaller in the long run. This is true as long as there is home bias $\alpha < 1$. The reason is that the associated transfer redistributes wealth from foreign to home consumers. This increases the demand for home goods because of home bias. In the neoclassical model with flexible prices, there would be an appreciation of the terms of trade and a reduction in the output of home goods because of a neoclassical wealth effect. With sticky prices, prices cannot adjust in the short term, and so this appreciation cannot take place right away, and so the output of home goods increases. In the long run, prices adjust and we get the neoclassical effect.

The lesson of this section is that we can partly overturn the conclusion of Proposition 3 when government spending is outside-financed. When the degree of home bias $1-\alpha$, is high, or when increases in government spending are very persistent, then local multipliers estimates that involve increases in government spending that are not self-financed are potentially substantially inflated compared to the counterfactual of self-financed increases in government spending.

8.2 Outside-Financed Fiscal Multipliers with Hand-to-Mouth

We now turn to the case where there are hand-to-mouth agents.

Proposition 10 (Outside-Financed Open Economy Multipliers, Incomplete Markets, Hand-to-Mouth) *With hand-to-mouth agents, when domestic government spending is outside-financed, the fiscal multipliers are given by the same expressions as in Proposition 8 with the difference that*

$$\alpha_s^{c,t,HM,PF} = \alpha_s^{c,t,HM,IM} + \delta_s^{c,t,PF},$$

where

$$\delta_s^{c,t,HM,PF} = \rho\left[\frac{1-\widetilde{\alpha}}{\widetilde{\alpha}} e^{\widetilde{\nu} t} - \lambda\widetilde{\sigma}\widetilde{\kappa}^{-1}\widetilde{\omega}\left(1-e^{\widetilde{\nu} t}\right)\right]\frac{1}{1-\Sigma\widetilde{\Omega}_c}\frac{\widetilde{\alpha}(1-\mathcal{G})\widetilde{\Omega}_f}{\rho}e^{-\rho(t+s)}.$$

When domestic government spending is outside-financed, the question of the incidence of the accompanying transfer across domestic optimizers and hand-to-mouth agents naturally arises. These distributive effects are entirely captured by the adjustment in the taxes t_t^r paid by hand-to-mouth agents.

From now on, we focus on the benchmark case where taxes and the accompanying per-period transfer are distributed equally on optimizers and hand-to-mouth agents and where the domestic government runs a balanced budget, because this case is the most relevant to think about most of the estimates in the local multipliers literature where regions correspond to states with limited de jure or de facto ability to borrow.

When domestic government spending is self-financed, we have $t_t^o = t_t^r = g_t$, and instead when government spending is outside-financed, we have $t_t^o = t_t^r = 0$. Comparing fiscal multipliers when government spending is self-financed vs outside-financed, the effect of reduced taxes on optimizers in the latter case is captured by the corrective term

$\delta_s^{c,t,PF}$, while the effect of reduced taxes on hand-to-mouth agents is captured by the reduction in t_t^r from g_t to zero. In particular, in the short run before prices can fully adjust, both effects increase fiscal multipliers, the first effect for reasons already discussed in the case without hand-to-mouth agents in Section 8.1, the second effect because hand-to-mouth agents have a higher marginal propensity to consume than optimizers.

The presence of hand-to-mouth agents magnifies the difference between self-financed and outside-financed fiscal multipliers for temporary government spending shocks, simply because hand-to-mouth agents spend more of the temporary implicit transfer from foreigners that separate these two experiments in the short run, the more so, the more temporary the government spending shock.

Overall, this analysis shows that when the average marginal propensity to consume on domestic goods, as captured by the fraction of hand-to-mouth agents χ and by the degree of home bias $1-\alpha$, is high, or when increases in government spending are very persistent, then local multipliers estimates that involve increases in government spending that are not self-financed are potentially substantially inflated compared to the counterfactual of self-financed increases in government spending.

9. TAKING STOCK: SOME SUMMARY MULTIPLIER NUMBERS

In this section, we provide numerical illustrations for the forces that we have identified in the chapter. We report summary multipliers $M^Y = 1 + M^C$ in liquidity traps and currency unions, computed as the ratio of the average response of output over the 2 years following the increase in spending to the average increase in government spending over the same period. Our baseline calibration features $\chi = 0$, $\sigma = 1$, $\epsilon = 6$, $\phi = 3$, and $\mathcal{G} = 0.3$ for liquidity traps and $\chi = 0$, $\sigma = 1$, $\eta = \gamma = 1$, $\epsilon = 6$, $\phi = 3$, $\mathcal{G} = 0.3$, and $\alpha = 0.4$ for currency unions. We take the government spending shock to be constant for $\tau_g = 1.25$ years (5 quarters) and zero afterwards.[t] We then explore variations with higher values of χ. In all these experiments, we maintain the assumption that taxes fall equally on hand-to-mouth agents and on optimizers, and that markets are incomplete. In the deficit financed experiments, taxes are increased (discretely) only after three years, and are then constant for 1.25 years before reverting to zero. The first part of Table 1 corresponds to the case of perfectly rigid prices $\lambda = 0$ (infinite price duration), the second part to $\lambda = 0.12$ (price duration of 2.9 years), and $\lambda = 1.37$ (price duration of 0.9 year).

We start with the case of perfectly rigid prices in the first part of Table 1. This table presents summary multipliers in liquidity traps and currency unions, depending on whether

[t] This shock has the same duration $\frac{\tau_g}{2} = \frac{1}{\rho_g}$ as an AR(1) with a coefficient with $\rho_g = 1.6$ (corresponding to a quarterly mean-reversion coefficient of 0.7), but dies off completely in finite time (after 1.6 years), leading to more reasonable values for liquidity trap multipliers when prices are somewhat flexible (the tail of the shock matters a great deal in this case because α_s^c and $\widetilde{\alpha}_s^{c,HM}$ increase exponentially with the horizon s).

Table 1 Summary output multipliers

	Liquidity trap						Currency union								
	Tax-financed			Deficit-financed			Tax-financed			Deficit-financed			Foreign-financed		
Rigid prices ($\lambda=0$)	$o=0$	$o=0.5$	$o=1$	$o=0$	$o=0.5$	$o=1$	$o=0$	$o=0.5$	$o=1$	$o=0$	$o=0.5$	$o=1$	$o=0$	$o=0.5$	$o=1$
$\chi=0$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.1160	1.1160	1.1160
$\chi=0.25$	4.5000	1.4804	1.0000	6.0000	1.8922	1.2386	1.6459	1.1956	1.0000	1.9474	1.3786	1.1314	2.0387	1.4823	1.2446
$\chi=0.5$	*	*	1.0000	*	*	1.7159	*	2.2835	1.0000	*	3.4861	1.3361	*	3.5041	1.4514
$\chi=0.75$	*	*	1.0000	*	*	3.1477	*	*	1.0000	*	*	1.7385	*	*	2.4971
Sticky prices ($\lambda=0.12$)	$o=0$	$o=0.5$	$o=1$	$o=0$	$o=0.5$	$o=1$	$o=0$	$o=0.5$	$o=1$	$o=0$	$o=0.5$	$o=1$	$o=0$	$o=0.5$	$o=1$
$\chi=0$	1.0542	1.0542	1.0542	1.0542	1.0542	1.0542	0.8968	0.8968	0.8968	0.8968	0.8968	0.8968	0.9550	0.9550	0.9550
$\chi=0.25$	6.9420	1.6437	1.0542	−191.4702	−1.0069	0.5347	1.2856	1.0321	0.8984	1.5476	1.2020	1.0241	1.5819	1.2410	1.0611
$\chi=0.5$	*	*	1.0542	−*	−*	−0.5044	*	1.5451	0.9009	*	2.5252	1.2233	*	2.3385	1.2302
$\chi=0.75$	*	*	1.0542	−*	−*	−3.6218	*	*	0.9083	*	*	1.5770	*	*	1.6241
Sticky prices ($\lambda\simeq1.37$)	$o=0$	$o=0.5$	$o=1$	$o=0$	$o=0.5$	$o=1$	$o=0$	$o=0.5$	$o=1$	$o=0$	$o=0.5$	$o=1$	$o=0$	$o=0.5$	$o=1$
$\chi=0$	1.8315	1.8315	1.8315	1.8315	1.8315	1.8315	0.6529	0.6529	0.6529	0.6529	0.6529	0.6529	0.6638	0.6638	0.6638
$\chi=0.25$	168.2368	4.5741	1.8315	−3.4965e8	−5153.3064	−242.9734	0.8142	0.7101	0.6542	0.9127	0.7795	0.7096	0.9266	0.7883	0.7141
$\chi=0.5$	*	*	1.8315	−*	−*	−732.5833	*	0.9238	0.6563	*	1.2767	0.7999	*	1.2515	0.7941
$\chi=0.75$	*	*	1.8315	−*	−*	−2201.4125	*	*	0.6612	*	*	0.9670	*	*	0.9559

or not they are tax-financed (taxes equal to government spending in every period), deficit-financed (taxes are raised only 3 years after the increase in spending, and then mean-revert at the same rate as spending), or outside-financed (no change in taxes). For all these cases, we also report multipliers for different values of the profit-offset coefficient o: 0, 0.5, and 1. This profit-offset coefficient is equal to the share of marginal profits per agent which is transferred to each hand-to-mouth agent: when it is equal to 0, hand-to-mouth agents are completely shielded from the impact of government spending on profits, and when it is equal to 1, they are impacted exactly like optimizers. This is important because with sticky prices and flexible wages, real wages increase following increases in government spending, so that profits increase less than proportionately with output, while labor income increases more than proportionately. With heterogeneous marginal propensities to consume, the incidence of this loss across agents matters for private spending, and hence for multipliers, and as we shall see below, these effects can be very large. While our analysis in the main text of the paper is confined to the case $o = 0$, the appendix gives a full treatment of the arbitrary o case. We also vary the fraction of hand-to-mouth agents χ between 0 and 0.75.

The results are as follows. We start with our baseline calibration. The multiplier is always 1 in a liquidity trap, independently of whether government spending is tax- or debt-financed. In a currency union, the multiplier is 1 independently of whether government spending is tax- or debt-financed, but it increases to 1.1 when it is outside-financed.

We then depart from the baseline increasing the fraction of hand-to-mouth agents χ from 0 to 0.25, 0.5, and 0.75. We start with the case of full profit offset $o = 1$ and explain the role of profit offset later. In a liquidity trap, the tax-financed multiplier remains at 1 irrespective of χ. The deficit-financed multiplier increases with χ to 1.2 ($\chi = 0.25$), 1.7 ($\chi = 0.5$), or 3.1 ($\chi = 0.75$). Turning to currency unions, the tax-financed multiplier is 1 irrespective of χ. The deficit-financed multiplier increases with χ to 1.1 ($\chi = 0.25$), 1.3 ($\chi = 0.5$), or 1.8 ($\chi = 0.75$). Finally the outside-financed multiplier increases with χ to 1.3 ($\chi = 0.25$), 1.5 ($\chi = 0.5$), or 2.7 ($\chi = 0.75$). Importantly, the difference between outside- and self-financed multipliers is now larger than in our baseline, and the deficit-financed multiplier is in between these two multipliers.

In general, lower values of the profit offset coefficient o lead to higher multipliers. This is because with no profit offset, the contemporaneous reduction in profits resulting from the increase in government spending acts like a redistribution from low marginal propensity to consume optimizers toward high marginal propensity to consume hand-to-mouth agents, which increases output (and vice versa for the increase in taxes). This effect, which can be very large, disappears with full profit offset. The * in Table 1 indicates that the feedback loop between output and the distributive effects of profits on agents with different marginal propensities to consume is so powerful that it "blows up". When it occurs, our formulas cease to apply and the correct interpretation is that multipliers are positive infinite.

We continue with the case of sticky but not perfectly rigid prices in the second and third parts of Table 1, where we run through the exact same experiments as in the first part of Table 1. The key differences are as follows. First, in the case of liquidity traps, tax-financed multipliers are a lot higher than with rigid prices, illustrating the power of the positive feedback loop between inflation and output. Deficit-financed multipliers can be a lot lower than with rigid prices and can actually be negative when there are enough hand-to-mouth agents because the positive feedback loop for front-loaded government spending is weaker than the more back-loaded negative one for taxes (in this case, lower profit offset reduces multipliers, potentially leading to negative infinite values indicated by $-*$). Second, in the case of currency unions, multipliers are lower than with rigid prices, but the difference is not as large as in the case of liquidity traps. This is because in this case, there is no feedback loop between output and inflation since inflation lowers spending instead of increasing it, because of its accumulated effect appreciates the terms of trade and rebalances spending away from home goods toward foreign goods.

Although this is not illustrated in the table, we briefly comment on the role of the persistence of shocks and of the openness of the economy. In liquidity traps, more persistent government spending shocks tend to increase tax-financed multipliers because of the feedback loop between output and inflation (in fact tax-financed multipliers can become infinite when prices are not entirely rigid, even without hand-to-mouth agents). They increase deficit-financed multipliers with no hand-to-mouth agents but can decrease them with enough hand-to-mouth agents and somewhat flexible prices because the feedback loop between output and inflation is more potent for back-loaded taxes than for front-loaded government spending. In currency unions, more persistent government spending shocks tend to decrease tax-financed and deficit-financed multipliers, but to increase outside-financed multipliers when prices are rigid enough. In currency unions, multipliers tend to increase when the economy is more closed (α is lower) when government spending is outside-financed and prices are not too flexible or when it is deficit-financed and larger than one (less leakage abroad).

Our simulations are illustrative and do not attempt to explore a wide range of possible parameters. For example, we have kept the fraction of hand-to-mouth agents at a modest level. Likewise, we only explore a relatively open economy. Overall, even within this limited range, our results show that fiscal multipliers are somewhat sensitive to various primitive parameters, as well as the nature of the fiscal experiment. Differences were found comparing completely rigid prices to standard degrees of price stickiness, especially for the liquidity trap case. The presence of hand-to-mouth agents also affects the responses significantly. Perhaps most surprisingly, distributional impacts appear to be crucial. First, there is the difference between tax-financed, deficit-financed, and outside-financed spending. Second, there is the difference in the responses obtained depending on the way profits are redistributed. As explained earlier, this effect relies on the model prediction that profits relative to labor earnings are countercyclical. Thus, this effect could be mitigated if wages, which are flexible in our standard New Keynesian model, were also assumed to be sticky.

Theoretically, in currency unions, outside-financed multipliers can be much larger than deficit-financed multipliers, especially when the economy is relatively closed and government spending shocks are relatively persistent. However, in our simulations with relatively open economies and relatively transitory government spending shocks (which capture the characteristics of many local multiplier studies), these differences are not very large. Since deficit-financed multipliers tend to be larger in liquidity traps than in currency unions (because there is less "leakage" abroad) with rigid enough prices, it would appear that outside-financed multipliers in currency unions (as estimated in the local multipliers literature) may provide a rough lower bound for national multipliers deficit-financed in liquidity traps with rigid enough prices. When prices are more flexible, the comparison is more delicate and the rough lower bound need not apply.

10. COUNTRY SIZE, AGGREGATION, AND FOREIGN GOVERNMENT SPENDING

So far, we have focused on the case where the country undertaking the fiscal stimulus is a small (infinitesimal) part of the currency union—this is implied by our modeling of countries as a continuum. Here, we relax this assumption. To capture country size, we interpret i as indexing regions and we imagine that countries $i \in [0, x]$ are part of a single country. They undertake the same fiscal stimulus g_t^i. We denote with a $- \ i \in (x, 1]$ the index of a typical region that is not undertaking fiscal stimulus so that $g_t^{-i} = 0$. We consider two situations: (1) monetary policy i_t^* at the union level achieves perfect inflation targeting (2) monetary policy at the union level is passive because the union is in a liquidity trap where interest rates i_t^* are at the zero lower bound. For simplicity, we focus on the Cole–Obstfeld case throughout.

10.1 Inflation Targeting at the Union Level

The aggregates variables satisfy

$$g_t^* = \int_0^1 g_t^i di = x g_t^i,$$

$$c_t^* = \int_0^1 c_t^i di = x c_t^i + (1-x) c_t^{-i},$$

$$\pi_t^* = \int_0^1 \pi_t^i di = x \pi_t^i + (1-x)] \pi_t^{-i}.$$

As long as the zero lower bound is not binding, monetary policy at the union level can be set to target zero inflation $\pi_t^* = 0$. The required interest rate i_t^* is

$$i_t^* - \rho = -\hat{\sigma}(1-\xi)x\dot{g}_t^i,$$

and the corresponding value of c_t^* is

$$c_t^* = -(1-\xi)xg_t^i.$$

The allocation for regions in the country undertaking the stimulus solves

$$\begin{aligned}\dot{\pi}_t^i &= \rho\pi_t^i - \kappa(c_t^i + (1-\xi)g_t^i),\\ \dot{c}_t^i &= -(1-\xi)x\dot{g}_t^i - \hat{\sigma}^{-1}\pi_t^i,\end{aligned}$$

$$c_0^i = -(1-\xi)xg_0^i.$$

Similarly the allocation for regions not undertaking the stimulus solves

$$\dot{\pi}_t^{-i} = \rho\pi_t^{-i} - \kappa c_t^{-i},$$

$$\dot{c}_t^{-i} = -(1-\xi)x\dot{g}_t^i - \hat{\sigma}^{-1}\pi_t^{-i},$$

$$c_0^{-i} = -(1-\xi)xg_0^i.$$

In the Cole–Obstfeld case, we define

$$\alpha_s^{c,t,CM*} = \begin{cases} \hat{\sigma}^{-1}\kappa(1-\xi)e^{-\nu s}\dfrac{1-e^{(\nu-\bar{\nu})(t+s)}}{\bar{\nu}-\nu} & s<0,\\ \hat{\sigma}^{-1}\kappa(1-\xi)e^{-\bar{\nu}s}\dfrac{1-e^{(\nu-\bar{\nu})t}}{\bar{\nu}-\nu} & s\geq 0.\end{cases}$$

Proposition 11 (Large Countries, Union-Wide Inflation Targeting) *Suppose that the zero bound is not binding at the union level and that monetary policy targets union-wide inflation $\pi_t^* = 0$. Then in the Cole–Obstfeld case, we have*

$$c_t^i = -x(1-\xi)g_t^i + (1-x)\int_{-t}^{\infty}\alpha_s^{c,t,CM}g_{t+s}^i ds,$$

$$c_t^{-i} = -(1-\xi)xg_t^i + x\int_{-t}^{\infty}\alpha_s^{c,t,CM*}g_t^i ds.$$

Let us first focus on the regions in the country undertaking the spending. This proposition shows that for regions in the country undertaking the stimulus, the effects on private spending on domestic goods are simply a weighted average of the effect $-(1-\xi)g_t^i$ that would arise if the country undertaking the stimulus could set monetary policy to target their own domestic inflation $\pi_t^i = 0$, and the effect that arises if the country is a small (infinitesimal) part of a currency union, with weights given by x and $1-x$, where x is the relative size of the country undertaking the stimulus.

Let us now turn to the regions in countries not undertaking the spending. There are both direct effects and indirect effects. The indirect effects work through inflation, which affect the terms of trade and, hence, the demand for the goods produced by these regions. To isolate the direct effects set $\kappa = 0$, so that there is no inflation and $\alpha_s^{c,t,CM*} = 0$. The demand for home goods is then equal to $c_t^{-i} = -(1-\xi)g_t^* = -(1-\xi)xg_t^i$. When spending rises in regions $i \in [0, x]$, it depresses private spending by agents of these regions, lowering the demand for output in regions $-\ i \in (x, 1]$. When $\kappa > 0$, the indirect effect works through inflation. The lower demand for goods in regions $-\ i \in (x, 1]$ creates deflation in these regions, which makes these economies more competitive. The lower prices then increase the demand for the goods produced by these regions.

Example 7 (Union-Wide Inflation Targeting, AR(1)) Suppose that $g_t^i = g^i e^{-\rho_g t}$, then we have

$$c_t^{-i} = -e^{\nu t}(1-\xi)xg^i \left[1 - \frac{1 - e^{-(\nu+\rho_g)t}}{\rho_g + \nu} \frac{\rho_g(\rho + \rho_g)}{\rho_g + \bar{\nu}}\right].$$

This implies that c_0^{-i} is negative if g^i is positive. If $\rho_g + \nu < 0$ then c_t^{-i} will remain negative. If instead $\rho_g + \nu > 0$ then c_t^{-i} starts out negative, but eventually switches signs.

This results suggests that a temporary increase in government spending abroad accompanied by monetary tightening to ensure no union-wide inflation induces a recession at home. This fits a common narrative regarding the post German reunification in the early 90s. The fiscal expansion was combined with a monetary contraction in Germany, so as to avoid inflation. The quasi-fixed exchange rate arrangements of the EMS forced other countries to follow suit and tighten monetary policy, negatively affecting their economic performance.

10.2 Zero Bound at the Union Level

If the zero bound binds at the union level, then c_t^* is given by

$$c_t^* = x \int_0^\infty \alpha_s^c g_{t+s}^i ds.$$

The allocation for regions in the country undertaking the stimulus solves

$$\dot{\pi}_t^i = \rho \pi_t^i - \kappa(c_t^i + (1-\xi)g_t^i),$$
$$\dot{c}_t^i = -\hat{\sigma}^{-1}\pi_t^i,$$
$$c_0^i = x \int_0^\infty \alpha_s^c g_{t+s}^i ds.$$

Similarly the allocation for regions not undertaking the stimulus solves

$$\dot{\pi}_t^{-i} = \rho \pi_t^{-i} - \kappa c_t^{-i},$$
$$\dot{c}_t^{-i} = -\hat{\sigma}^{-1}\pi_t^{-i},$$

$$c_0^{-i} = x \int_0^{\infty} \alpha_s^c g_{t+s}^i ds.$$

Proposition 12 (Large Countries, Union-Wide Zero Bound) *Suppose that the zero bound is binding at the union level, then in the Cole–Obstfeld case, we have*

$$c_t^i = x \int_0^{\infty} \alpha_s^c g_{t+s}^i ds + (1-x) \int_{-t}^{\infty} \alpha_s^{c,t,CM} g_{t+s}^i ds,$$

$$c_t^{-i} = x e^{\nu t} \int_0^{\infty} \alpha_s^c g_s^i ds.$$

Similarly to Proposition 11, this proposition shows that for the country undertaking the stimulus, the effects on private spending on domestic goods are simply a weighted average of the effect $\int_0^{\infty} \alpha_s^c g_{t+s}^i ds$ that would arise if the country undertaking the stimulus were a closed economy at the zero lower bound, and the effect that arises if the country were a small (infinitesimal) part of a currency union, with weights given by x and $1 - x$, where x is the relative size of the country undertaking the stimulus.

In contrast to the inflation targeting case, when the zero lower bound binds, an increase in government spending by regions $i \in [0, x]$ increases the demand for the goods of regions $-i \in (x, 1]$. This is natural since we now have a general expansion in private demand because inflation reduces real interest rates.[u]

11. CONCLUSION

We have explored the economic response to changes in government spending in a few benchmark models. Relative to the existing literature, our contribution is to characterize the dynamics of these responses analytically in some detail, rather than summarizing the effects in a single "summary multiplier." We have done so by defining the multipliers to be the partial derivative of private spending at any point in time, to public spending at any other date. We have also attempted to be relatively exhaustive in incorporating various elements that are important, but sometimes missing in standard analyses. In particular, we considered both closed and open economies and incorporated hand-to-mouth agents in both these frameworks. Most importantly, our analysis is the first to emphasize different forms of financing for the government spending shock, including tax-financed, deficit-financed, and outside-financed. It is our hope that our approach and analysis will prove useful in interpreting and unifying the large theoretical and empirical research on fiscal multipliers.

[u] These findings on the spillover effects of fiscal policy complement the results in Cook and Devereux (2011) who focus on different configurations than us: they show that the spillover effects of fiscal policy at home on foreign when home is in a liquidity trap are negative with flexible exchange rates, but positive with fixed exchange rates. In this section, we focus on fixed exchange rates in a currency union and show how these spillover effects switch signs depending on whether the union is in a liquidity trap or targets inflation.

APPENDICES

Appendix A

This appendix derives the linear systems of equations to be solved for in order to derive fiscal multipliers in the following cases: liquidity trap; currency union with either complete markets (CM), incomplete markets (IM), and outside-financed government spending (PF). Appendix B then solves these systems equations to derive fiscal multipliers.

In both appendices, the general case with an arbitrary fraction χ of hand-to-mouth agents and with arbitrary profit offset o is derived first, followed by two special cases: no hand-to-mouth agents $\chi = 0$ (as in Sections 1–6) and no profit offset $o = 0$ (as in Sections 1–8).

Compared to the main text, the environment is generalized by allowing hand-to-mouth agents to receive a profit offset which redistributes a share of profits $o \in [0, 1]$ to hand-to-mouth agents:

$$P_t C_t^r = W_t N_t^r + \frac{o}{\chi}\Pi_t - P_t \underline{T}_t^r,$$

with

$$\begin{aligned} P_t \underline{T}_t^r &= P_t T_t^r - \frac{o}{\chi}\Pi_t, \\ \Pi_t &= P_{H,t} Y_t - w_t N_t. \end{aligned}$$

A.1 Liquidity Trap

Assume that $c_t^* = 0, i_t^* = \bar{r}_t$ for all $t \geq 0$. The log-linearized equations are

$$\begin{aligned} \dot{c}_t^o &= (1-\mathcal{G})\sigma^{-1}(i_t - \bar{r}_t - \pi_t), \\ c_t^r &= \frac{WN^r}{Y}\left(w_t + n_t^r\right) - \underline{t}_t^r, \\ w_t &= \frac{\sigma}{1-\mathcal{G}} c_t^r + \phi n_t^r, \\ w_t &= \frac{\sigma}{1-\mathcal{G}} c_t + \phi n_t, \\ c_t &= \chi c_t^r + (1-\chi) c_t^o, \\ n_t &= \chi n_t^r + (1-\chi) n_t^o, \\ \dot{\pi}_t &= \rho \pi_t - \kappa[c_t + (1-\xi) g_t], \\ \underline{t}_t^r &= t_t^r - o\left[\left(1 - \frac{1}{\mu}\right) n_t - \frac{1}{\mu} w_t\right], \end{aligned}$$

where w_t denotes real wages and μ is the steady state markup, with $\lambda = \rho_\delta(\rho + \rho_\delta), \kappa = \lambda(\hat{\sigma} + \phi), \xi = \dfrac{\hat{\sigma}}{\hat{\sigma} + \phi}$.

Combining and rearranging, we get

$$n_t^r = \phi^{-1}\left(w_t - \frac{\sigma}{1-\mathcal{G}}c_t^r\right),$$

$$c_t^r = \frac{WN^r}{Y}\left[\left(1+\phi^{-1}\right)\left(\frac{\sigma}{1-\mathcal{G}}c_t + \phi n_t\right) - \phi^{-1}\frac{\sigma}{1-\mathcal{G}}c_t^r\right] - \underline{t}_t^r,$$

$$c_t^r = \frac{\frac{WN^r}{Y}\left(1+\phi^{-1}\right)\left(\frac{\sigma}{1-\mathcal{G}}c_t + \phi n_t\right) - \underline{t}_t^r}{1+\phi^{-1}\frac{\sigma}{1-\mathcal{G}}\frac{WN^r}{Y}},$$

$$c_t\left[\frac{1-\chi\frac{WN^r}{Y}\frac{\sigma}{1-\mathcal{G}} + (1-\chi)\phi^{-1}\frac{\sigma}{1-\mathcal{G}}\frac{WN^r}{Y}}{1+\phi^{-1}\frac{\sigma}{1-\mathcal{G}}\frac{WN^r}{Y}}\right] = \chi\frac{\frac{WN^r}{Y}\left(1+\phi^{-1}\right)\phi n_t - \underline{t}_t^r}{1+\phi^{-1}\frac{\sigma}{1-\mathcal{G}}\frac{WN^r}{Y}} + (1-\chi)c_t^o,$$

$$c_t = \chi\frac{\frac{WN^r}{Y}\left(1+\phi^{-1}\right)\phi n_t - \underline{t}_t^r}{1-\chi\frac{WN^r}{Y}\frac{\sigma}{1-\mathcal{G}} + (1-\chi)\phi^{-1}\frac{\sigma}{1-\mathcal{G}}\frac{WN^r}{Y}} + (1-\chi)\frac{1+\phi^{-1}\frac{\sigma}{1-\mathcal{G}}\frac{WN^r}{Y}}{1-\chi\frac{WN^r}{Y}\frac{\sigma}{1-\mathcal{G}} + (1-\chi)\phi^{-1}\frac{\sigma}{1-\mathcal{G}}\frac{WN^r}{Y}}c_t^o,$$

$$c_t = \chi\frac{\phi(1+\phi)n_t - \frac{Y}{WN^r}\phi\underline{t}_t^r}{\frac{Y}{WN^r}\phi - \chi\frac{\sigma}{1-\mathcal{G}}\phi + (1-\chi)\frac{\sigma}{1-\mathcal{G}}} + (1-\chi)\frac{\frac{Y}{WN^r}\phi + \frac{\sigma}{1-\mathcal{G}}}{\frac{Y}{WN^r}\phi - \chi\frac{\sigma}{1-\mathcal{G}}\phi + (1-\chi)\frac{\sigma}{1-\mathcal{G}}}c_t^o,$$

$$c_t = \chi(1-\mathcal{G})\frac{\phi(1+\phi)n_t - \mu\phi\underline{t}_t^r}{(1-\mathcal{G})\mu\phi + \sigma - \chi\sigma(1+\phi)} + (1-\chi)\frac{(1-\mathcal{G})\mu\phi + \sigma}{(1-\mathcal{G})\mu\phi + \sigma - \chi\sigma(1+\phi)}c_t^o,$$

and finally

$$c_t = \Theta_n n_t - \Theta_\tau \underline{t}_t^r + \bar{\sigma}^{-1}\sigma\frac{1}{1-\mathcal{G}}c_t^o,$$

where

$$\bar{\sigma}^{-1} = \sigma^{-1}(1-\chi)(1-\mathcal{G})\frac{(1-\mathcal{G})\mu\phi + \sigma}{\phi(1-\mathcal{G})\mu + \sigma - \chi\sigma(1+\phi)},$$

$$\Theta_n = \chi(1-\mathcal{G})\frac{(1+\phi)\phi}{\phi(1-\mathcal{G})\mu + \sigma - \chi\sigma(1+\phi)},$$

$$\Theta_\tau = \chi(1-\mathcal{G})\frac{\mu\phi}{\phi(1-\mathcal{G})\mu + \sigma - \chi\sigma(1+\phi)},$$

Differentiating, we get

$$c_t = \Theta_n \dot{n}_t - \Theta_\tau \underline{t}_t^r + \bar{\sigma}^{-1}(i_t - \bar{r}_t - \pi_t),$$

and using $\dot{n}_t = \dot{c}_t + \dot{g}_t$, we find the Euler equation

$$\dot{c}_t = \tilde{\sigma}^{-1}(i_t - \bar{r}_t - \pi_t) + \tilde{\Theta}_n \dot{g}_t - \tilde{\Theta}_\tau \underline{t}_t^r,$$

where

$$\tilde{\sigma}^{-1} = \frac{\bar{\sigma}^{-1}}{1 - \Theta_n},$$

$$\tilde{\Theta}_n = \frac{\Theta_n}{1 - \Theta_n},$$

$$\tilde{\Theta}_\tau = \frac{\Theta_\tau}{1 - \Theta_n}.$$

By definition of t_t^r and using the expression for the wage,

$$\underline{t}_t^r = t_t^r - \frac{o}{\chi}\left[\left(1 - \frac{1}{\mu}\right)(c_t + g_t) - \frac{1}{\mu}\left[\frac{\sigma}{1 - \mathcal{G}} c_t + \phi(c_t + g_t)\right]\right].$$

Thus,

$$\underline{t}_t^r = t_t^r + \psi_c c_t + \psi_n g_t,$$

where

$$\psi_c = -\frac{o}{\chi}\left[1 - \frac{1}{\mu}\left(\frac{\sigma}{1 - \mathcal{G}} + (1 + \phi)\right)\right],$$

$$\psi_n = -\frac{o}{\chi}\left[1 - \frac{1}{\mu}(1 + \phi)\right].$$

Using the Euler equation and the expression for t_t^r, we get

$$[1 - \Theta_n + \Theta_\tau \psi_c]\dot{c}_t = -\bar{\sigma}^{-1}\pi_t + [\Theta_n - \Theta_\tau \psi_n]\dot{g}_t - \Theta_\tau \dot{t}_t^r.$$

Thus,

$$\dot{c}_t = -\underline{\tilde{\sigma}}^{-1}\pi_t + \underline{\tilde{\Theta}}_n \dot{g}_t - \underline{\tilde{\Theta}}_\tau \dot{t}_t^r,$$

where

$$\underline{\tilde{\sigma}}^{-1} = \frac{1}{\underline{\tilde{\Theta}}_c}\bar{\sigma}^{-1},$$

$$\underline{\tilde{\Theta}}_n = \frac{1}{\underline{\tilde{\Theta}}_c}[\Theta_n - \Theta_\tau \psi_n],$$

$$\underline{\tilde{\Theta}}_\tau = \frac{1}{\underline{\tilde{\Theta}}_c}\Theta_\tau,$$

$$\underline{\tilde{\Theta}}_c = 1 - \Theta_n + \Theta_\tau \psi_c.$$

Special case: no hand-to-mouth agents $\chi = 0$

The log-linear system is

$$\begin{aligned}\dot{c}_t &= -\hat{\sigma}^{-1}\pi_t,\\ \dot{\pi}_t &= \rho\pi_t - \kappa[c_t + (1-\xi)g_t],\end{aligned}$$

for all $t \geq 0$.

Special case: no profit offset $o = 0$

The log-linear system is

$$\begin{aligned}\dot{c}_t &= -\widetilde{\sigma}^{-1}\pi_t + \widetilde{\Theta}_n\dot{g}_t - \widetilde{\Theta}_\tau\dot{t}^r_t,\\ \dot{\pi}_t &= \rho\pi_t - \kappa[c_t + (1-\xi)g_t].\end{aligned}$$

for all $t \geq 0$.

A.2 Currency Union

Assume that $c^*_t = 0, i^*_t = \bar{r}_t$ for all $t \geq 0$. The log-linearized equations are

$$c^o_t = (1-\mathcal{G})\theta + \frac{(1-\alpha)(1-\mathcal{G})}{\sigma}s_t,$$

$$y_t = (1-\alpha)\hat{c}_t + (1-\mathcal{G})\alpha\left[\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma}\right]s_t + g_t,$$

$$y_t = n_t,$$

$$\dot{c}^o_t = -(1-\mathcal{G})\sigma^{-1}(\pi_{H,t} + \alpha\dot{s}_t),$$

$$c^r_t = \frac{1}{\mu}\left(w_t + n^r_t\right) - \underline{t}^r_t,$$

$$w_t = \frac{\sigma}{1-\mathcal{G}}c^r_t + \phi n^r_t,$$

$$w_t = \frac{\sigma}{1-\mathcal{G}}\hat{c}_t + \phi n_t,$$

$$\hat{c}_t = \chi c^r_t + (1-\chi)c^o_t,$$

$$n_t = \chi n^r_t + (1-\chi)n^o_t,$$

$$\dot{\pi}_{H,t} = \rho\pi_{H,t} - \lambda(w_t + \alpha s_t),$$

$$\int_0^{+\infty} e^{-\rho t} nx_t dt = -\mathrm{nfa}_0,$$

$$\underline{t}^r_t = t^r_t - \frac{o}{\chi}\left[\left(1-\frac{1}{\mu}\right)n_t + \alpha p_{H,t} - \frac{1}{\mu}w_t\right],$$

with $\mathrm{nfa}_0 = 0$ in the IM case and $\mathrm{nfa}_0 = \int_0^{+\infty} e^{-\rho t}g_t dt$ in the PF case, where $\omega = \sigma\gamma + (1-\alpha)(\sigma\eta - 1)$. Note that we have denoted total consumption of home agents by $\hat{c}_t$ to avoid a confusion with c_t, the total consumption of home goods by private agents (both home and foreign).

Using the expressions for the wage, aggregate consumption, and labor,

$$\hat{c}_t = \Theta_n n_t - \Theta_\tau \underline{t}_t^r + \bar{\sigma}^{-1} \sigma \frac{1}{1-\mathcal{G}} c_t^o,$$

where Θ_n, Θ_τ, and $\bar{\sigma}$ have been defined above. Differentiating the Backus–Smith condition, we get (we could have gotten this equation directly from the definition of s_t)

$$\dot{s}_t = -\pi_{H,t}.$$

Now we can get to an equation involving total (home + foreign) consumption of the domestic good $c_t = y_t - g_t$ which yields

$$c_t = (1-\alpha)\hat{c}_t + (1-\mathcal{G})\alpha \left[\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma} \right] s_t.$$

Differentiating, we get

$$\dot{c}_t = (1-\alpha)\dot{\hat{c}}_t + (1-\mathcal{G})\alpha \left[\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma} \right] \dot{s}_t,$$

then combining with the equation for $\hat{c}_t$,

$$\dot{c}_t = (1-\alpha) \left[\Theta_n \dot{n}_t - \Theta_\tau \underline{t}_t^r + \bar{\sigma}^{-1} \sigma \frac{1}{1-\mathcal{G}} \dot{c}_t^o \right] + (1-\mathcal{G})\alpha \left[\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma} \right] \dot{s}_t,$$

and replacing $n_t = c_t + g_t$,

$$\dot{c}_t = (1-\alpha) \left[\Theta_n (\dot{c}_t + \dot{g}_t) - \Theta_\tau \underline{t}_t^r + \bar{\sigma}^{-1} \sigma \frac{1}{1-\mathcal{G}} \dot{c}_t^o \right] + (1-\mathcal{G})\alpha \left[\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma} \right] \dot{s}_t,$$

and rearranging

$$\dot{c}_t = \widetilde{\Theta}_n \dot{g}_t - \widetilde{\Theta}_\tau \underline{t}_t^r + \frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n} \sigma \frac{1}{1-\mathcal{G}} \dot{c}_t^o + \frac{1}{1-(1-\alpha)\Theta_n} \frac{\alpha(1-\mathcal{G})(\omega+1-\alpha)}{\sigma} \dot{s}_t,$$

where

$$\widetilde{\Theta}_n = \frac{(1-\alpha)\Theta_n}{1-(1-\alpha)\Theta_n},$$

$$\widetilde{\Theta}_\tau = \frac{(1-\alpha)\Theta_\tau}{1-(1-\alpha)\Theta_n},$$

then using the Euler equation for optimizers

$$\dot{c}_t = \widetilde{\Theta}_n \dot{g}_t - \widetilde{\Theta}_\tau \dot{\underline{t}}_t^r - \frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n} [\pi_{H,t} + \alpha \dot{s}_t] + \frac{1}{1-(1-\alpha)\Theta_n} \frac{\alpha(1-\mathcal{G})(\omega+1-\alpha)}{\sigma} \dot{s}_t,$$

and finally combining with the expression for $\dot{s}_t = -\pi_{H,t}$

$$\dot{c}_t = \widetilde{\Theta}_n \dot{g}_t - \widetilde{\Theta}_\tau \dot{\underline{t}}_t^r - \frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}[(1-\alpha)\pi_{H,t}]$$
$$-\frac{1}{1-(1-\alpha)\Theta_n}\frac{\alpha(1-\mathcal{G})(\omega+1-\alpha)}{\sigma}\dot{\pi}_{H,t},$$

which we can rewrite as

$$\dot{c}_t = \widetilde{\Theta}_n \dot{g}_t - \widetilde{\Theta}_\tau \dot{\underline{t}}_t^r - \frac{\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}\left[(1-\alpha)^2 + \alpha\frac{\bar{\sigma}}{\sigma}(1-\mathcal{G})(\omega+1-\alpha)\right]\dot{\pi}_{H,t},$$
$$\dot{c}_t = \widetilde{\Theta}_n \dot{g}_t - \widetilde{\Theta}_\tau \dot{\underline{t}}_t^r - \widetilde{\sigma}^{-1}\pi_{H,t},$$

where

$$\widetilde{\sigma}^{-1} = \frac{\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}\left[(1-\alpha)^2 + \alpha\frac{\bar{\sigma}}{\sigma}(1-\mathcal{G})(\omega+1-\alpha)\right].$$

This is our Euler equation.[v]

To derive an initial condition, we use

$$c_t = (1-\alpha)\hat{c}_t + (1-\mathcal{G})\alpha\left[\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma}\right]s_t,$$
$$\hat{c}_t = \Theta_n n_t - \Theta_\tau \underline{t}_t^r + \bar{\sigma}^{-1}\sigma\frac{1}{1-\mathcal{G}}c_t^o,$$
$$c_t^o = (1-\mathcal{G})\theta + \frac{(1-\alpha)(1-\mathcal{G})}{\sigma}s_t,$$

and

$$n_t = c_t + g_t,$$

to get

$$c_t = \widetilde{\Theta}_n g_t - \widetilde{\Theta}_\tau \underline{t}_t^r + \frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}\sigma\frac{1}{1-\mathcal{G}}\left((1-\mathcal{G})\theta + \frac{(1-\alpha)(1-\mathcal{G})}{\sigma}s_t\right)$$
$$+\frac{(1-\mathcal{G})\alpha\left[\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma}\right]}{1-(1-\alpha)\Theta_n}s_t,$$

and apply it at $t=0$ with $s_0 = 0$ to get

[v] We can check that when there are no hand-to-mouth consumers, this boils down to

$$\dot{c}_t = -\sigma^{-1}(1-\mathcal{G})[1+\alpha(\omega-1)]\pi_t,$$

which is exactly the expression that we found.

$$c_0 = \widetilde{\Theta}_n g_0 - \widetilde{\Theta}_\tau \underline{t}_0^r + \frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n} \sigma \frac{1}{1-\mathcal{G}}(1-\mathcal{G})\theta.$$

Hence with complete markets, this boils down to the simple condition

$$c_0 = \widetilde{\Theta}_n g_0 - \widetilde{\Theta}_\tau \underline{t}_0^r.$$

Finally we need to compute

$$mc_t = w_t + p_t - p_{H,t} = w_t + \alpha s_t,$$

We have

$$w_t = \frac{\sigma}{1-\mathcal{G}} \hat{c}_t + \phi n_t,$$
$$w_t = \frac{\sigma}{1-\mathcal{G}} \hat{c}_t + \phi (c_t + g_t),$$

which using

$$\hat{c}_t = \Theta_n n_t - \Theta_\tau \underline{t}_t^r + \bar{\sigma}^{-1} \sigma \frac{1}{1-\mathcal{G}} c_t^o,$$

we can rewrite as

$$w_t = \frac{\sigma}{1-\mathcal{G}} \left(\Theta_n (c_t + g_t) - \Theta_\tau \underline{t}_t^r + \bar{\sigma}^{-1} \sigma \frac{1}{1-\mathcal{G}} c_t^o \right) + \phi (c_t + g_t),$$
$$w_t = \frac{\sigma}{1-\mathcal{G}} \left[\Theta_n (c_t + g_t) - \Theta_\tau \underline{t}_t^r + \bar{\sigma}^{-1} \sigma \frac{1}{1-\mathcal{G}} \left((1-\mathcal{G})\theta + \frac{(1-\alpha)(1-\mathcal{G})}{\sigma} s_t \right) \right] + \phi (c_t + g_t),$$

so that

$$\begin{aligned} w_t + \alpha s_t = {} & \left(\frac{\sigma \Theta_n}{1-\mathcal{G}} + \phi \right) (c_t + g_t) - \frac{\sigma}{1-\mathcal{G}} \Theta_\tau \underline{t}_t^r + \left(\frac{\sigma}{1-\mathcal{G}} \right)^2 \bar{\sigma}^{-1} (1-\mathcal{G})\theta \\ & + \left[\alpha + \left(\frac{\sigma}{1-\mathcal{G}} \right)^2 \bar{\sigma}^{-1} \frac{(1-\alpha)(1-\mathcal{G})}{\sigma} \right] s_t, \end{aligned}$$

which using

$$\begin{aligned} c_t = {} & \widetilde{\Theta}_n g_t - \widetilde{\Theta}_\tau \underline{t}_t^r + \frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n} \sigma \frac{1}{1-\mathcal{G}} (1-\mathcal{G})\theta \\ & + \left[\frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n} \sigma \frac{1}{1-\mathcal{G}} \frac{(1-\alpha)(1-\mathcal{G})}{\sigma} + \frac{(1-\mathcal{G})\alpha \left[\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma} \right]}{1-(1-\alpha)\Theta_n} \right] s_t, \end{aligned}$$

i.e.

$$s_t = \frac{c_t - \widetilde{\Theta}_n g_t + \widetilde{\Theta}_\tau \underline{t}_t^r - \dfrac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}\sigma\dfrac{1}{1-\mathcal{G}}(1-\mathcal{G})\theta}{\dfrac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}\sigma\dfrac{1}{1-\mathcal{G}}\dfrac{(1-\alpha)(1-\mathcal{G})}{\sigma} + \dfrac{(1-\mathcal{G})\alpha\left[\dfrac{\omega}{\sigma}+\dfrac{1-\alpha}{\sigma}\right]}{1-(1-\alpha)\Theta_n}},$$

we can rewrite as

$$\begin{aligned} w_t + \alpha s_t = {} & \left(\frac{\sigma\Theta_n}{1-\mathcal{G}} + \phi\right)(c_t + g_t) - \frac{\sigma}{1-\mathcal{G}}\Theta_\tau \underline{t}_t^r + \left(\frac{\sigma}{1-\mathcal{G}}\right)^2 \bar{\sigma}^{-1}(1-\mathcal{G})\theta \\ & + \frac{\alpha + \dfrac{\sigma}{1-\mathcal{G}}\bar{\sigma}^{-1}(1-\alpha)}{\dfrac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}(1-\alpha) + \dfrac{(1-\mathcal{G})\alpha\left[\dfrac{\omega}{\sigma}+\dfrac{1-\alpha}{\sigma}\right]}{1-(1-\alpha)\Theta_n}} \\ & \times \left[c_t - \widetilde{\Theta}_n g_t + \widetilde{\Theta}_\tau \underline{t}_t^r - \frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}\sigma\frac{1}{1-\mathcal{G}}(1-\mathcal{G})\theta\right], \end{aligned}$$

$$\begin{aligned} w_t + \alpha s_t = {} & \left(\frac{\sigma\Theta_n}{1-\mathcal{G}} + \phi\right)(c_t + g_t) - \frac{\sigma}{1-\mathcal{G}}\Theta_\tau \underline{t}_t^r + \left(\frac{\sigma}{1-\mathcal{G}}\right)^2 \bar{\sigma}^{-1}(1-\mathcal{G})\theta \\ & + \frac{\alpha + \dfrac{\sigma}{1-\mathcal{G}}\bar{\sigma}^{-1}(1-\alpha)}{\dfrac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}(1-\alpha) + \dfrac{(1-\mathcal{G})\alpha\left[\dfrac{\omega}{\sigma}+\dfrac{1-\alpha}{\sigma}\right]}{1-(1-\alpha)\Theta_n}} \\ & \times \left[c_t - \widetilde{\Theta}_n g_t + \widetilde{\Theta}_\tau \underline{t}_t^r - \frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}\sigma\frac{1}{1-\mathcal{G}}(1-\mathcal{G})\theta\right]. \end{aligned}$$

We can then replace this expression in to get the New Keynesian Phillips Curve

$$\dot{\pi}_{H,t} = \rho\pi_{H,t} - \lambda(w_t + \alpha s_t).$$

The system is summarized by

$$\begin{aligned} \dot{c}_t &= \widetilde{\Theta}_n \dot{g}_t - \widetilde{\Theta}_\tau \underline{t}_t^r - \widetilde{\sigma}^{-1}\pi_{H,t}, \\ \dot{\pi}_{H,t} &= \rho\pi_{H,t} - \lambda(w_t + \alpha s_t), \\ c_0 &= \widetilde{\Theta}_n g_0 - \widetilde{\Theta}_\tau \underline{t}_0^r + \frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}\sigma\frac{1}{1-\mathcal{G}}(1-\mathcal{G})\theta, \end{aligned}$$

and the nfa condition, where

$$
\begin{aligned}
w_t + \alpha s_t = {} & \left(\frac{\sigma\Theta_n}{1-\mathcal{G}} + \phi\right)(c_t + g_t) - \frac{\sigma}{1-\mathcal{G}}\Theta_\tau \underline{t}_t^r + \left(\frac{\sigma}{1-\mathcal{G}}\right)^2 \bar{\sigma}^{-1}(1-\mathcal{G})\theta \\
& + \frac{\alpha + \frac{\sigma}{1-\mathcal{G}}\bar{\sigma}^{-1}(1-\alpha)}{\frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}(1-\alpha) + \frac{(1-\mathcal{G})\alpha\left[\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma}\right]}{1-(1-\alpha)\Theta_n}} \\
& \times \left[c_t - \widetilde{\Theta}_n g_t + \widetilde{\Theta}_\tau \underline{t}_t^r - \frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}\sigma\frac{1}{1-\mathcal{G}}(1-\mathcal{G})\theta\right].
\end{aligned}
$$

Define $\widetilde{\kappa}$ by

$$
\widetilde{\kappa} = \lambda\left[\frac{\sigma\Theta_n}{1-\mathcal{G}} + \phi + \frac{\alpha + \frac{\sigma}{1-\mathcal{G}}\bar{\sigma}^{-1}(1-\alpha)}{\frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}(1-\alpha) + \frac{(1-\mathcal{G})\alpha\left[\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma}\right]}{1-(1-\alpha)\Theta_n}}\right].
$$

Define $\widetilde{\xi}$ by

$$
\widetilde{\kappa}\left(1-\widetilde{\xi}\right) = \lambda\left[\frac{\sigma\Theta_n}{1-\mathcal{G}} + \phi - \frac{\alpha + \frac{\sigma}{1-\mathcal{G}}\bar{\sigma}^{-1}(1-\alpha)}{\frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}(1-\alpha) + \frac{(1-\mathcal{G})\alpha\left[\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma}\right]}{1-(1-\alpha)\Theta_n}}\widetilde{\Theta}_n\right].
$$

Define $\widetilde{\alpha}$ by

$$
\widetilde{\alpha} = 1 - \frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}\sigma\frac{1}{1-\mathcal{G}}.
$$

Define $\widetilde{\omega}$ by

$$
\begin{aligned}
\widetilde{\omega} = {} & \frac{1}{(1-\mathcal{G})\widetilde{\sigma}\widetilde{\alpha}} \\
& \times\left[\left(\frac{\sigma}{1-\mathcal{G}}\right)^2\bar{\sigma}^{-1}(1-\mathcal{G}) - \frac{\alpha + \frac{\sigma}{1-\mathcal{G}}\bar{\sigma}^{-1}(1-\alpha)}{\frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}(1-\alpha) + \frac{(1-\mathcal{G})\alpha\left(\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma}\right)}{1-(1-\alpha)\Theta_n}}\frac{(1-\alpha)\bar{\sigma}^{-1}\sigma}{1-(1-\alpha)\Theta_n}\right].
\end{aligned}
$$

Define $\widetilde{\widetilde{\Theta}}_\tau$ by

$$\widetilde{\widetilde{\Theta}}_\tau = \frac{\lambda}{\widetilde{\widetilde{\kappa}}}\left[-\frac{\sigma}{1-\mathcal{G}}\Theta_\tau + \frac{\alpha + \dfrac{\sigma}{1-\mathcal{G}}\bar{\sigma}^{-1}(1-\alpha)}{\dfrac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}(1-\alpha) + \dfrac{(1-\mathcal{G})\alpha\left[\dfrac{\omega}{\sigma} + \dfrac{1-\alpha}{\sigma}\right]}{1-(1-\alpha)\Theta_n}}\widetilde{\Theta}_\tau\right].$$

Define Γ_1 by

$$\Gamma_1 = (1-\alpha)^2\bar{\sigma}^{-1} + (1-\mathcal{G})\alpha\left(\frac{\omega}{\sigma} + \frac{1-\alpha}{\sigma}\right),$$

Then we can rewrite the system as

$$\begin{aligned}
\dot{\pi}_{H,t} &= \rho\pi_{H,t} - \widetilde{\kappa}\left(c_t + \left(1-\widetilde{\xi}\right)g_t\right) - (1-\mathcal{G})\lambda\widetilde{\sigma}\widetilde{\alpha}\widetilde{\omega}\theta - \widetilde{\kappa}\widetilde{\widetilde{\Theta}}_\tau \underline{t}_t^r,\\
\dot{c}_t &= -\widetilde{\sigma}^{-1}\pi_{H,t} + \widetilde{\Theta}_n\dot{g}_t - \widetilde{\Theta}_\tau\dot{\underline{t}}_t^r,
\end{aligned}$$

with an initial condition

$$c_0 = (1-\mathcal{G})(1-\widetilde{\alpha})\theta + \widetilde{\Theta}_n g_0 - \widetilde{\Theta}_\tau \underline{t}_0^r,$$

and the nfa condition.

For net exports we get

$$\begin{aligned}
nx_t &= -(1-\mathcal{G})\alpha s_t + y_t - \hat{c}_t - g_t,\\
nx_t &= (1-\mathcal{G})\left[\alpha\frac{\omega}{\sigma} + \alpha\frac{1-\alpha}{\sigma} - \alpha\right]s_t - \alpha\hat{c}_t,\\
nx_t &= (1-\mathcal{G})\left[\alpha\frac{\omega}{\sigma} + \alpha\frac{1-\alpha}{\sigma} - \alpha\right]s_t\\
&\quad -\alpha\left[\Theta_n(c_t+g_t) - \Theta_\tau\underline{t}_t^r + \bar{\sigma}^{-1}\sigma\frac{1}{1-\mathcal{G}}c_t^o\right],
\end{aligned}$$

and finally

$$\begin{aligned}
nx_t &= (1-\mathcal{G})\left[\alpha\frac{\omega}{\sigma} + \alpha\frac{1-\alpha}{\sigma} - \alpha\right]s_t\\
&\quad -\alpha\left[\Theta_n(c_t+g_t) - \Theta_\tau\underline{t}_t^r + \bar{\sigma}^{-1}\sigma\frac{1}{1-\mathcal{G}}\left((1-\mathcal{G})\theta + \frac{(1-\alpha)(1-\mathcal{G})}{\sigma}s_t\right)\right],
\end{aligned}$$

where

$$s_t = \frac{c_t - \widetilde{\Theta}_n g_t + \widetilde{\Theta}_\tau \underline{t}_t^r - \frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}\sigma\frac{1}{1-\mathcal{G}}(1-\mathcal{G})\theta}{\frac{(1-\alpha)\bar{\sigma}^{-1}}{1-(1-\alpha)\Theta_n}\sigma\frac{1}{1-\mathcal{G}}\frac{(1-\alpha)(1-\mathcal{G})}{\sigma} + \frac{(1-\mathcal{G})\alpha\left[\frac{\omega}{\sigma}+\frac{1-\alpha}{\sigma}\right]}{1-(1-\alpha)\Theta_n}}.$$

Using the Euler equation,

$$p_{H,t} = -s_0 - \frac{1}{\Gamma_1}[1-(1-\alpha)\Theta_n](c_t - c_0) + \frac{1}{\Gamma_1}(1-\alpha)\Theta_n(g_t - g_0) - (1-\alpha)\frac{1}{\Gamma_1}\Theta_\tau\left(\underline{t}_t^r - \underline{t}_0^r\right).$$

Using the initial condition for consumption,

$$p_{H,t} = -\frac{1}{\Gamma_1}[1-(1-\alpha)\Theta_n]c_t + \frac{1}{\Gamma_1}(1-\alpha)\Theta_n g_t - (1-\alpha)\frac{1}{\Gamma_1}\Theta_\tau \underline{t}_t^r + (1-\alpha)\frac{1}{\Gamma_1}\bar{\sigma}^{-1}\sigma\theta,$$

since $s_0 = 0$.

By definition of $\underline{t}_t^r$ and using the expressions for output, for prices and for the real wage,

$$\begin{aligned}\underline{t}_t^r = t_t^r - \frac{o}{\chi}\Bigg[&\left(1-\frac{1}{\mu}\right)(c_t + g_t)\\ &+\alpha\left[-\frac{1}{\Gamma_1}[1-(1-\alpha)\Theta_n]c_t + \frac{1}{\Gamma_1}(1-\alpha)\Theta_n g_t - (1-\alpha)\frac{1}{\Gamma_1}\Theta_\tau \underline{t}_t^r + (1-\alpha)\frac{1}{\Gamma_1}\bar{\sigma}^{-1}\sigma\theta\right]\\ &-\frac{1}{\mu}\left[\frac{\sigma}{1-\mathcal{G}}\frac{1}{1-\alpha}\left[c_t - (1-\mathcal{G})\alpha\left(\frac{\omega}{\sigma}+\frac{1-\alpha}{\sigma}\right)s_t\right] + \phi n_t\right]\Bigg].\end{aligned}$$

Using the expression for the terms-of-trade,

$$\begin{aligned}\underline{t}_t^r = t_t^r - \frac{o}{\chi}\Bigg[&\left(1-\frac{1}{\mu}\right)(c_t + g_t)\\ &+\alpha\left[-\frac{1}{\Gamma_1}[1-(1-\alpha)\Theta_n]c_t + \frac{1}{\Gamma_1}(1-\alpha)\Theta_n g_t - (1-\alpha)\frac{1}{\Gamma_1}\Theta_\tau \underline{t}_t^r + (1-\alpha)\bar{\sigma}^{-1}\sigma\theta\right]\\ &-\frac{1}{\mu}\frac{\sigma}{1-\mathcal{G}}\frac{1}{1-\alpha}c_t\\ &+\frac{1}{\mu}\sigma\frac{\alpha}{1-\alpha}\left(\frac{\omega}{\sigma}-\frac{1-\alpha}{\sigma}\right)\frac{1}{\Gamma_1}\left[[1-(1-\alpha)\Theta_n]c_t - (1-\alpha)\Theta_n g_t + (1-\alpha)\Theta_\tau \underline{t}_t^r - (1-\alpha)\bar{\sigma}^{-1}\sigma\theta\right]\\ &-\frac{1}{\mu}\phi n_t\Bigg].\end{aligned}$$

Thus,

$$\underline{t}_t^r = \psi_\tau t_t^r + \psi_c c_t + \psi_n g_t + \psi_\theta \theta,$$

where

$$\psi_c = -\frac{o}{\chi}\frac{1}{\hat{\psi}_\tau}\left[\left(1-\frac{1}{\mu}\right) - \alpha\frac{1}{\Gamma_1}[1-(1-\alpha)\Theta_n] - \frac{1}{\mu}\left[\frac{\sigma}{1-\alpha}\left[\frac{1}{1-\mathcal{G}} - \alpha\left(\frac{\omega}{\sigma}+\frac{1-\alpha}{\sigma}\right)\frac{1}{\Gamma_1}[1-(1-\alpha)\Theta_n]\right]+\phi\right]\right],$$

$$\psi_n = -\frac{o}{\chi}\frac{1}{\hat{\psi}_\tau}\left[\left(1-\frac{1}{\mu}\right) + \alpha\frac{1}{\Gamma_1}(1-\alpha)\Theta_n - \frac{1}{\mu}\left[\sigma\alpha\left(\frac{\omega}{\sigma}+\frac{1-\alpha}{\sigma}\right)\frac{1}{\Gamma_1}\Theta_n + \phi\right]\right],$$

$$\psi_\theta = -\frac{o}{\chi}\frac{1}{\hat{\psi}_\tau}\left[\alpha\frac{1}{\Gamma_1}(1-\alpha)\bar{\sigma}^{-1}\sigma - \frac{1}{\mu}\sigma\alpha\left(\frac{\omega}{\sigma}+\frac{1-\alpha}{\sigma}\right)\frac{1}{\Gamma_1}\bar{\sigma}^{-1}\sigma\right],$$

$$\psi_\tau = \frac{1}{\hat{\psi}_\tau},$$

$$\hat{\psi}_\tau = 1 - \frac{o}{\chi}\left[\alpha\frac{1}{\Gamma_1}(1-\alpha)\Theta_\tau - \frac{1}{\mu}\sigma\alpha\left(\frac{\omega}{\sigma}+\frac{1-\alpha}{\sigma}\right)\frac{1}{\Gamma_1}\Theta_\tau\right].$$

Using the Euler equation and the expression for $\underline{t}_t^r$,

$$[1-(1-\alpha)\Theta_n + (1-\alpha)\Theta_\tau\psi_c]\dot{c}_t = \Gamma_1\pi_{H,t} + (1-\alpha)[\Theta_n - \Theta_\tau\psi_n]\dot{g}_t - (1-\alpha)\Theta_\tau\psi_\tau t_t^r.$$

Thus,

$$\dot{c}_t = -\underline{\widetilde{\sigma}}^{-1}\pi_{H,t} + \underline{\widetilde{\Theta}}_n\dot{g}_t - \underline{\widetilde{\Theta}}_\tau\dot{t}_t^r,$$

where

$$\begin{aligned}
\underline{\widetilde{\sigma}}^{-1} &= \frac{1}{\underline{\widetilde{\Theta}}_c}\,\Gamma_1,\\
\underline{\widetilde{\Theta}}_n &= (1-\alpha)\frac{1}{\underline{\widetilde{\Theta}}_c}\,[\Theta_n - \Theta_\tau\psi_n],\\
\underline{\widetilde{\Theta}}_\tau &= (1-\alpha)\frac{1}{\underline{\widetilde{\Theta}}_c}\,\Theta_\tau\psi_\tau,\\
\underline{\widetilde{\Theta}}_c &= 1-(1-\alpha)\Theta_n + (1-\alpha)\Theta_\tau\psi_c.
\end{aligned}$$

Using the New Keynesian Phillips Curve and the expression for $\underline{t}_t^r$,

$$\dot{\pi}_{H,t} = \rho\pi_{H,t} - \widetilde{\kappa}\left[c_t + \left(1-\widetilde{\xi}\right)g_t\right] - (1-\mathcal{G})\lambda\widetilde{\sigma}\widetilde{\alpha}\widetilde{\omega}\theta - \widetilde{\kappa}\widetilde{\widetilde{\Theta}}_\tau\left[\psi_\tau t_t^r + \psi_c c_t + \psi_n g_t + \psi_\theta\theta\right].$$

Thus,

$$\dot{\pi}_{H,t} = \rho\pi_{H,t} - \underline{\widetilde{\kappa}}_c c_t - \underline{\widetilde{\kappa}}_n g_t - \underline{\widetilde{\kappa}}_\theta\theta - \underline{\widetilde{\kappa}}_\tau t_t^r,$$

where

$$\underline{\widetilde{\kappa}}_c = \widetilde{\kappa}\left(1 + \widetilde{\widetilde{\widetilde{\Theta}}}_\tau \psi_c\right),$$

$$\underline{\widetilde{\kappa}}_n = \widetilde{\kappa}\left(1 - \widetilde{\xi} + \widetilde{\widetilde{\widetilde{\Theta}}}_\tau \psi_n\right),$$

$$\underline{\widetilde{\kappa}}_\theta = (1 - \mathcal{G})\lambda\widetilde{\sigma}\widetilde{\alpha}\widetilde{\omega} + \widetilde{\kappa}\widetilde{\widetilde{\widetilde{\Theta}}}_\tau \psi_\theta,$$

$$\underline{\widetilde{\kappa}}_\tau = \widetilde{\kappa}\widetilde{\widetilde{\widetilde{\Theta}}}_\tau \psi_\tau.$$

Using the initial condition for consumption and the expression for $\underline{t}_t^r$,

$$c_0 = \Upsilon\theta + \underline{\widetilde{\Theta}}_n g_0 - \underline{\widetilde{\Theta}}_\tau t_0^r,$$

where

$$\Upsilon = \frac{1}{\underline{\widetilde{\widetilde{\Theta}}}_c}\left[(1 - \mathcal{G})(1 - \widetilde{\alpha})[1 - (1 - \alpha)\Theta_n] - (1 - \alpha)\Theta_\tau \psi_\theta\right].$$

Using the expressions for net exports and for $\underline{t}_t^r$,

$$nx_t = \alpha(1 - \mathcal{G})\left[\frac{1}{1 - \alpha}\left(\frac{\omega}{\sigma} + \frac{1 - \alpha}{\sigma}\right) - 1\right]s_t - \frac{\alpha}{1 - \alpha}c_t.$$

Using the expression for the terms-of-trade,

$$\begin{aligned} nx_t = {} & \alpha(1 - \mathcal{G})\left[\frac{1}{1 - \alpha}\left(\frac{\omega}{\sigma} + \frac{1 - \alpha}{\sigma}\right) - 1\right] \\ & \times \frac{1}{\Gamma_1}\left[[1 - (1 - \alpha)\Theta_n]c_t - (1 - \alpha)\Theta_n g_t + (1 - \alpha)\Theta_\tau \underline{t}_t^r - (1 - \alpha)\bar{\sigma}^{-1}\sigma\theta\right] \\ & - \frac{\alpha}{1 - \alpha}c_t. \end{aligned}$$

Thus,

$$\begin{aligned} nx_t = {} & \Omega_c c_t - (1 - \mathcal{G})\frac{\Gamma_2}{\Gamma_1} \\ & \times \left[[(1 - \alpha)\Theta_n - (1 - \alpha)\Theta_\tau \psi_n]g_t - (1 - \alpha)\Theta_\tau \psi_\tau t_t^r + \left[(1 - \alpha)\bar{\sigma}^{-1}\sigma - (1 - \alpha)\Theta_\tau \psi_\theta\right]\theta\right], \end{aligned}$$

where

$$\begin{aligned} \Omega_c &= (1 - \mathcal{G})\frac{\Gamma_2}{\Gamma_1}[1 - (1 - \alpha)\Theta_n + (1 - \alpha)\Theta_\tau \psi_c] - \frac{\alpha}{1 - \alpha}, \\ \Gamma_2 &= \alpha\left[\frac{1}{1 - \alpha}\left(\frac{\omega}{\sigma} + \frac{1 - \alpha}{\sigma}\right) - 1\right]. \end{aligned}$$

Using the expressions for the Pareto weight and for net exports,

$$\theta = \int_0^{+\infty} e^{-\rho s}\left[\widetilde{\boldsymbol{\Omega}}_c c_s + \widetilde{\boldsymbol{\Omega}}_n g_s + \widetilde{\boldsymbol{\Omega}}_\tau t_s^r\right] ds + \widetilde{\boldsymbol{\Omega}}_f \mathrm{nfa}_0,$$

where

$$\begin{aligned}
\widetilde{\boldsymbol{\Omega}}_c &= \rho \frac{\Gamma_1}{\Gamma_2}\frac{\boldsymbol{\Omega}_c}{1-\mathcal{G}}\frac{1}{(1-\alpha)\bar{\sigma}^{-1}\sigma - \Theta_\tau \Psi_\theta},\\
\widetilde{\boldsymbol{\Omega}}_n &= -\rho \frac{(1-\alpha)\Theta_n - (1-\alpha)\Theta_\tau \Psi_n}{(1-\alpha)\bar{\sigma}^{-1}\sigma - \Theta_\tau \Psi_\theta},\\
\widetilde{\boldsymbol{\Omega}}_\tau &= \rho \frac{(1-\alpha)\Theta_\tau \Psi_\tau}{(1-\alpha)\bar{\sigma}^{-1}\sigma - \Theta_\tau \Psi_\theta},\\
\widetilde{\boldsymbol{\Omega}}_f &= \rho \frac{\Gamma_1}{\Gamma_2}\frac{1}{1-\mathcal{G}}\frac{1}{(1-\alpha)\bar{\sigma}^{-1}\sigma - \Theta_\tau \Psi_\theta}.
\end{aligned}$$

Special case: no hand-to-mouth agents $\chi = 0$

The log-linear system is

$$\begin{aligned}
\dot{c}_t &= -\hat{\sigma}^{-1}\pi_{H,t},\\
\dot{\pi}_{H,t} &= \rho \pi_{H,t} - \kappa[c_t + (1-\xi)g_t] - (1-\mathcal{G})\lambda \hat{\sigma}\alpha\omega\theta,
\end{aligned}$$

for all $t \geq 0$, with

$$c_0 = (1-\mathcal{G})(1-\alpha)\theta$$

And

$$\theta = \int_0^{+\infty} e^{-\rho s}\rho(1-\mathcal{G})\frac{\omega - \sigma}{\omega + (1-\alpha)(1-\sigma)} c_s ds + \rho \frac{1}{\alpha}\frac{\alpha\omega + 1 - \alpha}{\omega + (1-\alpha)(1-\sigma)}\frac{1}{1-\mathcal{G}}\mathrm{nfa}_0,$$

where $\kappa = \lambda[\phi + \hat{\sigma}]$, $\xi = \dfrac{\hat{\sigma}}{\phi + \hat{\sigma}}$, $\hat{\sigma} = \dfrac{\sigma}{1-\mathcal{G}}\dfrac{1}{(1-\alpha)+\alpha\omega}$.

Special case: no profit offset $o = 0$

The log-linear system is

$$\begin{aligned}
\dot{c}_t &= -\widetilde{\sigma}^{-1}\pi_{H,t} + \widetilde{\boldsymbol{\Theta}}_n \dot{g}_t - \widetilde{\boldsymbol{\Theta}}_\tau \dot{t}_t^r,\\
\dot{\pi}_{H,t} &= \rho \pi_{H,t} - \widetilde{\kappa}\left[c_t + \left(1 - \widetilde{\xi}\right)g_t\right] - (1-\mathcal{G})\lambda \widetilde{\sigma}\widetilde{\alpha}\widetilde{\omega}\theta - \widetilde{\kappa}\widetilde{\widetilde{\boldsymbol{\Theta}}}_\tau t_t^r,
\end{aligned}$$

for all $t \geq 0$, with

$$c_0 = \frac{1}{1-\boldsymbol{\Theta}_n}[(1-\mathcal{G})(1-\widetilde{\alpha})[1-(1-\alpha)\boldsymbol{\Theta}_n]]\theta + \widetilde{\boldsymbol{\Theta}}_n g_0 - \widetilde{\boldsymbol{\Theta}}_\tau t_0^r$$

and

$$\theta = \int_0^{+\infty} e^{-\rho s} \left[\rho \frac{\Gamma_1}{\Gamma_2} \frac{(1-\mathcal{G})\frac{\Gamma_2}{\Gamma_1}[1-(1-\alpha)\Theta_n] - \frac{\alpha}{1-\alpha}}{1-\mathcal{G}} \frac{1}{(1-\alpha)\bar{\sigma}^{-1}\sigma} c_s - \rho \frac{\Theta_n}{\bar{\sigma}^{-1}\sigma} g_s + \rho \frac{\Theta_\tau}{\bar{\sigma}^{-1}\sigma} t_s^r \right] ds$$
$$+ \rho \frac{\Gamma_1}{\Gamma_2} \frac{1}{1-\mathcal{G}} \frac{1}{(1-\alpha)\bar{\sigma}^{-1}\sigma} \mathrm{nfa}_0.$$

Appendix B

This appendix derives the solutions to the linear systems obtained in Appendix A. The same special cases are considered.

B.1 Liquidity Trap

Define

$$\widetilde{\nu} = \frac{\rho - \sqrt{\rho^2 + 4\kappa \underline{\widetilde{\sigma}}^{-1}}}{2}, \quad \widetilde{\overline{\nu}} = \frac{\rho + \sqrt{\rho^2 + 4\kappa \underline{\widetilde{\sigma}}^{-1}}}{2}.$$

The equilibrium is completely characterized by the following:

$$\dot{X}_t = A X_t + B_t,$$

where

$$X_t = [\pi_t, c_t]^t, \quad A = \begin{bmatrix} \rho & -\kappa \\ -\underline{\widetilde{\sigma}}^{-1} & 0 \end{bmatrix}, \quad B_t = -\kappa(1-\xi) g_t E_1 + \left[\underline{\widetilde{\Theta}}_n \dot{g}_t - \underline{\widetilde{\Theta}}_\tau \dot{t}_t^r \right] E_2,$$

for all $t \geq 0$.

The (unique) solution that satisfies saddle-path stability writes:

$$X_t = \int_t^{+\infty} \kappa(1-\xi) g_s e^{-A(s-t)} E_1 \, ds - \int_t^{+\infty} \left(\underline{\widetilde{\Theta}}_n \dot{g}_s - \underline{\widetilde{\Theta}}_\tau \dot{t}_s^r \right) e^{-A(s-t)} E_2 \, ds.$$

Equivalently, integrating the relevant objects by part,

$$\begin{aligned} X_t = & \int_t^{+\infty} \kappa(1-\xi) g_s e^{-A(s-t)} E_1 \, ds + \left(\underline{\widetilde{\Theta}}_n g_t - \underline{\widetilde{\Theta}}_\tau t_t^r \right) E_2 \\ & - \int_t^{+\infty} \left(\underline{\widetilde{\Theta}}_n g_s - \underline{\widetilde{\Theta}}_\tau t_s^r \right) A e^{-A(s-t)} E_2 \, ds \end{aligned}.$$

Thus,

$$c_t = \int_t^{+\infty} \kappa(1-\xi) E_2^t e^{-A(s-t)} E_1 \, ds + \left(\underline{\widetilde{\Theta}}_n g_t - \underline{\widetilde{\Theta}}_\tau t_t^r \right) - \int_t^{+\infty} \left(\underline{\widetilde{\Theta}}_n g_s - \underline{\widetilde{\Theta}}_\tau t_s^r \right) E_2^t A e^{-A(s-t)} E_2 \, ds.$$

Note that

$$E_2^t e^{-At} E_1 = \widetilde{\underline{\sigma}}^{-1} \frac{e^{-\tilde{\nu} t} - e^{-\bar{\tilde{\nu}} t}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} \;,\;\; E_2^t A e^{-At} E_2 = -\kappa(1-\xi)\widetilde{\underline{\sigma}}^{-1} \frac{e^{-\tilde{\nu} t} - e^{-\bar{\tilde{\nu}} t}}{\bar{\widetilde{\nu}} - \widetilde{\nu}},$$

for all $t \geq 0$.

Thus,

$$\begin{aligned} c_t = {} & \widetilde{\underline{\Theta}}_n g_t - \widetilde{\underline{\Theta}}_\tau t_t^r \\ & + \kappa \widetilde{\underline{\sigma}}^{-1}\left(1 - \xi + \widetilde{\underline{\Theta}}_n\right) \int_t^{+\infty} \frac{e^{-\tilde{\nu}(s-t)} - e^{-\bar{\tilde{\nu}}(s-t)}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} g_s ds \\ & - \kappa(1-\xi)\widetilde{\underline{\sigma}}^{-1} \widetilde{\underline{\Theta}}_\tau \int_t^{+\infty} \frac{e^{-\tilde{\nu}(s-t)} - e^{-\bar{\tilde{\nu}}(s-t)}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} t_s^r ds. \end{aligned}$$

Therefore,

$$c_t = \widetilde{\underline{\Theta}}_n g_t - \widetilde{\underline{\Theta}}_\tau t_t^r + \int_0^{+\infty} \alpha_s^{c,HM} g_{t+s} ds - \int_0^{+\infty} \gamma_s^{c,HM} t_{t+s}^r ds,$$

where

$$\begin{aligned} \alpha_s^{c,HM} &= \left(1 + \frac{\widetilde{\underline{\Theta}}_n}{1-\xi}\right) \widetilde{\alpha}_s^{c,HM} \;,\;\; \gamma_s^{c,HM} = \frac{\widetilde{\underline{\Theta}}_\tau}{1-\xi} \widetilde{\alpha}_s^{c,HM}, \\ \widetilde{\alpha}_s^{c,HM} &= \kappa \widetilde{\underline{\sigma}}^{-1}(1-\xi) e^{-\bar{\tilde{\nu}} s} \frac{e^{(\bar{\tilde{\nu}} - \tilde{\nu})s} - 1}{\bar{\widetilde{\nu}} - \widetilde{\nu}}. \end{aligned}$$

Special case: no hand-to-mouth agents $\chi = 0$

Define

$$\nu = \frac{\rho - \sqrt{\rho^2 + 4\kappa\hat{\sigma}^{-1}}}{2} \;,\;\; \bar{\nu} = \frac{\rho + \sqrt{\rho^2 + 4\kappa\hat{\sigma}^{-1}}}{2}.$$

We have

$$c_t = \int_0^{+\infty} \alpha_s^{c,HM} g_{t+s} ds,$$

where

$$\alpha_s^{c,HM} = \kappa\hat{\sigma}^{-1}(1-\xi) e^{-\bar{\nu} s} \frac{e^{(\bar{\nu} - \nu)s} - 1}{\bar{\nu} - \nu}.$$

Special case: no profit offset $o = 0$

We have

$$c_t = \widetilde{\Theta}_n g_t - \widetilde{\Theta}_\tau t_t^r + \int_0^{+\infty} \alpha_s^{c,HM} g_{t+s} ds - \int_0^{+\infty} \gamma_s^{c,HM} t_{t+s}^r ds,$$

where

$$\alpha_s^{c,HM} = \left(1 + \frac{\widetilde{\Theta}_n}{1-\xi}\right)\widetilde{\alpha}_s^{c,HM} \ , \ \ \gamma_s^{c,HM} = \frac{\widetilde{\Theta}_\tau}{1-\xi}\widetilde{\alpha}_s^{c,HM},$$

$$\widetilde{\alpha}_s^{c,HM} = \kappa\widetilde{\sigma}^{-1}(1-\xi)e^{-\bar{\widetilde{\nu}}s}\frac{e^{(\bar{\widetilde{\nu}}-\widetilde{\nu})s}-1}{\bar{\widetilde{\nu}}-\widetilde{\nu}}.$$

B.2 Currency Union

The IM and PF cases are considered here. The results for the CM case are obtained by direct analogy.

Define

$$\widetilde{\nu} = \frac{\rho - \sqrt{\rho^2 + 4\widetilde{\kappa}_c\underline{\widetilde{\sigma}}^{-1}}}{2} \ , \ \ \bar{\widetilde{\nu}} = \frac{\rho + \sqrt{\rho^2 + 4\widetilde{\kappa}_c\underline{\widetilde{\sigma}}^{-1}}}{2}.$$

The equilibrium is completely characterized by the following:

$$\dot{X}_t = AX_t + B_t,$$

With

$$E_2^t X_0 = \Upsilon\theta + \underline{\widetilde{\Theta}}_n g_0 - \underline{\widetilde{\Theta}}_\tau t_0^r,$$

$$\theta = \int_0^{+\infty} e^{-\rho s}\left[\underline{\widetilde{\Omega}}_c c_t + \underline{\widetilde{\Omega}}_n g_t + \underline{\widetilde{\Omega}}_\tau t_t^r\right] ds + \underline{\widetilde{\Omega}}_f \mathrm{nfa}_0,$$

where

$$X_t = [\pi_t, c_t]^t \ , \ \ A = \begin{bmatrix} \rho & -\widetilde{\kappa}_c \\ -\underline{\widetilde{\sigma}}^{-1} & 0 \end{bmatrix} \ , \ \ B_t = -\left(\underline{\widetilde{\kappa}}_n g_t + \underline{\widetilde{\kappa}}_\theta \theta + \underline{\widetilde{\kappa}}_\tau t_t^r\right)E_1 + \left[\underline{\widetilde{\Theta}}_n \dot{g}_t - \underline{\widetilde{\Theta}}_\tau \dot{t}_t^r\right]E_2,$$

for all $t \geq 0$.

The (unique) solution that satisfies saddle-path stability writes:

$$X_t = \alpha_{\widetilde{\nu}} e^{\widetilde{\nu}t} X_{\widetilde{\nu}} + \int_t^{+\infty}\left(\underline{\widetilde{\kappa}}_n g_s + \underline{\widetilde{\kappa}}_\theta \theta + \underline{\widetilde{\kappa}}_\tau t_s^r\right)e^{-A(s-t)}E_1 ds - \int_t^{+\infty}\left(\underline{\widetilde{\Theta}}_n \dot{g}_s - \underline{\widetilde{\Theta}}_\tau \dot{t}_s^r\right)e^{-A(s-t)}E_2 ds,$$

with

$$E_2^t X_0 = \Upsilon\theta + \underline{\widetilde{\Theta}}_n g_0 - \underline{\widetilde{\Theta}}_\tau t_0^r,$$
$$\theta = \int_0^{+\infty} e^{-\rho s}\left[\widetilde{\Omega}_c c_t + \widetilde{\Omega}_n g_t + \widetilde{\Omega}_\tau t_t^r\right] ds + \widetilde{\Omega}_f \mathrm{nfa}_0,$$

where $\alpha_{\tilde{\nu}} \in \mathbb{R}$.

Equivalently, integrating the relevant objects by part,

$$X_t = \alpha_{\tilde{\nu}} e^{\tilde{\nu} t} X_{\tilde{\nu}} + \int_t^{+\infty} \left(\underline{\widetilde{\kappa}}_n g_s + \underline{\widetilde{\kappa}}_\theta \theta + \underline{\widetilde{\kappa}}_\tau t_s^r\right) e^{-A(s-t)} E_1 ds + \left(\underline{\widetilde{\Theta}}_n g_t - \underline{\widetilde{\Theta}}_\tau t_t^r\right) E_2$$
$$- \int_t^{+\infty} \left(\underline{\widetilde{\Theta}}_n g_s - \underline{\widetilde{\Theta}}_\tau t_s^r\right) A e^{-A(s-t)} E_2 ds,$$

with

$$E_2^t X_0 = \Upsilon\theta + \underline{\widetilde{\Theta}}_n g_0 - \underline{\widetilde{\Theta}}_\tau t_0^r,$$
$$\theta = \int_0^{+\infty} e^{-\rho s}\left[\widetilde{\Omega}_c c_t + \widetilde{\Omega}_n g_t + \widetilde{\Omega}_\tau t_t^r\right] ds + \widetilde{\Omega}_f \mathrm{nfa}_0.$$

Thus,

$$\Upsilon\theta - \int_0^{+\infty} \left(\underline{\widetilde{\kappa}}_n g_s + \underline{\widetilde{\kappa}}_\theta \theta + \underline{\widetilde{\kappa}}_\tau t_s^r\right) E_2^t e^{-As} E_1 ds$$
$$+ \int_0^{+\infty} \left(\underline{\widetilde{\Theta}}_n g_s - \underline{\widetilde{\Theta}}_\tau t_s^r\right) A E_2^t e^{-As} E_2 ds = \alpha_{\tilde{\nu}}.$$

Therefore,

$$c_t = \left[\Upsilon\theta - \int_0^{+\infty} \left(\underline{\widetilde{\kappa}}_n g_s + \underline{\widetilde{\kappa}}_\theta \theta + \underline{\widetilde{\kappa}}_\tau t_s^r\right) E_2^t e^{-As} E_1 ds + \int_0^{+\infty} \left(\underline{\widetilde{\Theta}}_n g_s - \underline{\widetilde{\Theta}}_\tau t_s^r\right) E_2^t A e^{-As} E_2 ds\right] e^{\tilde{\nu} t}$$
$$+ \int_t^{+\infty} \left(\underline{\widetilde{\kappa}}_n g_s + \underline{\widetilde{\kappa}}_\theta \theta + \underline{\widetilde{\kappa}}_\tau t_s^r\right) E_2^t e^{-A(s-t)} E_1 ds + \left(\underline{\widetilde{\Theta}}_n g_t - \underline{\widetilde{\Theta}}_\tau t_t^r\right)$$
$$- \int_t^{+\infty} \left(\underline{\widetilde{\Theta}}_n g_s - \underline{\widetilde{\Theta}}_\tau t_s^r\right) E_2^t A e^{-A(s-t)} E_2 ds.$$

Equivalently,

$$c_t = \left[\Upsilon e^{\tilde{\nu} t} - \underline{\widetilde{\kappa}}_\theta \left[e^{\tilde{\nu} t} \int_0^{+\infty} E_2^t e^{-As} E_1 ds - \int_t^{+\infty} E_2^t e^{-A(s-t)} E_1 ds\right]\right] \theta + \underline{\widetilde{\Theta}}_n g_t - \underline{\widetilde{\Theta}}_\tau t_t^r$$
$$- \underline{\widetilde{\kappa}}_n \left[e^{\tilde{\nu} t} \int_0^{+\infty} E_2^t e^{-As} E_1 g_s ds - \int_t^{+\infty} E_2^t e^{-A(s-t)} E_1 g_s ds\right]$$
$$+ \underline{\widetilde{\Theta}}_n \left[e^{\tilde{\nu} t} \int_0^{+\infty} E_2^t A e^{-As} E_2 g_s ds - \int_t^{+\infty} E_2^t A e^{-A(s-t)} E_2 g_s ds\right]$$
$$- \underline{\widetilde{\kappa}}_\tau \left[e^{\tilde{\nu} t} \int_0^{+\infty} E_2^t e^{-As} E_1 t_s^r ds - \int_t^{+\infty} E_2^t e^{-A(s-t)} E_1 t_s^r ds\right]$$
$$- \underline{\widetilde{\Theta}}_\tau \left[e^{\tilde{\nu} t} \int_0^{+\infty} E_2^t e^{-As} A E_2 t_s^r ds - \int_t^{+\infty} E_2^t A e^{-A(s-t)} E_2 t_s^r ds\right].$$

Note that

$$E_2^t e^{-At} E_1 = \underline{\widetilde{\sigma}}^{-1} \frac{e^{-\tilde{\nu}t} - e^{-\bar{\tilde{\nu}}t}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} \ , \quad E_2^t A e^{-At} E_2 = -\widetilde{\kappa}_c \underline{\widetilde{\sigma}}^{-1} \frac{e^{-\tilde{\nu}t} - e^{-\bar{\tilde{\nu}}t}}{\bar{\widetilde{\nu}} - \widetilde{\nu}},$$

for all $t \geq 0$.

Thus,

$$c_t = \left[\frac{\Upsilon e^{\tilde{\nu}t} - \underline{\widetilde{\kappa}_\theta} \underline{\widetilde{\sigma}}^{-1} \left(e^{\tilde{\nu}t} - 1 \right) \int_0^{+\infty} e^{-\tilde{\nu}s} - e^{-\bar{\tilde{\nu}}} s}{\bar{\widetilde{\nu}} - \widetilde{\nu} ds} \right] \theta + \underline{\widetilde{\Theta}}_n g_t - \underline{\widetilde{\Theta}}_\tau t_t^r$$

$$- \left(\underline{\widetilde{\kappa}}_n + \underline{\widetilde{\kappa}}_c \underline{\widetilde{\Theta}}_n \right) \underline{\widetilde{\sigma}}^{-1} \left[e^{\tilde{\nu}t} \int_0^{+\infty} \frac{e^{-\tilde{\nu}s} - e^{-\bar{\tilde{\nu}}s}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} g_s ds - \int_t^{+\infty} \frac{e^{-\tilde{\nu}(s-t)} - e^{-\bar{\tilde{\nu}}(s-t)}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} g_s ds \right]$$

$$- \left(\underline{\widetilde{\kappa}}_\tau - \underline{\widetilde{\kappa}}_c \underline{\widetilde{\Theta}}_\tau \right) \underline{\widetilde{\sigma}}^{-1} \left[e^{\tilde{\nu}t} \int_0^{+\infty} \frac{e^{-\tilde{\nu}s} - e^{-\bar{\tilde{\nu}}s}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} t_s^r ds - \int_t^{+\infty} \frac{e^{-\tilde{\nu}(s-t)} - e^{-\bar{\tilde{\nu}}(s-t)}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} t_s^r ds \right].$$

Using the expression for the Pareto weight θ,

$$c_t = \left[\Upsilon e^{\tilde{\nu}t} - \underline{\widetilde{\kappa}_\theta} \underline{\widetilde{\sigma}}^{-1} \left(e^{\tilde{\nu}t} - 1 \right) \frac{\widetilde{\nu}^{-1} - \bar{\widetilde{\nu}}^{-1}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} \right] \left(\int_0^{+\infty} e^{-\rho s} \left[\widetilde{\Omega}_c c_s + \widetilde{\Omega}_n g_s + \widetilde{\Omega}_\tau t_s^r \right] ds + \widetilde{\Omega}_f \mathrm{nfa}_0 \right)$$

$$+ \underline{\widetilde{\Theta}}_n g_t - \underline{\widetilde{\Theta}}_\tau t_t^r$$

$$- \left(\underline{\widetilde{\kappa}}_n + \underline{\widetilde{\kappa}}_c \underline{\widetilde{\Theta}}_n \right) \underline{\widetilde{\sigma}}^{-1} \left[e^{\widetilde{\nu t}} \int_0^{+\infty} \frac{e^{-\tilde{\nu}s} - e^{-\bar{\tilde{\nu}}s}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} g_s ds - \int_t^{+\infty} \frac{e^{-\tilde{\nu}(s-t)} - e^{-\bar{\tilde{\nu}}(s-t)}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} g_s ds \right]$$

$$- \left(\underline{\widetilde{\kappa}}_\tau - \underline{\widetilde{\kappa}}_c \underline{\widetilde{\Theta}}_\tau \right) \underline{\widetilde{\sigma}}^{-1} \left[e^{\tilde{\nu}t} \int_0^{+\infty} \frac{e^{-\tilde{\nu}s} - e^{-\bar{\tilde{\nu}}s}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} t_s^r ds - \int_t^{+\infty} \frac{e^{-\tilde{\nu}(s-t)} - e^{-\bar{\tilde{\nu}}(s-t)}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} t_s^r ds \right].$$

From Fubini's Theorem, assuming that the integrals are finite,

$$\int_0^{+\infty} e^{-\rho t} \int_t^{+\infty} \frac{e^{-\tilde{\nu}(s-t)} - e^{-\bar{\tilde{\nu}}(s-t)}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} x_s ds dt = \int_0^{+\infty} e^{-\rho s} \int_0^s \frac{e^{(\rho - \tilde{\nu})(s-t)} - e^{(\rho - \bar{\tilde{\nu}})(s-t)}}{\bar{\widetilde{\nu}} - \widetilde{\nu}} dt x_s ds$$

$$= -\frac{1}{\bar{\widetilde{\nu}}} - \widetilde{\nu} \int_0^{+\infty} e^{-\rho s} \left[(\rho - \widetilde{\nu})^{-1} \left(1 - e^{(\rho - \tilde{\nu})s} \right) - \left(\rho - \bar{\widetilde{\nu}} \right)^{-1} \left(1 - e^{(\rho - \tilde{\nu})s} \right) \right] x_s ds,$$

for each $x \in \{g, t^r\}$.

Note that $\frac{\tilde{\nu}^{-1} - \bar{\tilde{\nu}}^{-1}}{\bar{\tilde{\nu}} - \tilde{\nu}} = -\underline{\widetilde{\kappa}}_c^{-1} \underline{\widetilde{\sigma}}$ by definition of $\widetilde{\nu}, \bar{\widetilde{\nu}}$. Thus,

$$\int_0^{+\infty} e^{-\rho t} c_t dt = \frac{1}{1 - \Sigma \widetilde{\Omega}_c} \left(\int_0^{+\infty} \varsigma_n^t g_s ds + \int_0^{+\infty} \varsigma_\tau^t t_s^r ds + \varsigma_f \mathrm{nfa}_0 \right),$$

where

$$\begin{aligned}
\varsigma_n^t &= \Sigma e^{-\rho t}\widetilde{\Omega}_n + e^{-\rho t}\underline{\widetilde{\Theta}}_n - \left(\underline{\widetilde{\kappa}}_n + \underline{\widetilde{\kappa}}_c\underline{\widetilde{\Theta}}_n\right)\underline{\widetilde{\sigma}}^{-1}\left[\frac{1}{\rho-\widetilde{\nu}}\frac{e^{-\tilde{\nu}t}-e^{-\bar{\bar{\tilde{\nu}}}t}}{\bar{\bar{\widetilde{\nu}}}-\widetilde{\nu}}\right.\\
&\qquad\left.+\frac{1}{\bar{\bar{\widetilde{\nu}}}-\widetilde{\nu}}e^{-\rho t}\left[(\rho-\widetilde{\nu})^{-1}\left(1-e^{(\rho-\tilde{\nu})t}\right)-\left(\rho-\bar{\bar{\widetilde{\nu}}}\right)^{-1}\left(1-e^{(\rho-\bar{\bar{\tilde{\nu}}})t}\right)\right]\right],\\
\varsigma_\tau^t &= \Sigma e^{-\rho t}\widetilde{\Omega}_\tau - e^{-\rho t}\underline{\widetilde{\Theta}}_\tau - \left(\underline{\widetilde{\kappa}}_\tau - \underline{\widetilde{\kappa}}_c\underline{\widetilde{\Theta}}_\tau\right)\underline{\widetilde{\sigma}}^{-1}\left[\frac{1}{\rho-\widetilde{\nu}}\frac{e^{-\tilde{\nu}t}-e^{-\bar{\bar{\tilde{\nu}}}t}}{\bar{\bar{\widetilde{\nu}}}-\widetilde{\nu}}\right.\\
&\qquad\left.+\frac{1}{\bar{\bar{\widetilde{\nu}}}-\widetilde{\nu}}e^{-\rho t}\left[(\rho-\widetilde{\nu})^{-1}\left(1-e^{-(\rho-\tilde{\nu})t}\right)-\left(\rho-\bar{\bar{\widetilde{\nu}}}\right)^{-1}\left(1-e^{(\rho-\bar{\bar{\tilde{\nu}}})t}\right)\right]\right],\\
\varsigma_f &= \Sigma\widetilde{\Omega}_f,\\
\Sigma &= \Upsilon\frac{1}{\rho-\widetilde{\nu}} - \underline{\widetilde{\kappa}}_\theta\underline{\widetilde{\kappa}}_c^{-1}\left(\frac{1}{\rho}-\frac{1}{\rho-\widetilde{\nu}}\right).
\end{aligned}$$

Therefore,

$$c_t = \underline{\widetilde{\Theta}}_n g_t - \underline{\widetilde{\Theta}}_\tau t_t^r + \int_{-t}^{+\infty}\alpha_s^{c,t,HM,IM} g_{t+s}ds - \int_{-t}^{+\infty}\gamma_s^{c,t,HM,IM} t_{t+s}^r ds,$$

where

$$\begin{aligned}
\alpha_s^{c,t,HM,IM} &= \alpha_s^{c,t,HM,CM} + \delta_s^{c,t,HM,IM} + \delta_s^{c,t,HM,PF},\\
\gamma_s^{c,t,HM,IM} &= \gamma_{s0}^{c,t,HM,CM} + \epsilon_s^{c,t,HM,IM},
\end{aligned}$$

with

$$\begin{aligned}
\alpha_s^{c,t,HM,CM} &= -\left(\underline{\widetilde{\kappa}}_n + \underline{\widetilde{\kappa}}_c\underline{\widetilde{\Theta}}_n\right)\underline{\widetilde{\sigma}}^{-1}\left[e^{\widetilde{\nu t}}\frac{e^{-\tilde{\nu}(t+s)}-e^{-\bar{\bar{\tilde{\nu}}}(t+s)}}{\bar{\bar{\widetilde{\nu}}}-\widetilde{\nu}} - \mathbb{1}_{s\geq 0}\frac{e^{-\tilde{\nu}s}-e^{-\bar{\bar{\tilde{\nu}}}s}}{\bar{\bar{\widetilde{\nu}}}-\widetilde{\nu}}\right],\\
\gamma_s^{c,t,HM,CM} &= \left(\underline{\widetilde{\kappa}}_\tau - \underline{\widetilde{\kappa}}_c\underline{\widetilde{\Theta}}_\tau\right)\underline{\widetilde{\sigma}}^{-1}\left[e^{\tilde{\nu}t}\frac{e^{-\tilde{\nu}(t+s)}-e^{-\bar{\bar{\tilde{\nu}}}(t+s)}}{\bar{\bar{\widetilde{\nu}}}-\widetilde{\nu}} - \mathbb{1}_{s\geq 0}\frac{e^{-\tilde{\nu}s}-e^{-\bar{\bar{\tilde{\nu}}}s}}{\bar{\bar{\widetilde{\nu}}}-\widetilde{\nu}}\right],\\
\delta_s^{c,t,HM,IM} &= \left(\frac{1}{1-\Sigma\widetilde{\Omega}_c}\widetilde{\Omega}_c\varsigma_n^{t+s} + e^{-\rho(t+s)}\widetilde{\Omega}_n\right)\left[\Upsilon e^{\tilde{\nu}t} - \underline{\widetilde{\kappa}}_\theta\underline{\widetilde{\kappa}}_c^{-1}\left(1-e^{\tilde{\nu}t}\right)\right],\\
\epsilon_s^{c,t,HM,IM} &= -\left(\frac{1}{1-\Sigma\widetilde{\Omega}_c}\widetilde{\Omega}_c\varsigma_\tau^{t+s} + e^{-\rho(t+s)}\widetilde{\Omega}_\tau\right)\left[\Upsilon e^{\tilde{\nu}t} - \underline{\widetilde{\kappa}}_\theta\underline{\widetilde{\kappa}}_c^{-1}\left(1-e^{\tilde{\nu}t}\right)\right],
\end{aligned}$$

and $\delta_s^{c,t,HM,PF} = 0$ in IM case, and

$$\delta_s^{c,t,HM,PF} = \left(\frac{1}{1-\Sigma\widetilde{\Omega}_c}\widetilde{\Omega}_c\varsigma_f + \widetilde{\Omega}_f\right)\left[\Upsilon e^{\tilde{\nu}t} - \underline{\widetilde{\kappa}}_\theta\underline{\widetilde{\kappa}}_c^{-1}\left(1-e^{\tilde{\nu}t}\right)\right]e^{-\rho(t+s)}$$

in PF case.

We can reexpress these as

$$\alpha_s^{c,t,HM,CM} = -\left(\underline{\widetilde{\kappa}}_n + \underline{\widetilde{\kappa}}_c \underline{\widetilde{\Theta}}_n\right)\underline{\widetilde{\sigma}}^{-1}\left[e^{\tilde{\nu}t}\frac{e^{-\tilde{\nu}(t+s)} - e^{-\bar{\bar{\tilde{\nu}}}(t+s)}}{\bar{\bar{\widetilde{\nu}}} - \widetilde{\nu}} - \mathbb{1}_{s\geq 0}\frac{e^{-\tilde{\nu}s} - e^{-\bar{\bar{\tilde{\nu}}}s}}{\bar{\bar{\widetilde{\nu}}} - \widetilde{\nu}}\right],$$

$$\gamma_s^{c,t,HM,CM} = \left(\underline{\widetilde{\kappa}}_\tau - \underline{\widetilde{\kappa}}_c \underline{\widetilde{\Theta}}_\tau\right)\underline{\widetilde{\sigma}}^{-1}\left[e^{\tilde{\nu}t}\frac{e^{-\tilde{\nu}(t+s)} - e^{-\bar{\bar{\tilde{\nu}}}(t+s)}}{\bar{\bar{\widetilde{\nu}}} - \widetilde{\nu}} - \mathbb{1}_{s\geq 0}\frac{e^{-\tilde{\nu}s} - e^{-\bar{\bar{\tilde{\nu}}}s}}{\bar{\bar{\widetilde{\nu}}} - \widetilde{\nu}}\right],$$

$$\begin{aligned}\delta_s^{c,t,HM,IM} = {} & \left[\Upsilon e^{\tilde{\nu}t} - \underline{\widetilde{\kappa}}_\theta \underline{\widetilde{\kappa}}_c^{-1}\left(1 - e^{\tilde{\nu}t}\right)\right] \times \\ & \frac{1}{1 - \Sigma \underline{\widetilde{\Omega}}_c}\Big[e^{-\rho(t+s)}\underline{\widetilde{\Omega}}_n + e^{-\rho(t+s)}\underline{\widetilde{\Omega}}_c \underline{\widetilde{\Theta}}_n \\ & + \underline{\widetilde{\Omega}}_c\left(\underline{\widetilde{\kappa}}_n + \underline{\widetilde{\kappa}}_c \underline{\widetilde{\Theta}}_n\right)\underline{\widetilde{\sigma}}^{-1}\frac{1}{\widetilde{\kappa}_c \underline{\widetilde{\sigma}}^{-1}} e^{-\rho(t+s)}\left(1 - e^{\tilde{\nu}(t+s)}\right)\Big],\end{aligned}$$

$$\begin{aligned}\epsilon_s^{c,t,HM,IM} = {} & -\left[\Upsilon e^{\tilde{\nu}t} - \underline{\widetilde{\kappa}}_\theta \underline{\widetilde{\kappa}}_c^{-1}\left(1 - e^{\tilde{\nu}t}\right)\right] \times \\ & \frac{1}{1 - \Sigma \underline{\widetilde{\Omega}}_c}\Big[e^{-\rho(t+s)}\underline{\widetilde{\Omega}}_\tau - e^{-\rho(t+s)}\underline{\widetilde{\Omega}}_c \underline{\widetilde{\Theta}}_\tau \\ & + \underline{\widetilde{\Omega}}_c\left(\underline{\widetilde{\kappa}}_\tau - \underline{\widetilde{\kappa}}_c \underline{\widetilde{\Theta}}_\tau\right)\underline{\widetilde{\sigma}}^{-1}\frac{1}{\widetilde{\kappa}_c \underline{\widetilde{\sigma}}^{-1}} e^{-\rho(t+s)}\left(1 - e^{\tilde{\nu}(t+s)}\right)\Big],\end{aligned}$$

and $\delta_s^{c,t,HM,PF} = 0$ in the IM case, and

$$\delta_s^{c,t,HM,PF} = \left[\Upsilon e^{\tilde{\nu}t} - \underline{\widetilde{\kappa}}_\theta \underline{\widetilde{\kappa}}_c^{-1}\left(1 - e^{\widetilde{\nu t}}\right)\right]\frac{1}{1 - \Sigma \underline{\widetilde{\Omega}}_c}\underline{\widetilde{\Omega}}_f e^{-\rho(t+s)}$$

in the PF case.

By direct analogy,

$$c_t = \underline{\widetilde{\Theta}}_n g_t - \underline{\widetilde{\Theta}}_\tau t_t^r + \int_{-t}^{+\infty} \alpha_s^{c,t,HM,CM} g_{t+s} ds - \int_0^{+\infty} \gamma_s^{c,t,HM,CM} t_{t+s}^r ds$$

in CM case.

Special case: no hand-to-mouth agents $\chi = 0$

Define

$$\nu = \frac{\rho - \sqrt{\rho^2 + 4\kappa\hat{\sigma}^{-1}}}{2}, \quad \bar{\nu} = \frac{\rho + \sqrt{\rho^2 + 4\kappa\hat{\sigma}^{-1}}}{2},$$

and

$$\hat{\Sigma} = (1 - \mathcal{G})(1 - \alpha)\frac{1}{\bar{\bar{\widetilde{\nu}}}} + (1 - \mathcal{G})\frac{\hat{\sigma}}{\phi + \hat{\sigma}}\alpha\omega\frac{1}{\bar{\bar{\widetilde{\nu}}}}\frac{\widetilde{\nu}}{\rho}.$$

Define

$$\alpha_s^{c,t,HM,CM} = -\kappa(1-\xi)\hat{\sigma}^{-1}\left[e^{\nu t}\frac{e^{-\nu(t+s)} - e^{-\bar{\nu}(t+s)}}{\bar{\nu}-\nu} - \mathbb{1}_{s\geq 0}\frac{e^{-\nu s} - e^{-\bar{\nu}s}}{\bar{\nu}-\nu}\right].$$

We have

$$c_t = \int_{-t}^{+\infty} \alpha_s^{c,t,HM,IM} g_{t+s} ds,$$

where

$$\alpha_s^{c,t,HM,IM} = \alpha_s^{c,t,HM,CM} + \delta_s^{c,t,HM,IM} + \delta_s^{c,t,HM,PF},$$

with

$$\delta_s^{c,t,HM,IM} = \rho\left[\frac{1-\alpha}{\alpha}e^{\tilde{\nu}t} - \frac{\hat{\sigma}}{\phi+\hat{\sigma}}\omega\left(1-e^{\tilde{\nu}t}\right)\right] \\ \times \frac{\alpha\dfrac{\omega-\sigma}{\omega+(1-\alpha)(1-\sigma)}}{1-\hat{\Sigma}\rho\dfrac{\omega-\sigma}{\omega+(1-\alpha)(1-\sigma)}\dfrac{1}{1-\mathcal{G}}}(1-\xi)e^{-\rho(t+s)}\left(1-e^{\tilde{\nu}(t+s)}\right),$$

and $\delta_s^{c,t,HM,PF} = 0$ in the IM case, and

$$\delta_s^{c,t,HM,PF} = \rho\left[\frac{1-\alpha}{\alpha}e^{\tilde{\nu}t} - \frac{\hat{\sigma}}{\phi+\hat{\sigma}}\omega\left(1-e^{\widetilde{\nu t}}\right)\right]\frac{\dfrac{\alpha\omega+1-\alpha}{\omega+(1-\alpha)(1-\sigma)}}{1-\hat{\Sigma}\rho\dfrac{\omega-\sigma}{\omega+(1-\alpha)(1-\sigma)}\dfrac{1}{1-\mathcal{G}}}e^{-\rho(t+s)}$$

in the PF case.

Special case: no profit offset $o = 0$

In that case we have

$$\Sigma = (1-\mathcal{G})(1-\widetilde{\alpha})\frac{1}{\bar{\bar{\widetilde{\nu}}}} + (1-\mathcal{G})\lambda\widetilde{\sigma}\widetilde{\alpha}\widetilde{\omega}\widetilde{\kappa}^{-1}\frac{1}{\rho}\frac{\widetilde{\nu}}{\bar{\bar{\widetilde{\nu}}}}.$$

Define

$$\widetilde{\alpha}_s^{c,t,HM,CM} = -\widetilde{\kappa}\left(1-\widetilde{\xi}\right)\widetilde{\sigma}^{-1}\left[e^{\tilde{\nu}t}\frac{e^{-\tilde{\nu}(t+s)} - e^{-\bar{\bar{\tilde{\nu}}}(t+s)}}{\bar{\bar{\widetilde{\nu}}}-\widetilde{\nu}} - \mathbb{1}_{s\geq 0}\frac{e^{-\tilde{\nu}s} - e^{-\bar{\bar{\tilde{\nu}}}s}}{\bar{\bar{\widetilde{\nu}}}-\widetilde{\nu}}\right].$$

We have

$$c_t = \widetilde{\Theta}_n g_t - \widetilde{\Theta}_\tau t_t^r + \int_{-t}^{+\infty} \alpha_s^{c,t,HM,IM} g_{t+s} ds - \int_{-t}^{+\infty} \gamma_s^{c,t,HM,IM} t_{t+s}^r ds,$$

where

$$\alpha_s^{c,t,HM,IM} = \alpha_s^{c,t,HM,CM} + \delta_s^{c,t,HM,IM} + \delta_s^{c,t,HM,PF},$$
$$\gamma_s^{c,t,HM,IM} = \gamma_s^{c,t,HM,CM} + \epsilon_s^{c,t,HM,IM},$$

with

$$\alpha_s^{c,t,HM,CM} = -\left(1 + \frac{\widetilde{\Theta}_n}{1-\widetilde{\xi}}\right)\widetilde{\kappa}\left(1-\widetilde{\xi}\right)\widetilde{\sigma}^{-1}\left[e^{\tilde{\nu}t}\frac{e^{-\tilde{\nu}(t+s)} - e^{-\bar{\tilde{\nu}}(t+s)}}{\bar{\tilde{\nu}} - \widetilde{\nu}} - \mathbb{1}_{s\geq 0}\frac{e^{-\tilde{\nu}s} - e^{-\bar{\tilde{\nu}}s}}{\bar{\tilde{\nu}} - \widetilde{\nu}}\right],$$

$$\gamma_s^{c,t,HM,CM} = -\frac{\widetilde{\Theta}_\tau - \widetilde{\widetilde{\Theta}}_\tau}{1-\widetilde{\xi}}\widetilde{\kappa}\left(1-\widetilde{\xi}\right)\widetilde{\sigma}^{-1}\left[e^{\tilde{\nu}t}\frac{e^{-\tilde{\nu}(t+s)} - e^{-\bar{\tilde{\nu}}(t+s)}}{\bar{\tilde{\nu}} - \widetilde{\nu}} - \mathbb{1}_{s\geq 0}\frac{e^{-\tilde{\nu}s} - e^{-\bar{\tilde{\nu}}s}}{\bar{\tilde{\nu}} - \widetilde{\nu}}\right],$$

$$\begin{aligned}
\delta_s^{c,t,HM,IM} = \rho &\left[\frac{1-\widetilde{\alpha}}{\widetilde{\alpha}}e^{\tilde{\nu}t} - \lambda\widetilde{\sigma}\widetilde{\kappa}^{-1}\widetilde{\omega}\left(1-e^{\tilde{\nu}t}\right)\right] \\
&\times \frac{\widetilde{\alpha}}{1-\Sigma\widetilde{\Omega}_c}\left[e^{-\rho(t+s)}\frac{(1-\mathcal{G})\widetilde{\Omega}_n}{\rho} + e^{-\rho(t+s)}\frac{(1-\mathcal{G})\widetilde{\Omega}_c}{\rho}\widetilde{\Theta}_n. \right. \\
&\left. + \frac{(1-\mathcal{G})\widetilde{\Omega}_c}{\rho}\left(1-\widetilde{\xi}\right)\left(1+\frac{\widetilde{\Theta}_n}{1-\widetilde{\xi}}\right)e^{-\rho(t+s)}\left(1-e^{\tilde{\nu}(t+s)}\right)\right],
\end{aligned}$$

$$\begin{aligned}
\epsilon_s^{c,t,HM,IM} = -\rho &\left[\frac{1-\widetilde{\alpha}}{\widetilde{\alpha}}e^{\tilde{\nu}t} - \lambda\widetilde{\sigma}\widetilde{\kappa}^{-1}\widetilde{\omega}\left(1-e^{\tilde{\nu}t}\right)\right] \\
&\times \frac{\widetilde{\alpha}}{1-\Sigma\widetilde{\Omega}_c}\left[e^{-\rho(t+s)}\frac{(1-\mathcal{G})\widetilde{\Omega}_\tau}{\rho} - e^{-\rho(t+s)}\frac{(1-\mathcal{G})\widetilde{\Omega}_c}{\rho}\widetilde{\Theta}_\tau. \right. \\
&\left. + \widetilde{\Omega}_c\left(1-\widetilde{\xi}\right)\frac{\widetilde{\widetilde{\Theta}}_\tau - \widetilde{\Theta}_\tau}{1-\widetilde{\xi}}e^{-\rho(t+s)}\left(1-e^{\tilde{\nu}(t+s)}\right)\right],
\end{aligned}$$

and $\delta_s^{c,t,HM,PF} = 0$ in the IM case, and

$$\delta_s^{c,t,HM,PF} = \rho\left[\frac{1-\widetilde{\alpha}}{\widetilde{\alpha}}e^{\tilde{\nu}t} - \lambda\widetilde{\sigma}\widetilde{\kappa}^{-1}\widetilde{\omega}\left(1-e^{\tilde{\nu}t}\right)\right]\frac{1}{1-\Sigma\widetilde{\Omega}_c}\frac{\widetilde{\alpha}(1-\mathcal{G})\widetilde{\Omega}_f}{\rho}e^{-\rho(t+s)}$$

in the PF case.

REFERENCES

Acconcia, A., Corsetti, G., Simonelli, S., 2011. Mafia and public spending: evidence on the fiscal multiplier from a quasi-experiment. CEPR Discussion Papers.

Auerbach, A., Gorodnichenko, Y., 2012. Fiscal multipliers in recession and expansion. In: Fiscal Policy After the Financial Crisis, NBER Chapters. National Bureau of Economic Research, Inc. NBER working paper #17447.

Barro, R.J., Redlick, C.J., 2009. Macroeconomic effects from government purchases and taxes. NBER Working Papers No. 15369, National Bureau of Economic Research, Inc.

Campbell, J.Y., Mankiw, N.G., 1989. Consumption, income and interest rates: reinterpreting the time series evidence. In: Blanchard, O., Fischer, S. (Eds.), NBER Macroeconomics Annual 1989, NBER Chapters, vol. 4. National Bureau of Economic Research, Inc, pp. 185–246.

Chodorow-Reich, G., Feiveson, L., Liscow, Z., Woolston, W., 2011. Does state fiscal relief during recessions increase employment? Evidence from the American Recovery and Reinvestment Act. Working Paper, University of California at Berkeley.

Christiano, L., Eichenbaum, M., Rebelo, S., 2011. When is the government spending multiplier large? J. Polit. Econ. 119 (1), 78–121.

Clemens, J., Miran, S., 2010. The effects of state budget cuts on employment and income. Working Paper, Harvard University.

Cohen, L., Coval, J.D., Malloy, C., 2010. Do powerful politicians cause corporate downsizing? NBER Working Papers, National Bureau of Economic Research, Inc.

Cole, H.L., Obstfeld, M., 1991. Commodity trade and international risk sharing: how much do financial markets matter? J. Monet. Econ. 28 (1), 3–24.

Cook, D., Devereux, M.B., 2011. Optimal fiscal policy in a world liquidity trap. Eur. Econ. Rev. 55 (4), 443–462.

Corsetti, G., Kuester, K., Muller, G.J., 2011. Floats, pegs and the transmission of fiscal policy. J. Econ. (Chin.) 14 (2), 5–38.

Dornbusch, R., 1980. Exchange rate economics: where do we stand? Brook. Pap. Econ. Act. 11 (1), 143–206.

Eggertsson, G.B., 2011. What fiscal policy is effective at zero interest rates? In: Acemoglu, D., Woodford, M. (Eds.), NBER Macroeconomics Annual 2010, NBER Chapters, vol. 25. National Bureau of Economic Research, Inc, pp. 59–112.

Erceg, C.J., Linde, J., 2012. Fiscal consolidation in an open economy. Am. Econ. Rev. 102 (3), 186–191.

Farhi, E., Werning, I., 2012a. Dealing with the trilemma: optimal capital controls with fixed exchange rates. NBER Working Papers No.18199, National Bureau of Economic Research, Inc.

Farhi, E., Werning, I., 2012b. Fiscal unions. NBER Working Papers No.18280, National Bureau of Economic Research, Inc.

Fishback, P.V., Kachanovskaya, V., 2010. In search of the multiplier for federal spending in the states during the great depression. NBER Working Papers, National Bureau of Economic Research, Inc.

Gali, J., Monacelli, T., 2005. Monetary policy and exchange rate volatility in a small open economy. Rev. Econ. Stud. 72 (3), 707–734.

Gali, J., Monacelli, T., 2008. Optimal monetary and fiscal policy in a currency union. J. Int. Econ. 76 (1), 116–132.

Gali, J., Lopez-Salido, J.D., Valles, J., 2007. Understanding the effects of government spending on consumption. J. Eur. Econ. Assoc. 5 (1), 227–270.

Gordon, R.J., Krenn, R., 2010. The end of the great depression 1939-41: policy contributions and fiscal multipliers. NBER Working Papers No. 16380, National Bureau of Economic Research, Inc.

Keynes, J., 1929. The german transfer problem. Econ. J. 39 (153), 1–7.

Mankiw, G.N., 2000. The savers-spenders theory of fiscal policy. Am. Econ. Rev. 90 (2), 120–125.

Nakamura, E., Steinsson, J., 2011. Fiscal stimulus in a monetary union: evidence from U.S. regions. NBER Working Papers No.17391, National Bureau of Economic Research, Inc.

Ohlin, B., 1929. The reparation problem: a discussion. Econ. J. 39 (154), 172–182.

Ramey, V.A., 2011. Can government purchases stimulate the economy? J. Econ. Lit. 49 (3), 673–685.

Serrato, J.C.S., Wingender, P., 2010. Estimating local multipliers. Working Paper, University of California at Berkeley.

Shoag, D., 2010. The impact of government spending shocks: evidence on the multiplier from state pension plan returns. Working Paper, Harvard University.

Werning, I., 2012. Managing a liquidity trap: monetary and fiscal policy. NBER Working Papers, National Bureau of Economic Research, Inc.

Woodford, M., 2011. Monetary policy and financial stability. Working Paper, Columbia University.

CHAPTER 32

What is a Sustainable Public Debt?

P. D'Erasmo[*], E.G. Mendoza[†,‡], J. Zhang[§]
[*]Federal Reserve Bank of Philadelphia, Philadelphia, PA, United States
[†]PIER, University of Pennsylvania, Philadelphia, PA, United States
[‡]NBER, Cambridge, MA, United States
[§]Federal Reserve Bank of Chicago, Chicago, IL, United States

Contents

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.013

Abstract

The question of what is a sustainable public debt is paramount in the macroeconomic analysis of fiscal policy. This question is usually formulated as asking whether the outstanding public debt and its projected path are consistent with those of the government's revenues and expenditures (ie, whether fiscal solvency conditions hold). We identify critical flaws in the traditional approach to evaluate debt sustainability, and examine three alternative approaches that provide useful econometric and model-simulation tools to analyze debt sustainability. The first approach is Bohn's nonstructural empirical framework based on a fiscal reaction function that characterizes the dynamics of sustainable debt and primary balances. The second is a structural approach based on a calibrated dynamic general equilibrium framework with a fully specified fiscal sector, which we use to quantify the positive and normative effects of fiscal policies aimed at restoring fiscal solvency in response to changes in debt. The third approach deviates from the others in assuming that governments cannot commit to repay their domestic debt and can thus optimally decide to default even if debt is sustainable in terms of fiscal solvency. We use these three approaches to analyze debt sustainability in the United States and Europe after the sharp increases in public debt following the 2008 crisis, and find that all three raise serious questions about the prospects of fiscal adjustment and its consequences.

Keywords

Debt sustainability, Fiscal reaction function, Fiscal austerity, Tax policy, Sovereign default

JEL Classification Codes:

E62, F34, F42, H21, H6, H87

1. INTRODUCTION

The question of what is a sustainable public debt has always been paramount in the macroeconomic analysis of fiscal policy, and the recent surge in the debt of many advanced and emerging economies has made it particularly critical. This question is often understood as equivalent to asking whether the government is solvent. That is, whether the outstanding stock of public debt matches the projected present discounted value of the primary fiscal balance, measuring both at the general government level and including all forms of fiscal revenue as well as all current expenditures, transfers and entitlement payments. This chapter revisits the question of public debt sustainability, identifies critical flaws in traditional ways to approach it, and discusses three alternative approaches that provide useful econometric and model-simulation tools to evaluate debt sustainability.

The first approach is an empirical approach proposed in Bohn's seminal work on fiscal solvency. The advantage of this approach is that it provides a straightforward and powerful method to conduct nonstructural empirical tests. These tests require only data on the primary balance, outstanding debt, and a few control variables. The data are then used to estimate linear and nonlinear *fiscal reaction functions* (FRFs), which map the response of the primary balance to changes in outstanding debt, conditional on the control variables. A positive, statistically significant response coefficient is a sufficient condition for the debt to be sustainable. A key lesson from Bohn's work, however, is that using this or other time-series econometric tools just to test for fiscal solvency is futile, because the intertemporal government budget constraint holds under very weak time-series assumptions that are generally satisfied in the data. In particular, Bohn (2007) showed that the constraint holds if either the debt or revenues and expenditures (including debt service) are integrated of *any* finite order. In light of this result, he proposed shifting the focus to analyzing the characteristics of the FRFs in order to study the dynamics of fiscal adjustment that have maintained solvency.

We provide new FRF estimation results for historical data spanning the 1791–2014 period for the United States, and for a cross-country panel of advanced and emerging economies for the period 1951–2013. The results are largely in line with previous findings showing that the response coefficient of the primary balance to outstanding debt is positive and statistically significant in most countries (ie, the sufficiency condition for debt sustainability is supported by the data).[a] On the other hand, the results provide clear evidence of a large structural shift in the response coefficients since the 2008 crisis, which is reflected in large negative residuals in the FRFs since 2009. The primary balances predicted by the FRF of the United States for the period 2008–14 are much larger than the observed ones, and the debt and primary balance dynamics that FRFs predict after 2014 for both the United States and European economies yield higher primary surpluses and lower debt ratios than what official projections show. Moreover, in the case of the United States, the pattern of consistent primary deficits since 2009 and continuing until at least 2020 in official projections, is unprecedented. In all previous episodes of large increases in public debt of comparable magnitudes (the Civil War, the two World Wars, and the Great Depression), the primary balance was in surplus 5 years after the debt peaked.

Using the estimated FRFs, we illustrate that there are multiple parameterizations of a FRF that support the same expected present discounted value of primary balances, and thus all of them make the same initial public debt position sustainable. However, these multiple reaction functions yield different short- and long-run dynamics of debt and primary balances, and therefore differ in terms of social welfare and their macro effects. At this point, this nonstructural approach reaches its limits. The standard Lucas-critique argument implies that estimated FRFs cannot be used to study the implications of fiscal policy changes. Hence, comparing different patterns of fiscal adjustment requires a structural framework that models explicitly the mechanisms and distortions by which tax and expenditure policies affect the economy, the structure of financial markets the government can access, and the implications of the government's inability to commit to repay its obligations.

The second approach to study debt sustainability that we examine picks up at this point. We use a calibrated two-country dynamic general equilibrium framework with

[a] Formally, the null hypothesis that the response coefficient is nonpositive is rejected at the standard confidence level.

a fully specified fiscal sector to study the effects of alternative fiscal strategies to restore fiscal solvency in the aftermath of large increases in debt, assuming that the government is committed to repay. The model is calibrated to data from the United States and Europe and used to quantify the positive and normative effects of fiscal policies that governments may use seeking to increase the present value of the primary fiscal balance by enough to match the increases in debt observed since 2008 (ie, by enough to restore fiscal solvency). This framework has many of the standard elements of the workhorse open-economy Neoclassical model with exogenous long-run balanced growth, but it includes modifications designed to make the model consistent with the observed elasticity of tax bases. As a result, the model captures more accurately the relevant tradeoffs between revenue-generating capacity and distortionary effects in the choice of fiscal instruments.

The results show that indeed alternative fiscal policy strategies that are equivalent in that they restore fiscal solvency, have very different effects on welfare and macro aggregates. Moreover, some fiscal policy setups fall short from producing the changes in the equilibrium present discounted value of primary balances that are necessary to match the observed increases in debt. This is particularly true for taxes on capital in the United States and labor taxes in Europe. The dynamic Laffer curves for these taxes (ie, Laffer curves in terms of the present discounted value of the primary fiscal balance) peak below the level required to make the higher post-2008 debts sustainable.

We also find that, in line with findings in the international macroeconomics literature, the fact that the United States and Europe are financially integrated economies implies that the revenue-generating capacity of taxation on capital income is adversely affected by international externalities.[b] At the prevailing tax structures, increases in US capital income taxes (assuming European taxes are constant) generate significantly smaller increases in the present value of US primary balances than if the United States implemented the same taxes under financial autarky. The model also predicts that at its current capital tax rate, Europe is in the inefficient side of its dynamic Laffer curve for the capital income tax. Hence, lowering its tax, assuming the United States keeps its capital tax constant, induces externalities that enlarge European fiscal revenues, and thus the present value of European primary balances rises significantly more than if Europe implemented the same taxes under financial autarky. This does not imply that debt is easier to sustain in Europe but that the incentives for tax competition are strong, and hence that the assumption that US taxes would remain invariant is unlikely to hold.

The results from the empirical and structural approach suggest that public debt sustainability analysis needs to be extended to consider the implications of the government's lack of commitment to repay domestic obligations. In particular, the evidence of

[b] There is a large empirical and theoretical literature on international taxation and tax competition examining the effects of these externalities. See for example, Frenkel et al. (1991), Huizinga et al. (2012), Klein et al. (2007), Mendoza and Tesar (1998, 2005), Persson and Tabellini (1995), and Sorensen (2003).

structural changes weakening the response of primary balances to debt post-2008, and the findings that tax increases may not be able to generate enough revenue to restore fiscal solvency and are hampered by international externalities, indicate that the risk of default on domestic public debt should be considered. In addition, the ongoing European debt crisis and the recurrent turmoil around federal debt ceiling debates in the United States demonstrate that domestic public debt is not in fact the risk-free asset that is generally taken to be. The first two approaches to study debt sustainability covered in this chapter are not useful for addressing this issue, because they are built on the premise that the government is committed to repay. Note also that the risk here is not that of external sovereign default, which is the subject of a different chapter in this Handbook and has been widely studied in the literature. Instead, the risk here is the one that Reinhart and Rogoff (2011) referred to as "the forgotten history of domestic debt:" Historically, there have been episodes in which governments have defaulted outright on their domestic public debt, and until very recently the macro literature had paid little attention to these episodes. Hence, the third approach we examine assumes that governments cannot commit to repay domestic debt, and decide optimally to default even if standard solvency conditions hold, and even when domestic debt holders enter in the payoff function of the sovereign making the default decision. Sustainable debt in this setup is the debt that can be supported as a market equilibrium with positive quantity and price, exposed with positive probability to a government default, and with actual episodes in which default is the equilibrium outcome.

In this framework, the government maximizes a social welfare function that assigns positive weight to the welfare of all domestic agents in the economy, including those who are holders of government debt. Defaulting on public debt is useful as a tool for redistributing resources across agents, but is also costly because debt effectively provides liquidity to credit-constrained agents and serves as a vehicle for tax-smoothing and self-insurance.[c] If default is costless, debt is unsustainable for a utilitarian government because default is always optimal. Debt can be sustainable if default carries a cost or if the government's social welfare function has a bias in favor of bond holders. In addition, this second assumption can be an equilibrium outcome under majority voting if the fraction of agents that do not own debt is sufficiently large, because these agents benefit from the consumption-smoothing ability that public debt issuance provides for them, and may thus choose a government biased in favor of bond holders over a utilitarian government. A quantitative application of this setup calibrated to data from Europe shows how the

[c] This view of default costs is motivated by the findings of Aiyagari and McGrattan (1998) on the social value of domestic public debt as the vehicle for self insurance in a model of heterogeneous agents assuming the government is committed to repay. Birkeland and Prescott (2006) show that public debt also has social value as a mechanism for tax smoothing when population growth declines, taxes distort labor, and intergenerational transfers fund retirement. Welfare when public debt is used to save for retirement is larger than in a tax-and-transfer system.

tradeoff between these costs and benefits of default determines sustainable debt. Domestic default occurs with low probability and returns on government debt carry default premia, and in the setup with a government biased in favor of bondholders the sustainable debt is large and rises with the concentration of debt ownership.

The rest of this chapter is organized as follows: Section 2 discusses the classic and empirical approaches to evaluate debt sustainability, including the new FRF estimation results. Section 3 focuses on the structural approach. It examines the quantitative predictions of the two-country dynamic general equilibrium model for the positive and normative effects of fiscal policies aimed at restoring fiscal solvency in response to large increases in debt, including the application to the case of the United States and Europe. Section 4 covers the domestic default approach, with the quantitative example based on European data. Section 5 provides a critical assessment of all three approaches and an outlook with directions for future research. Section 6 summarizes the main conclusions.

2. EMPIRICAL APPROACH

Several articles and conference volumes survey the large literature on indicators of public debt sustainability and empirical tests of fiscal solvency (eg, Buiter, 1985; Blanchard, 1990; Blanchard et al., 1990; Chalk and Hemming, 2000; IMF, International Monetary Fund, 2003; Afonso, 2005; Bohn, 2008; Neck and Sturm, 2008, and Escolano, 2010). These surveys generally start by formulating standard concepts of government accounting, and then build around them the arguments to construct indicators of debt sustainability or tests of fiscal solvency. We proceed here in a similar way, but adopting a general formulation following the analysis of government debt in the textbook by Ljungqvist and Sargent (2012). The advantage of this formulation is that it is explicit about the structure of asset markets, which as we show below turns out to be critical for the design of empirical tests of fiscal solvency.

Consider a simple economy in which output and total government outlays (ie, current expenditures and transfer payments) are exogenous functions of a vector of random variables s denoted $y(s_t)$ and $g(s_t)$, respectively. The exogenous state vector follows a standard discrete Markov process with transition probability matrix $\pi(s_{t+1},s_t)$. Taxes at date t depend on s_t and on the outstanding public debt, but since the latter is the result of the history of values of s up to and including date t, denoted s^t, taxes can be expressed as $\tau_t(s^t)$. In terms of asset markets, this economy has a full set of state-contingent Arrow securities with a j-step ahead equilibrium pricing kernel given by $Q_j(s_{t+j}|s_t) = MRS(c_{t+j},c_t)\pi^j(s_{t+j},s_t)$.[d]

[d] $MRS(c_{t+j},c_t) \equiv \beta^j u'(c(s_{t+j}))/u'(c(s_t))$ is the marginal rate of substitution in consumption between date $t+j$ and date t. Note also that in this simple economy the resource constraint implies that consumption is exogenous and given by $c(s_{t+j}) = y(s_{t+j}) - g(s_{t+j})$.

Public debt outstanding at the beginning of date t is denoted as $b_{t-1}(s_t|s^{t-1})$, which is the amount of date-t goods that the government promised at $t-1$ to deliver if the economy is in state s_t at date t with history s^{t-1}. The government's budget constraint can then be written as follows:

$$\sum_{s_{t+1}} Q_1(s_{t+1}|s_t)b_t(s_{t+1}|s^t)\pi(s_{t+1},s_t) - b_{t-1}(s_t|s^{t-1}) = g(s_t) - \tau_t(s^t).$$

Notice that there are no restrictions on what type of financial instruments the government uses to borrow. In particular, the typical case in which the government issues only risk-free debt is not ruled out. In this case, the above budget constraint reduces to the familiar form: $[b_t(s^t)/R_1(s_t)] - b_{t-1}(s^{t-1}) = g(s_t) - \tau_t(s^t)$, where $R_1(s_t)$ is the one-step-ahead risk-free real interest rate (which at equilibrium satisfies $R_1(s_t)^{-1} = E_t[MRS(c_{t+1},c_t)]$).

Imposing the no-Ponzi game condition $\lim\inf_{j\to\infty} E_t[MRS(c_{t+j},c_t)b_{t+j}] = 0$ on the above budget constraint, and using the equilibrium asset pricing conditions, yields the following intertemporal government budget constraint (IGBC):

$$b_{t-1} = pb_t + \sum_{j=1}^{\infty} E_t[MRS(c_{t+j},c_t)pb_{t+j}], \tag{1}$$

where $pb_t \equiv \tau_t - g_t$ is the primary fiscal balance. This IGBC condition is the familiar fiscal solvency condition that anchors the standard concept of debt sustainability: b_{t-1} is said to be sustainable if it matches the expected present discounted value of the stream of future primary fiscal balances. Hence, the two main goals of most of the empirical literature on public debt sustainability have been: (a) to construct simple indicators that can be used to assess debt sustainability, and (b) to develop formal econometric tests that can determine whether the hypothesis that IGBC holds can be rejected by the data.

2.1 Classic Debt Sustainability Analysis

Classic public debt sustainability analysis focuses on the long-run implications of a deterministic version of the IGBC. This approach uses the government budget constraint evaluated at steady state as a condition that relates the long-run primary fiscal balance as a share of GDP and the debt-output ratio, and defines the latter as the sustainable debt (see Buiter, 1985, Blanchard, 1990, and Blanchard et al., 1990). To derive this condition from the setup described earlier, first remove uncertainty from the government budget constraint with nonstate contingent debt to obtain: $[b_t/(1+r_t)] - b_{t-1} = -pb_t$. Then rewrite the equation with government bonds at face value instead of discount bonds: $b_t - (1+r_t)b_{t-1} = -pb_t$. Finally, apply a change of variables so that debt and primary balances are measured as GDP ratios, which implies that the effective interest rate becomes $r_t \equiv (1+i_t^r)/(1+\gamma_t) - 1$, where i_t^r is the real interest rate and γ_t is the growth rate of GDP (or alternatively use the nominal interest rate and the growth rate of nominal GDP). Solving for the steady-state debt ratio yields:

$$b^{ss} = \frac{pb^{ss}}{r} \approx \frac{pb^{ss}}{i^r - \gamma}. \tag{2}$$

Thus, the steady-state debt ratio $b^{ss}$is the annuity value of the steady state primary balance pb^{ss}, discounted at the long-run, growth-adjusted interest rate. In policy applications, this condition is used either as an indicator of the primary balance-output ratio needed to stabilize a given debt-output ratio (the so-called "debt stabilizing" primary balance), or as an indicator of the sustainable target debt-output ratio that a given primary balance-output ratio can support. There are also variations of this approach that use the constraint $b_t - (1 + r_t)b_{t-1} = -pb_t$ to construct estimates of primary balance targets needed to produce desired changes in debt at shorter horizons than the steady state. For instance, imposing the condition that the debt must decline ($b_t - b_{t-1} < 0$), implies that the primary balance must yield a surplus that is at least as large as the growth-adjusted debt service: $pb_t \geq r_t b_{t-1}$.

The Classic Approach was developed in the 1980s but remains a tool widely used in policy assessments of sustainable debt. In particular, Annex VI of IMF (2013) instructs IMF economists to use a variation of the Blanchard ratio, called the Exceptional Fiscal Performance Approach, as one of three methodologies for estimating maximum sustainable public debt ranges (the other two methodologies introduce uncertainty and are discussed later in this section). This variation determines a country's maximum sustainable primary balance and "appropriate" levels of i^r and γ, and then applies them to the Blanchard ratio to estimate the maximum level of debt that the country can sustain.

The main flaw of the Classic Approach is that it only *defines* what long-run debt is for a given long-run primary balance (or vice versa) if stationarity holds, or *defines* lower bounds on the short-run dynamics of the primary balance. It does not actually connect the outstanding initial debt of a particular period b_{t-1} with b^{ss}, where the latter should be $\lim_{j\to\infty} b_{t+j}$ starting from b_{t-1}, and thus it cannot actually guarantee that b_{t-1} is sustainable in the sense of satisfying the IGBC. In fact, as we show below, for a given b_{t-1} there are multiple dynamic paths of the primary balance that satisfy IGBC. A subset of these paths converges to stationary debt positions, with different values of b^{ss} that vary widely depending on the primary balance dynamics, and there is even a subset of these paths for which the debt diverges to infinity but is still consistent with IGBC!

A second important flaw of the Classic approach is the absence of uncertainty and considerations about the asset market structure. Policy institutions have developed several methodologies that introduce uncertainty into debt sustainability analysis. For example, Barnhill and Kopits (2003) proposed incorporating uncertainty by adapting the value-at-risk (VaR) methodology of the financial industry to debt instruments issued by governments. Their methodology aims to quantify the probability of a negative net worth position for the government. Other methodologies described in IMF (2013) use stochastic time-series simulation tools to examine debt dynamics, estimating models for the individual components of the primary balance or nonstructural vector-autoregression models that include these variables jointly with key macroeconomic aggregates (eg, output growth, inflation) and a set of exogenous variables. The goal is to compute probability density functions of possible debt-output ratios based on forward simulations of the time-series models. The distributions are then used to make assessments of sustainable debt in terms of the probability that the simulated debt ratios are greater or equal than a critical value, or to construct "fan charts" summarizing the confidence intervals of the future evolution of debt. More recently, Ostry et al. (2015) use the fiscal reaction functions estimated by Ghosh et al. (2013) and discussed later in this section to construct measures of "fiscal space," which are intended to show the space a country has for increasing its debt ratio while still satisfying the IGBC.

IMF (2013) proposes two other stochastic tools as part of the framework for quantifying maximum sustainable debt (complementing the deterministic Exceptional Fiscal Performance estimates discussed earlier). The first is labeled the Early Warning Approach. This method computes a threshold debt ratio above which a country is likely to experience a debt crisis. The threshold is optimized with respect to the type-1 (false alarms of crises) and type-2 (missed warnings of crises) errors it produces, by minimizing the sum of the ratio of missed crises to total crises periods and false alarms to total noncrises periods. The second tool, labeled the Uncertainty Approach, is actually the same as the method proposed by Mendoza and Oviedo (2009), to which we turn next.[e]

The stochastic methods reviewed above have the significant shortcoming that, as with the Blanchard ratio, they cannot guarantee that their sustainable debt estimates satisfy the IGBC. Moreover, they introduce uncertainty without taking into account the fact that typically government debt is in the form of non-state-contingent instruments. The setup proposed by Mendoza and Oviedo (2006, 2009) addresses these two shortcomings. In this setup, the government issues non-state-contingent debt facing stochastic Markov processes for government revenues and outlays (ie, asset markets are incomplete). The key assumption is that the government is committed to repay, which imposes a constraint on public debt akin to Ayagari's Natural Debt Limit for private debt in Bewley models of heterogeneous agents with incomplete markets.

Following the simple version of this framework presented in Mendoza and Oviedo (2009), assume that output follows a deterministic trend, with an exogenous growth rate given by γ, and that the real interest rate is constant. Assume also that the government

[e] IMF (2013) refers to this approach as "a derivative of the exceptional fiscal performance approach and relies on the same underlying concepts and equations." As we explain, however, Blanchard ratios and their variations differ significantly from the debt limits and debt dynamics characterized by Mendoza and Oviedo (2009).

keeps its outlays smooth, unless it finds itself unable to borrow more, and when this happens it cuts its outlays to minimum tolerable levels.[f] Since the government cannot have its outlays fall below this minimum level, it does not hold more debt than the amount it could service after a long history in which $pb(s^t)$ remains at its worst possible realization (ie, the primary balance obtained with the worst realization of revenues, $\tau^{\min}$, and public outlays cut to their tolerable minimum $g^{\min}$), which can happen with positive probability. This situation is defined as a state of fiscal crisis and it sets and upper bound on debt denoted the "Natural Public Debt Limit" (NPDL), which is given by the growth-adjusted annuity value of the primary balance in the state of fiscal crisis:

$$b_t \leq NPDL \equiv \frac{\tau^{\min} - g^{\min}}{i^r - \gamma}. \tag{3}$$

This result together with the government budget constraint yields a law of motion for debt that follows this simple rule: $b_t = \min\left[NPDL, (1 + r_t)b_{t-1} - pb_t\right] \geq \bar{b}$, where $\bar{b}$ is an assumed lower bound for debt that can be set to zero for simplicity (ie, the government cannot become a net creditor).[g]

Notice that NPDL is lower for governments that have (a) higher variability in public revenues (ie, lower $\tau^{\min}$ in the support of the Markov process of revenues), (b) less flexibility to adjust public outlays (higher $g^{\min}$), or (c) lower growth rates and/or higher real interest rates. The stark differences between NPDL and b^{ss} from the classic debt sustainability analysis are also important to note. The expressions are similar, but the two methods yield sharply different implications for debt sustainability: The classic approach will always identify as sustainable debt ratios that are unsustainable according to the NPDL, because in practice b^{ss} uses the average primary fiscal balance, instead of its worst realization, and as a result it yields a long-run debt ratio that violates the NPDL. Moreover, while b^{ss} cannot be related to the IGBC, the debt rule $b_t = \max\left[NPDL, (1 + r_t)b_{t-1} - pb_t\right] \geq \bar{b}$ always satisfies the IGBC, because debt is bounded above at the NPDL, which guarantees that the no-Ponzi game condition cannot be violated. Note also, however, that the NPDL is a measure of the largest debt that a government can maintain, and not an estimate of the long-run average debt ratio or of the stationary debt ratio.

[f] This is a useful assumption to keep the setup simple, but is not critical. Mendoza and Oviedo (2006) model government expenditures entering a CRRA utility function as an optimal decision of the government, and here the curvature of the utility function imposes the debt limit in the same way as in Bewley models.

[g] This debt rule has an equivalent representation as a lower bound on the primary balance: $pb_t \geq (1 + r_t)b_{t-1} - NPDL$. On the date of a fiscal crisis, b_t hits NPDL. The next period, if the lowest realization of revenues is drawn again, pb_{t+1} hits $\tau^{\min} - g^{\min}$. Debt and the primary balance remain unchanged until higher revenue realizations are drawn, and the larger surpluses reduce the debt. See section III.3 of Mendoza and Oviedo (2009) for stochastic simulations of a numerical example.

The NPDL can be turned into a policy indicator by characterizing the probabilistic processes of the components of the primary balance together with some simplifying assumptions. On the revenue side, the probabilistic process of tax revenues reflects the uncertainty affecting tax rates and tax bases. This uncertainty includes domestic tax policy variability, the endogenous response of the economy to that variability, and other factors that can be largely exogenous to the domestic economy (eg, the effects of fluctuations in commodity prices and commodity exports on government revenues). On the expenditure side, government expenditures adjust partly in response to policy decisions, but the manner in which they respond varies widely across countries, as the literature on procyclical fiscal policy in emerging economies has shown (eg, see Alesina and Tabellini, 2005; Kaminsky et al., 2005; Talvi and Vegh, 2005).

The quantitative analysis in Mendoza and Oviedo (2009) treats the revenue and expenditures processes as exogenous, and calibrates them to 1990–2005 data from four Latin American economies.[h] Since the value of the expenditure cuts that each country can commit to is unobservable, they calculate instead the implied cuts in government outlays, relative to each country's average (ie, $g^{\min} - E[g]$), that would be needed so that each country's NPDL is consistent with the largest debt ratio observed in the sample. The largest debt ratios are around 55% for all four countries (Brazil, Colombia, Costa Rica, and Mexico), but the cuts in outlays that make these debt ratios consistent with the NPDL range from 3.8 percentage points of GDP for Costa Rica to 6.2 percentage points for Brazil. This is the case largely because revenues in Brazil have a coefficient of variation of 12.8%, vs 7% in Costa Rica, and hence to support a similar NPDL at a much higher revenue volatility requires higher $g^{\min}$. Mendoza and Oviedo also showed that the time-series dynamics of debt follow a random walk with boundaries at NPDL and $\bar{b}$.

2.2 Bohn's Debt Sustainability Framework

In a series of influential articles published between 1995 and 2011, Henning Bohn made four major contributions to the empirical literature on debt sustainability tests:

1. *IGBC tests that discount future primary balances at the risk-free rates are misspecified, because the correct discount factors are determined by the state-contingent equilibrium pricing kernel (Bohn, 1995).*[i] Tests affected by this problem include those reported in several well-known empirical studies (eg, Hamilton and Flavin, 1986, Hansen et al., 1991, and Gali, 1991). Following Ljungqvist and Sargent (2012), this mispecification

[h] Mendoza and Oviedo (2006) endogenize the choice of government outlays and decentralize the private and public borrowing decisions in a small open economy model with nonstate-contingent assets.

[i] Lucas (2012) raised a similar point in a different context. She argued that the relevant discount rate for government flows should not be the risk-free rate but a cost of capital that incorporates the market risk associated with government activities.

error is easy to illustrate by using the equilibrium risk-free rates ($R_{t+j}^{-1} = E_t\left[MRS(c_{t+j}, c_t)\right]$) to rewrite the IGBC as follows:

$$b_{t-1} = pb_t + \sum_{j=1}^{\infty}\left[\frac{E_t[pb_{t+j}]}{R_{t+j}} + cov_t\left(MRS(c_{t+j}, c_t), pb_{t+j}\right)\right]. \tag{4}$$

Hence, discounting the primary balances at the risk-free rates is only correct if

$$\sum_{j=1}^{\infty} cov_t\left(MRS(c_{t+j}, c_t), pb_{t+j}\right) = 0.$$

This would be true under one of the following assumptions: (a) perfect foresight, (b) risk-neutral private agents, or (c) primary fiscal balances that are uncorrelated with future marginal utilities of consumption. All of these assumptions are unrealistic, and (c) in particular runs contrary to the strong empirical evidence showing that primary balances are not only correlated with macro fluctuations, but show a strikingly distinct pattern across industrial and developing countries: primary balances are procyclical in industrial countries, and acyclical or countercyclical in developing countries. Moreover, Bohn (1995) also showed examples in which this misspecification error leads to incorrect inferences that reject fiscal solvency when it actually does hold. For instance, a rule that maintains g/y and b/y constant in a balanced-growth economy with i.i.d. output growth violates the mispeficied IGBC if mean output growth is greater or equal than the interest rate, but it does satisfy condition (1).

2. *Testing for debt sustainability is futile, because the IGBC holds under very weak assumptions about the time-series processes of fiscal data that are generally satisfied. The IGBC holds if either debt or revenue and spending inclusive of debt service are integrated of finite but arbitrarily high order (Bohn, 2007).* This invalidates several fiscal solvency tests based on specific stationarity and cointegration conditions (eg, Hamilton and Flavin, 1986; Trehan and Walsh, 1988; Quintos, 1995), because neither a particular order of integration of the debt data, nor the cointegration of revenues and government outlays is necessary for debt sustainability. As Bohn explains in the proof of this result, the reason is intuitive: In the forward conditional expectation that forms the no-Ponzi game condition, the j^{th} power of the discount factor asymptotically dominates the expectation $E_t(b_{t+j})$ as $j \rightarrow \infty$ if the debt is integrated of any finite order. This occurs because $E_t(b_{t+j})$ is *at most* a polynomial of order n if b is integrated of order n, while the discount factor is exponential in j, and exponential growth dominates polynomial growth. But perhaps of even more significance is the implication that, since integration of finite order is indeed a very weak condition, testing for fiscal solvency or debt sustainability per se is not useful: The data are all but certain to reject the hypotheses that debt or revenue and spending inclusive of debt service are nonstationary after differencing the

data a finite number of times (usually only once!). Bohn (2007) concluded that, in light of this result, using econometric tools to try and identify in the data fiscal reaction functions that support fiscal solvency and studying their dynamics is "more promising for understanding deficit problems."

3. *A linear fiscal reaction function (FRF) with a statistically significant, positive (conditional) response of the primary balance to outstanding debt is sufficient for the IGBC to hold (Bohn, 1998, 2008).* Proposition 1 in Bohn (2008) demonstrates that this linear FRF is sufficient to satisfy the IGBC:

$$pb_t = \mu_t + \rho b_{t-1} + \varepsilon_t,$$

for all t, where $\rho > 0$, μ_t is a set of additional determinants of the primary balance, which typically include an intercept and proxies for temporary fluctuations in output and government expenditures, and ε_t is i.i.d. The proof only requires that μ_t be bounded and that the present value of GDP be finite. Intuitively, the argument of the proof is that with pb changing by the positive factor ρ when debt rises, the growth of the debt j periods ahead is lowered by $(1-\rho)^j$. Formally, for any small $\rho > 0$, the following holds as $j \to \infty$: $E_t[MRS(c_{t+j}, c_t)b_{t+j}] \approx (1-\rho)^j b_t \to 0$, which in turn implies that the NPG condition and thus the IGBC hold. Note also that while debt sustainability holds for any $\rho > 0$, the long-run behavior of the debt ratio differs sharply depending on the relative values of the mean r and ρ. To see why, combine the FRF and the government budget constraint to obtain the law of motion of the debt ratio $b_t = -\mu_t + (1 + r_t - \rho)b_{t-1} + \varepsilon_t$. Hence, debt is stationary only if $\rho > r$, otherwise it explodes, but as long as $\rho > 0$ it does so at a slow enough pace to still satisfy IGBC.[j] In addition, the IGBC holds for the same value of initial debt for any $\rho > 0$, but, if $\rho > r$, debt converges to a higher long-run average as ρ falls.

The above results also show why the steady-state debt b^{ss} of the classic debt sustainability analysis is not useful for assessing debt sustainability: With the linear FRF, multiple well-defined long-run averages of debt are consistent with debt sustainability, each determined by the particular value of the response coefficient in the range $\rho > r$, and even exploding debt is consistent with debt sustainability if $0 < \rho < r$. Moreover, in the limit as $r \to 0$, the Blanchard ratio of the classic analysis predicts that debt diverges to infinity ($b^{ss} \to \infty$ if pb^{ss} is finite), while the linear FRF predicts that both b and pb are mean-reverting to well-defined long-run averages given by $-\mu/\rho$ and 0. Similarly, notions of a "maximal sustainable interest rate" are meaningless from

[j] Bohn (2007) shows that this result holds for any of the following three assumptions about the interest rate process: (i) $r_t = r$ for all t, (ii) r_t is a stochastic process that is serially uncorrelated with $E_t[r_{t+1}] = r$, or (iii) r_t is any stochastic process with mean r subject only to implicit restrictions such that $b_t = \frac{1}{1+r}E_t[pb_{t+1} + b_{t+1} - (r_{t+1} - r)b_t]$.

the perspective of assessing whether the debt satisfies the IGBC, because $\rho > 0$ is sufficient for IGBC to hold regardless of the value of r.[k]

4. *Empirical tests of the linear FRF based on historical U.S. data and various subsamples reject the hypothesis that $\rho \leq 0$, so IGBC holds (Bohn, 1998, 2008).* In his 2008 article, Bohn constructed a dataset going back to 1791, the start of US public debt after the Funding Act of 1790, and found that the response coefficient estimated with 1793–2003 data is positive and significant, ranging from 0.1 to 0.12. Moreover, looking deeper into the fiscal dynamics he found that economic growth has been sufficient to cover the entire servicing costs of US public debt, but there are structural breaks in the response coefficient. The 1793–2003 estimates are about twice as large as those obtained in Bohn (1998) using data for 1916–2005, which is a period that emphasizes the cold-war era of declining debt but high military spending.

Bohn's framework has been applied to cross-country datasets by Mendoza and Ostry (2008) and extended to include a nonlinear specification allowing for default risk by Ghosh et al. (2013).[l] Mendoza and Ostry found estimates of response coefficients for a panel of industrial countries that are similar to those Bohn (1998) obtained for the United States. In addition, they found that the solvency condition holds for a panel that includes both industrial and developing countries, as well as in a subpanel that includes only the latter. They also found, however, that cross-sectional breaks are present in the data at particular debt thresholds. In the combined panel and the subpanels with only advanced or only developing economies, there are high-debt country groups for which the response coefficient is not statistically significantly different from zero. Ghosh et al. found that the response coefficients fall sharply at high debt levels, and obtained estimates of fiscal space that measure the distance between observed debt ratios and the largest debt ratios that can be supported given debt limits implied by the presence of default risk.

2.3 Estimated Fiscal Reaction Functions and Their Implications

We provide below new estimation results for linear FRFs for the United States using historical data from 1791 to 2014, and for a cross-country panel using data for the 1951–2013 period. Some of the results are in line with the findings of previous studies, but the key difference is that there is a significant break in the response of the primary

[k] This is not the case if the government cannot commit to repay its debt. In external sovereign default models in the vein of Eaton and Gersovitz (1981), for example, the interest rate is an increasing, convex function of the debt stock, and there exists a debt level at which rationing occurs because future default on newly issued debt becomes a certain event.

[l] The same approach has also been used to test for external solvency (ie, whether the present discounted value of the balance of trade matches the observed net foreign asset position). Durdu et al. (2013) conducted cross-country empirical tests using data for 50 countries over the 1970–2006 period and found that the data cannot reject the hypothesis of external solvency, which in this case is measured as a negative response of net exports to net foreign assets.

balance to debt after 2008. We then use the estimation results and the historical data to put in perspective the current fiscal situation of the United States and Europe. In particular, we show that: (a) primary balance adjustment in the United States is lagging significantly behind what has been observed in the aftermath of previous episodes of large increases in debt, (b) observed primary deficits have been much larger than what the FRFs predict, and (c) hypothetical scenarios with alternative response coefficients produce sharply different patterns of transitional dynamics and long-run debt ratios, but they are all consistent with the same observed initial debt ratios (ie, IGBC holds for all of them).

2.3.1 FRF Estimation Results

Table 1 shows estimation results for the FRF of the United States using historical data for the 1791–2014 period. The table shows results for five regression models similar to those estimated in Bohn (1998, 2008). Column (1) shows the base model, which uses as regressors the initial debt ratio, the cyclical component of output, and temporary military expenditures as a measure of transitory fluctuations in government expenditures.[m] Column (2) introduces a nonlinear spline coefficient when the debt is higher than the mean. Column (3) introduces an AR(1) error term. Column (4) adds the squared mean deviation of the debt ratio. Column (5) includes a time trend. Columns (6) and (7) provide modifications that are important for showing the structural instability of the FRF post-2008: Column (6) reruns the base model truncating the sample in last year of the sample used in Bohn (2008) and Column (7) uses a sample that ends in 2008. The signs of the debt, output gap and military expenditures coefficients are the same as in Bohn's regressions, and in particular the response coefficient estimates are generally positive, which satisfies the sufficiency condition for debt sustainability.

In Columns (1)–(5), the point estimates of ρ range between 0.077 and 0.105, which are lower than Bohn's 2008 estimates based on 1793–2003 data, but higher than his (1998) estimates based on 1916–95 data. The ρ estimates are always statistically significant, although only at the 90% confidence level in the base and squared-debt models.

Column (6) shows that if we run the linear FRF over the same sample period as in Bohn (2008), the results are very similar to his (see in particular Column 1 of table 7 in his paper).[n] The point estimate of ρ is 0.105, compared with 0.121 in Bohn's study (both statistically significant at the 99% confidence level). But in our base model of Column (1) we found that using the full sample that runs through 2014 the point estimate of ρ falls to 0.078. Moreover, excluding the post-2008-crisis data in Column (7), the results

[m] We follow Bohn in measuring this temporary component as the residual of an AR(2) process for military expenditures.

[n] They only differ because we defined military expenditures as the sum of expenditures by the Department of Defense and the Veterans Administration for the full sample, excluding international relations, while Bohn includes Veterans starting in 1940 and adds international relations.

Table 1 Fiscal reaction function of the United States: 1792–2014

Model:	**Base model**	**Asymmetric response**	**AR(1) term**	**Debt squared**	**Time trend**	**Bohn's sample (1793–2003)**	**Prerecession (1793–2008)**
Coefficient	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Constant	0.00648	0.00540	0.00974	0.00653	0.00601	0.00485	0.00470
	(0.004)	(0.003)*	(0.008)	(0.004)	(0.006)	(0.003)*	(0.003)
Initial debt d_t^*	0.07779	0.08689	0.10477	0.07715	0.07674	0.10498	0.10188
	(0.040)*	(0.030)***	(0.032)***	(0.038)*	(0.035)**	(0.023)***	(0.022)***
GDP gap	0.07404	0.07300	0.15330	0.07390	0.07490	0.07987	0.07407
	(0.078)	(0.079)	(0.043)***	(0.079)	(0.077)	(0.086)	(0.086)
Military expenditure	−0.72302	−0.72001	−0.98955	−0.72320	−0.72462	−0.77835	−0.76857
	(0.133)***	(0.136)***	(0.110)***	(0.133)***	(0.135)***	(0.135)***	(0.135)***
$\max(0, d_t^* - \bar{d})$		−0.14487					
		(0.061)					
AR(1)			0.89154				
			(0.029)***				
$(d_t^* - \bar{d})^2$				0.00261			
				(0.044)			
Time trend					6.89×10^{-06}		
					(5.9×10^{-05})		
s.e.	0.0239	0.0240	0.198	0.0120	0.0240	0.0210	0.0209
Adj. R-squared:	0.606	0.605	0.901	0.614	0.605	0.695	0688
Observations:	223	223	222	223	223	213	217

Note: HAC standard errors shown in parenthesis, 2-lag window prewhitening. "*,""**," and "***" denote that the corresponding coefficient is statistically significant at the 90%, 95%, and 99% confidence levels. Output gap is percent deviation from Hodrick–Prescott trend. Military expenditure includes all Department of Defense and Department of Veterans Affairs outlays.

are very similar to those obtained with the same sample period as Bohn's. Hence, these results suggest that the addition of the post-2008 data, a tumultous period in the fiscal stance of the United States, produces a structural shift in the FRF.[o] Testing formally for this hypothesis, we found that Chow's forecast test rejects strongly the null hypothesis of no structural change in the value of ρ when the post-2008 data are added. Hence, the decline in the estimate of ρ from 0.102 to 0.078 is statistically significant. This change in the response of the primary balance to higher debt ratios may seem small, but it implies that the primary balance adjustment is about 25% smaller, and as we show later this results in large changes in the short- and long-run dynamics of debt.

The regressions with nonlinear features (Column (2) with the debt spline at the mean debt ratio, and Column (4) with the squared deviation from the mean debt ratio) are very different from Bohn's estimates. In Bohn (1998), the FRF with the same spline term has a negative point estimate $\rho = -0.015$ and a large, positive spline coefficient of 0.105 when debt is above its mean, so that for above-average debt ratios the response of the primary balance is stronger than for below-average debt ratios, and becomes positive with a net effect of 0.09, which is consistent with debt sustainability. In contrast, Table 1 shows a ρ estimate of 0.09 with a spline coefficient of -0.14. Hence, these results suggest that the response of the primary balance is weaker for above-average debt ratios, and the net effect is negative at -0.05, which violates the linear FRF's sufficiency condition for debt sustainability. The spline coefficient is not, however, statistically significant. For the squared-debt regressions, Bohn (2008) estimated a positive coefficient of 0.02, while the coefficient shown in Table 1 is only 0.003 (both not statistically significant). Thus, both the debt-spline and debt-squared regressions are also consistent with the possibility of a structural change in the FRF. In particular, the stronger primary balance response at higher debt ratios that Bohn identified in his 1998 and 2008 studies changed to a much weaker response once the data up to 2014 are introduced. The rationale for this is that the large debt increases since 2008 have been accompanied by adjustments in the primary balance that differ sharply from what has been observed in previous episodes of large debt increases, as we illustrate below.

Tables 2–4 show the results of cross-country panel regressions similar to those reported by Mendoza and Ostry (2008) and Ghosh et al. (2013), but expanded to include data for the 1951–2013 period for 25 advanced and 33 emerging economies. The first six columns of results in these tables show three pairs of regression models. Each pair uses a different measure of government expenditures, since the measure based on military expenditures used in the US regressions is unavailable and/or less relevant as a measure of the temporary component of government expenditures in the international dataset.

[o] Bohn (2008) also found evidence of structural shifts when contrasting his results for 1784–2003 with his 1916–95 results, with sharply lower response coefficients for the shorter sample, which he attributed to the larger weight of the cold-war era (in which debt declined while military spending remained high).

Table 2 Fiscal reaction functions of advanced economies (1951–2013)

	All advanced economies					
Model	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Constant	11.23917	1.76019	−1.02696	−0.07294	−1.42979	0.02521
	(3.134)***	(0.037)***	(0.472)**	(0.195)	(2.651)	(0.222)
Previous debt	0.06916	0.01461	0.01983	0.00295	0.02750	−0.00076
d_{t-1}	(0.013)***	(0.001)***	(0.010)**	(0.005)	(0.010)***	(0.005)
GDP gap	0.17053	0.28046	0.31501	0.34696	0.34939	0.40503
	(0.050)***	(0.058)***	(0.065)***	(0.060)***	(0.073)***	(0.073)***
Government	−0.35654	−0.06305				
Expenditure	(0.078)***	(0.013)***				
Government			−0.10449	−0.12511		
Expenditure gap			(0.031)***	(0.031)***		
Govt consumption					−0.20579	−0.33638
Gap (Nat. Acc.)					(0.064)***	(0.070)***
Country AR(1)	Yes	No	Yes	No	Yes	No
s.e.	1.603	2.814	1.709	2.813	1.796	2.884
Adj. R-squared:	0.766	0.277	0.755	0.306	0.733	0.304
Observations:	1285	1346	1218	1273	1139	1186
Countries:	25	25	25	25	25	25

Note: All regressions include country fixed effect and White cross-section corrected standard errors and covariances. Standard errors shown in parenthesis. "*,""**," and "***" denote that the corresponding coefficient is statistically significant at the 90%, 95%, and 99% confidence levels. Output, government expenditure, and government consumption gaps are percent deviation from Hodrick–Prescott trend.

Table 3 Fiscal reaction functions of emerging economies (1951–2013)

Model	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Constant	9.99549	1.32486	−2.38214	−1.88325	−2.33727	−1.70461
	(1.473)***	(0.409)***	(0.462)***	(0.284)***	(0.544)***	(0.322)***
Previous debt	0.03806	0.05657	0.05452	0.04519	0.05280	0.04376
d_{t-1}	(0.009)***	(0.006)***	(0.006)***	(0.005)***	(0.008)***	(0.006)***
GDP gap	0.03698	0.07352	0.15962	0.15509	0.07568	0.06831
	(0.029)	(0.027)***	(0.034)***	(0.027)***	(0.042)*	(0.030)**
Government	−0.44322	−0.15638				
Expenditure	(0.049)***	(0.020)***				
Government			−0.11986	−0.12420		
Expenditure gap			(0.012)***	(0.012)***		
Govt					−0.01302	−0.02662
Consumption						
Gap (Nat. Acc.)					(0.018)	(0.014)*
Country AR(1)	Yes	No	Yes	No	Yes	No
s.e.	1.854	2.630	1.772	2.450	2.072	2.795
Adj. R-squared:	0.666	0.346	0.698	0.437	0.589	0.321
Observations:	1071	1144	977	1035	967	1022
Countries:	33	33	33	33	33	33

Note: All regressions include country fixed effect and White cross-section corrected standard errors and covariances. Standard errors shown in parenthesis. "*,""**," and "***" denote that the corresponding coefficient is statistically significant at the 90%, 95%, and 99% confidence levels. Output, government expenditure, and government consumption gaps are percent deviation from Hodrick–Prescott trend.

Table 4 Fiscal reaction functions for advanced and emerging economies (1951–2013)

Model	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Constant	10.53960	1.50777	−2.23188	−0.65482	−2.29040	−0.57649
	(1.528)***	(0.357)***	(0.400)***	(0.160)***	(0.466)***	(0.172)***
Previous debt	0.05138	0.02962	0.04576	0.01634	0.04661	0.01500
d_{t-1}	(0.007)***	(0.004)***	(0.006)***	(0.004)***	(0.006)***	(0.004)***
GDP gap	0.07864	0.12611	0.20956	0.20590	0.16205	0.15198
	(0.031)**	(0.030)***	(0.043)***	(0.032)***	(0.051)***	(0.036)***
Government	−0.40043	−0.08823				
Expenditure	(0.047)***	(0.015)***				
Government			−0.11558	−0.12788		
Expenditure Gap			(0.014)***	(0.016)***		
Govt consumption					−0.03764	−0.07534
Gap (Nat. Acc.)					(0.021)*	(0.020)***
Country AR(1)	Yes	No	Yes	No	Yes	No
s.e.	1.729	2.796	1.756	2.727	1.970	2.915
Adj R-squared:	0.720	0.275	0.718	0.328	0.656	0.254
Observations:	2356	2490	2195	2308	2106	2208
Countries:	58	58	58	58	58	58

Note: All regressions include country fixed effect and White cross-section corrected standard errors and covariances. Standard errors shown in parenthesis. “*,”“**,” and “***” denote that the corresponding coefficient is statistically significant at the 90%, 95%, and 99% confidence levels. Output, government expenditure, and government consumption gaps are percent deviation from Hodrick–Prescott trend.

Models (1) and (2) use total real government outlays (ie, current expenditures plus all other noninterest expenditures, including transfer payments), models (3) and (4) use the cyclical component of total real outlays, and models (5) and (6) follow Mendoza and Ostry (2008) and use the cyclical component of real government absorption from the national accounts (ie, real current government expenditures). Models (1), (3) and (5) include country-specific AR(1) terms, which Mendoza and Ostry also found important to consider, while model (2), (4), and (6) do not.

Two caveats about the measures of government expenditures used in these regressions. First, they are less representative of unexpected increases in government expenditures, particularly the HP cyclical component because of the double-sided nature of the HP filter. Second, since the primary balance is the difference between total revenues and expenditures, adding the latter as a regressor implies that revenues are the only endogenous component of the dependent variable that can respond to changes in debt. This is less true when we use only the cyclical component of expenditures and/or use only current expenditures instead of total outlays, but it remains a potential limitation. Interestingly, the coefficients on government expenditures do have the same sign as in the US regressions with temporary military expenditures (although they are about half the size), and they are statistically significant at the 99% confidence level. These caveats do imply, however, that the coefficients on government expenditures cannot be interpreted as measuring only the response of the primary balance to unexpected increases in government expenditures, but can reflect also differences in the cyclical stance of fiscal policies and in the degree of access to debt markets (see Mendoza and Ostry, 2008 for a discussion of these issues).

Table 2 shows that, as in Mendoza and Ostry, considering the country-specific AR(1) terms in the cross-country panel is important. The advanced economies' response coefficients are higher and with significantly smaller standard errors when the autocorrelation of error terms is corrected. Hence, we focus the rest of the discussion of the panel results on the results with AR(1) terms.

The advanced economies' response coefficients of the primary balance on debt in the AR(1) models are positive and statistically significant in general. The coefficients are smaller in the regressions that use cyclical components of either total outlays or current expenditures (models (3) and (5)) than in the one that uses the level of government outlays (model (1)), but across the first two the ρ coefficients are similar (0.02 vs 0.028). Following again Mendoza and Ostry, we focus on the regressions that use the cyclical components of current government expenditures.

Comparing the FRFs with country AR(1) terms and using the cyclical component of current government expenditures across the three panel datasets, Tables 2–4 show that the estimates of ρ are 0.028 for advanced economies, 0.053 for emerging economies, and 0.047 for the combined panel. Mendoza and Ostry obtained estimates of 0.02 for advanced economies and 0.036 for both emerging economies and the combined panel.

The results are somewhat different, but the two are consistent in producing larger values of ρ for emerging economies and the combined panel than for advanced economies.

The difference in the response coefficients across advanced and emerging economies highlights important features of their debt dynamics. Condition (4) suggests that countries with procyclical fiscal policy (ie, acyclical or countercyclical primary balances) can sustain higher debt ratios than countries with countercyclical fiscal policy (ie, procyclical primary balances). Yet we observe the opposite in the data: Advanced economies conduct countercyclical fiscal policy and show higher average debt ratios than emerging economies, which display procyclical or acyclical fiscal policy (ie, significantly lower primary balance-output gap correlations). Indeed, the higher ρ of the emerging economies implies that these countries converge to lower mean debt ratios in the long run. As Mendoza and Ostry (2008) concluded, this higher ρ is not an indicator of "more sustainable" fiscal policies in emerging economies, but evidence of the fact that past increases in debt of a given magnitude in these countries require a stronger conditional response of the primary balance, and hence less reliance on debt markets, than in advanced economies.

2.3.2 Implications for Europe and the United States

Public debt and fiscal deficits rose sharply in several advanced economies after the 2008 global financial crisis, in response to both expansionary fiscal policies and policies aimed at stabilizing financial systems. To put in perspective the magnitude of this recent surge in debt, it is useful to examine Bohn's historical dataset of public debt and primary balances for the United States. Defining a public debt crisis as a year-on-year increase in the public debt ratio larger than twice the historical standard deviation, which is equivalent to more than 8.15 percentage points in Bohn's dataset, we identify five debt crisis events (see Fig. 1): The two world wars (World War I with an increase of 28.7 percentage points of GDP over 1918–19 and World War II with 59.3 percentage points over 1943–45), the Civil War (19.7 percentage points over 1862–63), the Great Depression (18.5 percentage points over 1932–33), and the Great Recession (22.3 percentage points over 2009–10). The Great Recession episode is the third largest, ahead of the Civil War and the Great Depression episodes.

Fig. 2 illustrates the short-run dynamics of the US primary fiscal balance after each of the five debt crises. Each crisis started with large deficits, ranging from 4% of GDP for the Great Depression to nearly 20% of GDP for World War II, but the Great Recession episode is unique in that the primary balance remains in deficit 4 years after the crisis. In the three war-related crises, a large primary deficit turned into a small surplus within 3 years. By contrast, the latest baseline scenario from the Congressional Budget Office (*Updated Budget and Economic Outlook: 2015–2025*, January 2015), projects that the US primary balance will continue in deficit for the next 10 years. The primary deficit is projected to shrink to 0.6% of GDP in 2018 and then hover near 1% through 2025. In addition,

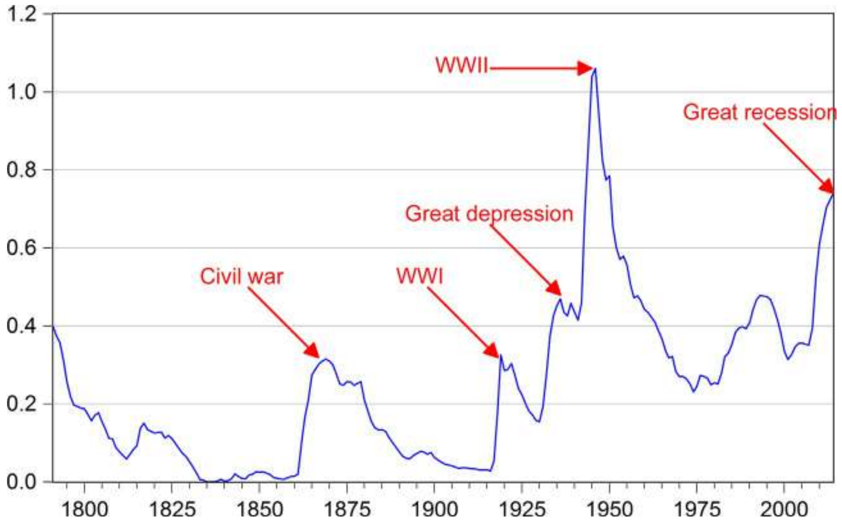


Fig. 1 US government debt as percentage of GDP.

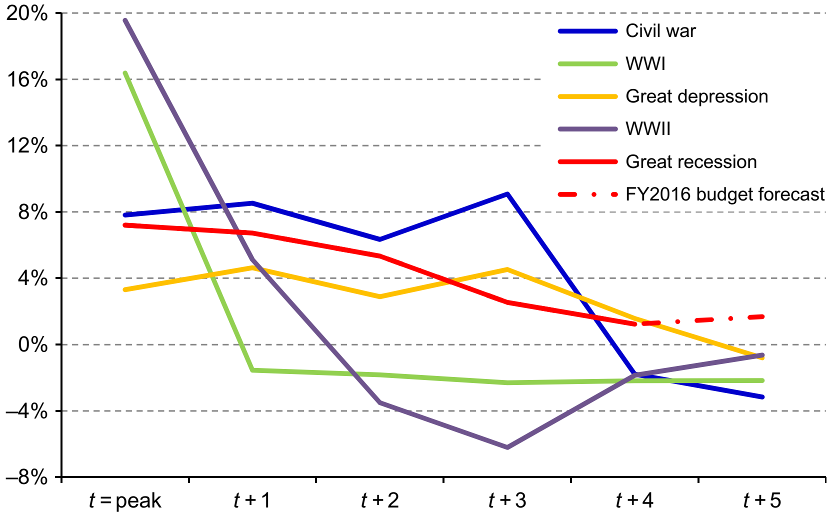


Fig. 2 US government deficits after debt crises.

relative to the Great Depression, the first three deficits of the Great Recession were nearly twice as large, and by 5 years after the debt crisis of the Great Depression the United States had a primary surplus of nearly 1% of GDP. In summary, the post-2008 increase in public debt has been of historic proportions, and the absence of primary surpluses in both the 4 years after the surge in debt and the projections for 2015–25 is *unprecedented* in US history.

Many advanced European economies have not fared much better. Weighted by GDP, the average public debt ratio of the 15 largest European economies rose from 38% to 58% between 2007 and 2011. The increase was particularly large in the five countries at the center of the European debt crisis (Greece, Ireland, Italy, Portugal and Spain), where the debt ratio weighted by GDP rose from 75% to 105%, but even in some of the largest European economies public debt rose sharply (by 33 and 27 percentage points in the United Kingdom and France, respectively).

The estimated FRFs can be used to examine the implications of these rapid increases in public debt ratios for debt sustainability and for the short- and long-run dynamics of debt and deficits. Consider first the regression residuals. Fig. 3 shows the residuals of the US fiscal reaction function estimated in the base model (1) of Table 1, and Fig. 4 shows rolling residuals from the same regression. These two plots show that the residuals for 2008–14 are significantly negative, and much larger in absolute value than the residuals in the rest of the sample period. In fact, the residuals for 2009–11 are twice as large as the corresponding minus-two-standard-error bound. Thus, the primary deficits observed during the post-2008 years have been much larger than what the FRFs predicted, even after accounting for the larger deficits that the FRFs allow on account of the depth of the recession and expansionary government expenditures. These large residuals are of course consistent with the results documented earlier showing evidence of structural change in the FRF when the post-2008 data are added.

The structural change in the FRF can also be illustrated by comparing the actual primary balances from 2009 to 2014 and the government-projected primary balances for

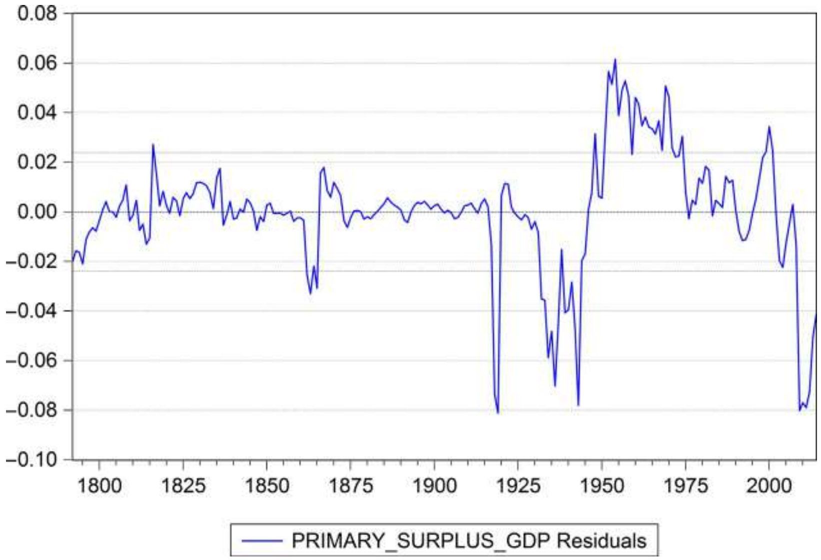


Fig. 3 Residuals for the US fiscal reaction function. *Note*: This residuals correspond to the Base Model (1) in Table 1. The dotted lines are at two s.d. above and below zero.

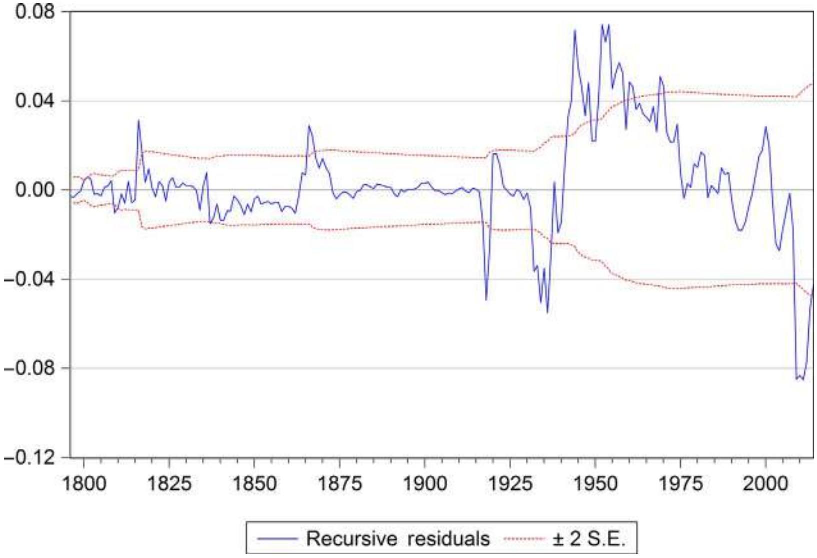


Fig. 4 Rolling residuals for the US fiscal reaction function. *Note*: For each sample 1791-*t*, the baseline specification, model (1) in Table 1, is estimated and the residual at time *t* is reported together with the 2 standard deviation band for the errors in that sample.

2015 to 2020 in the *President's Budget for Fiscal Year 2016* with the out-of-sample forecast that the FRF estimated with data up to 2008 in Column (7) of Table 1 produces (see Fig. 5). To construct this forecast, we use the observed realizations of the cyclical components of output and government expenditures from 2009 to 2014, and for 2015 to 2020 we use again data from the projections in the *President's Budget*.

As Fig. 5 shows, for the period 2009–14, the primary balance showed deficits significantly larger than what the FRF predicted, and also much larger than the deficit at the minus-two-standard-error bound of the forecast band. The mean forecast of the FRF predicted a rising primary surplus from zero to about 4% of GDP between 2009 and 2014, while the data showed deficits narrowing from 8% to about 2% of GDP. In addition, the primary deficits projected in the *President's Budget* are also much larger than predicted by the mean forecast of the FRF, with the projections at or below the minus-two-standard error band. Bohn (2011) warned that already by 2011 there were signs of a likely structural break, because his estimated FRFs called for primary surpluses when the debt ratio surpassed 55–60%, while the 2012 *Budget* projected large and persistent primary deficits at debt ratios much higher than those.

The estimated FRF results can also be used to study projected time-series paths for public debt and the primary balance as of the latest actual observations (2014). To simulate the debt dynamics, we use the law of motion for public debt that results from combining the government budget constraint and the FRF mentioned earlier: $b_t = -\mu_t + (1 + r_t - \rho)b_{t-1} + \varepsilon_t$. We consider baseline scenarios in which we use estimated

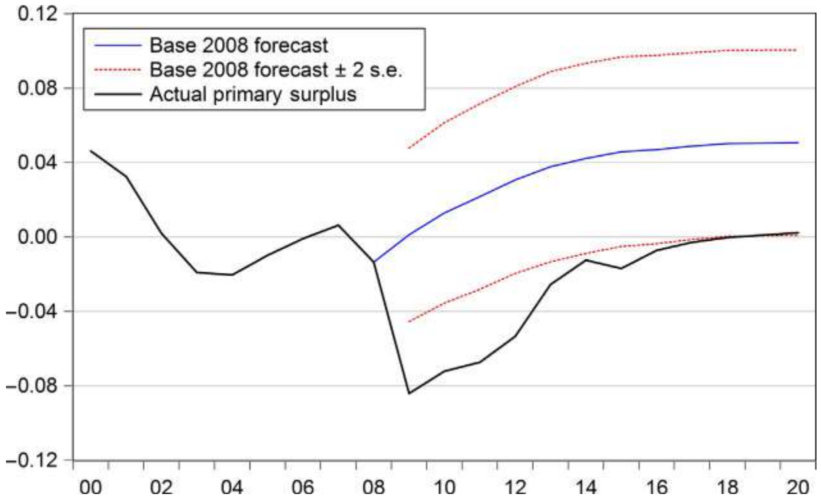


Fig. 5 US primary surplus actual value and 2008 based forecast. *Note*: The forecast is based on model (7) in Table 1 which has the sample restricted to 1791–2008. Given actual values of debt-to-GDP ratio, GDP gap, and military expenditure a forecast of the primary surplus to GDP ratio is generated for the sample 2009–20. Actual variables from 2015 onward correspond to estimates included in The president's budget for fiscal year 2016. Chow's forecast test rejects the null hypothesis of no structural change starting in 2009 with 99.9% confidence.

ρ coefficients for Europe and the United States, and simulate forward starting from the 2014 observations. For the United States, we used model (3) in Table 1. For Europe, we use model (5) from Table 2 and take a simple cross-section average among European industrialized countries. Projections of the future values of the fluctuations in output and government expenditures are generated with simple univariate AR models. In addition, we compare these baseline projection scenarios with scenarios in which we lower the response coefficient to half of the regression estimates or lower the intercept of the FRFs. Recall from the earlier discussion that changing these parameters, as long as $\rho > 0$, generates the same present discounted value of the primary balance as the baseline scenarios, but as we show below the transitional dynamics and long-run debt ratios they produce are very different. These simulations also require assumptions about the values of the real interest rate and the growth rate that determine $1 + r$. For simplicity, we assume that $r = 0$, which rules out the range in which debt can grown infinitely large but still be consistent with the IGBC (ie, the range $0 < \rho < r$), and it also implies that primary balances converge to zero in the long run.[p]

[p] Real interest rates on government debt and rates of output growth in large industrial countries are low but with expectations of an eventual increase. Rather than taking a stance on the difference between the two, we just assumed here that they are equal.

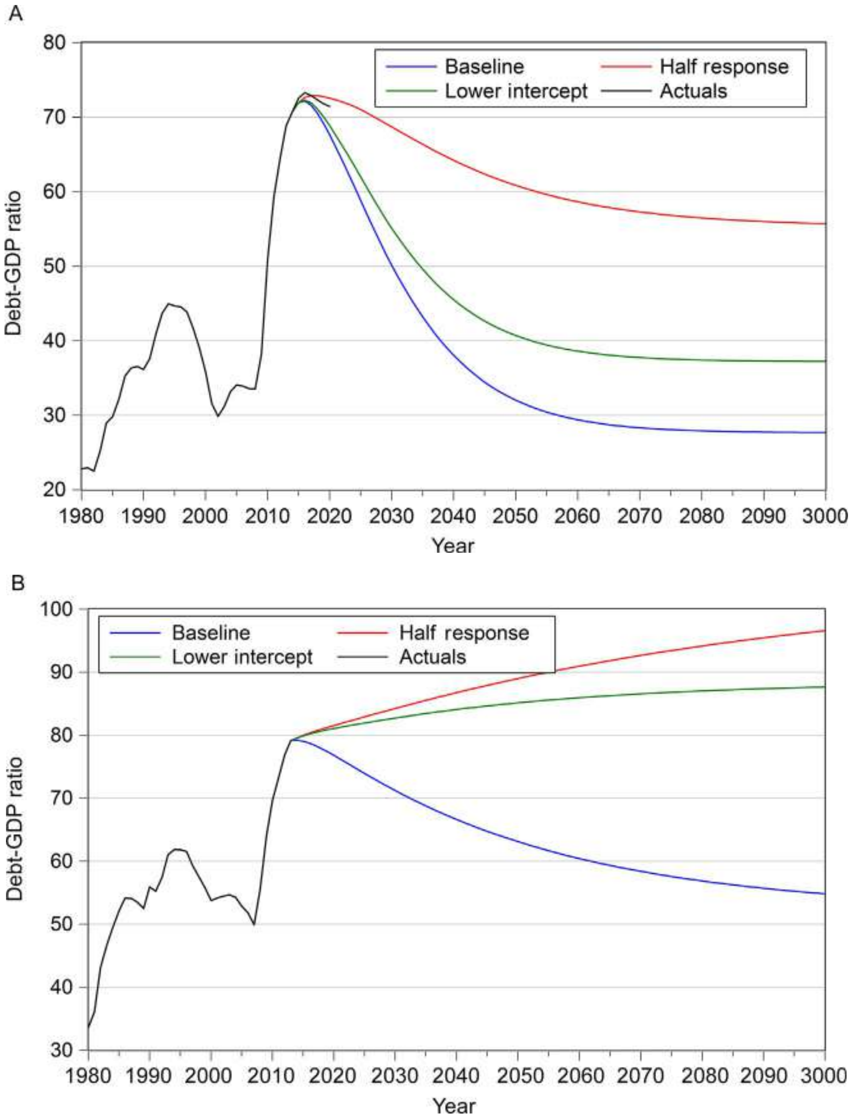


Fig. 6 Debt-to-GDP actuals and simulations since 2014. (A) US debt to GDP. (B) Europe debt to GDP. *Note*: For the United States: Model (3) in Table 1 is used in conjunction with estimated AR(2) processes for the output gap and military expenditure, plus the government budget constraint. For Europe: Model (5) in Table 2 is used in conjunction with estimated AR(1) processes for the output gap and government consumption gap in each country, and a simple average among advanced European countries is taken.

Fig. 6 and 7 show the projected paths of debt ratios and primary balances for the base-line and the alternative scenarios, for both the United States and Europe. These plots show that under the baseline scenario the countries should be reporting primary surpluses that will decline monotonically over time, and should therefore display a monotonically

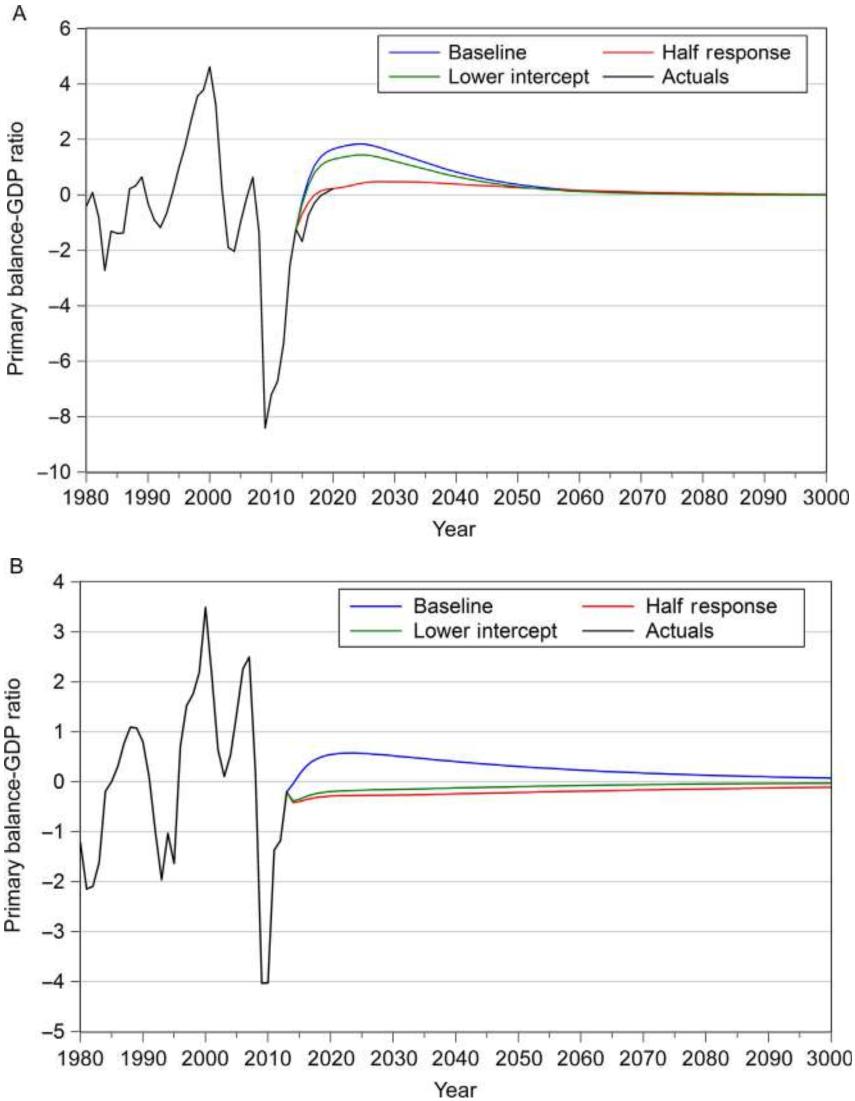


Fig. 7 Primary balance to GDP actuals and simulations since 2014. (A) US primary balance to GDP. (B) Europe primary balance to GDP. *Note*: For details on the construction of this simulations see note on Fig. 6.

declining path for the debt ratio converging back to the average observed in the sample period of the FRF estimates. With lower ρ or lower intercept, the initial surpluses can be significantly smaller or even turned into deficits, but the long-run mean debt ratio would increase significantly. In the case of the United Sates, for example, the long-run average of the debt ratio would rise from 29% in the baseline case to around 57% in the scenario with lower ρ.

All the debt and primary balance paths shown in Figs. 6 and 7 satisfy the same IGBC, and therefore make the same initial debt ratio sustainable, but clearly their macroeconomic implications cannot be the same. Unfortunately, at this point the FRF approach reaches its limits. To evaluate the positive and normative implications of alternative paths of fiscal adjustment, we need a structural framework that can be used to quantify the implications of particular revenue and expenditure policies for equilibrium allocations and prices and for social welfare.

3. STRUCTURAL APPROACH

This section presents a two-country dynamic general equilibrium framework of fiscal adjustment, and uses it to quantify the positive and normative effects of alternative fiscal policy strategies to restore fiscal solvency (ie, maintain debt sustainability) in the United States and Europe after the recent surge in public debt ratios. The structure of the model is similar to the Neoclassical models widely studied in the large quantitative literature on optimal taxation, the effects of tax reforms, and international tax competition (see, for example, Lucas, 1990, Chari et al., 1994, Cooley and Hansen, 1992, Mendoza and Tesar, 1998, 2005, Prescott, 2004, Trabandt and Uhlig, 2011, etc.). In particular, we use the two-country model proposed by Mendoza et al. (2014), which introduces modifications to the Neoclassical model that allow it to match empirical estimates of the elasticity of tax bases to change in tax rates. This is done by introducing endogenous capacity utilization and by limiting the tax allowance for depreciation of physical capital to approximate the allowance reflected in the data.[q]

3.1 Dynamic Equilibrium Model

Consider a world economy that consists of two countries or regions: home (H) and foreign (F). Each country is inhabited by an infinitely-lived representative household, and has a representative firm that produces a single tradable good using as inputs labor, l, and units of utilized capital, $\widetilde{k} = mk$ (where k is installed physical capital and m is the utilization rate). Capital and labor are immobile across countries, but the countries are perfectly integrated in goods and asset markets. Trade in assets is limited to one-period discount bonds denoted by b and sold at a price q. Assuming this simple asset-market structure is without loss of generality, because the model is deterministic.

Following King et al. (1988), growth is exogenous and driven by labor-augmenting technological change that occurs at a rate γ. Accordingly, stationarity of all variables (except labor and leisure) is induced by dividing them by the level

[q] Dynamic models of taxation that consider endogenous capacity utilization include the theoretical analysis of optimal capital income taxes by Ferraro (2010) and the quantitative analysis of the effects of taxes in an RBC model by Greenwood and Huffman (1991).

of this technological factor.[r] The stationarity-inducing transformation of the model also requires discounting utility flows at the rate $\widetilde{\beta} = \beta(1+\gamma)^{1-\sigma}$, where β is the standard subjective discount factor and σ is the coefficient of relative risk aversion of CRRA preferences, and adjusting the laws of motion of k and b so that the date $t+1$ stocks grow by the balanced-growth factor $1+\gamma$.

We describe below the structure of preferences, technology and the government sector of the home country. The same structure applies to the foreign country, and when needed foreign country variables are identified by an asterisk.

3.1.1 Households, Firms, and Government

3.1.1.1 Households

The preferences of the representative home household are standard:

$$\sum_{t=0}^{\infty} \widetilde{\beta}^{t} \frac{\left(c_t(1-l_t)^a\right)^{1-\sigma}}{1-\sigma}, \sigma > 1, a > 0, \text{ and } 0 < \widetilde{\beta} < 1. \tag{5}$$

The period utility function is CRRA in terms of a CES composite good made of consumption, c_t, and leisure, $1 - l_t$ (assuming a unit time endowment). $\frac{1}{\sigma}$ is the intertemporal elasticity of substitution in consumption, and a governs both the Frisch and intertemporal elasticities of labor supply for a given value of σ.[s]

The household takes as given proportional tax rates on consumption, labor income and capital income, denoted τ_C, τ_L, and τ_K, respectively, lump-sum government transfers or entitlement payments, denoted by e_t, the rental rates of labor w_t and capital services r_t, and the prices of domestic government bonds and international-traded bonds, q_t^g and q_t.[t]

The household rents $\widetilde{k}$ and l to firms, and makes the investment and capacity utilization decisions. As is common in models with endogenous utilization, the rate of depreciation of the capital stock increases with the utilization rate, according to a convex function $\delta(m) = \chi_0 m^{\chi_1}/\chi_1$, with $\chi_1 > 1$and $\chi_0 > 0$ so that $0 \leq \delta(m) \leq 1$.

Investment incurs quadratic adjustment costs:

$$\phi(k_{t+1}, k_t, m_t) = \frac{\eta}{2}\left(\frac{(1+\gamma)k_{t+1} - (1-\delta(m_t))k_t}{k_t} - z\right)^2 k_t,$$

[r] The assumption that growth is exogenous implies that tax policies do not affect long-run growth, in line with the empirical findings of Mendoza et al. (1997).

[s] We are using the standard functional form of the utility function from the canonical exogenous balanced growth model as in King et al. (1988) and many RBC applications. This function implies a constant Frisch elasticity for $\sigma = 1$. See Trabandt and Uhlig (2011) for a generalized formulation of the utility function that maintains the constant Frisch elasticity when $\sigma > 1$, and a discussion of the role of the Frisch elasticity in the use of Neoclassical models to quantify the macroeconomic effects of tax changes.

[t] The gross yields in these bonds are simply the reciprocal of these prices.

where the coefficient η determines the speed of adjustment of the capital stock, while z is a constant set equal to the long-run investment-capital ratio, so that at steady state the capital adjustment cost is zero.

The household chooses intertemporal sequences of consumption, leisure, investment inclusive of adjustment costs x, international bonds, domestic government bonds d, and utilization to maximize (5) subject to a sequence of period budget constraints given by:

$$\begin{aligned}(1+\tau_c)c_t + x_t + (1+\gamma)(q_t b_{t+1} + q_t^g d_{t+1}) =& (1-\tau_L)w_t l_t + (1-\tau_K)r_t m_t k_t \\ &+\theta\tau_K\overline{\delta}k_t + b_t + d_t + e_t,\end{aligned} \tag{6}$$

and the following law of motion for the capital stock:

$$x_t = (1+\gamma)k_{t+1} - (1-\delta(m_t))k_t + \phi(k_{t+1}, k_t, m_t),$$

for $t = 0, \ldots, \infty$, given the initial conditions $k_0 > 0$, b_0, and d_0.

The left-hand-side of equation (6) includes all the uses of household income, and the right-hand-side includes all the sources of income net of income taxes. We impose a standard no-Ponzi-game condition on households, and hence the present value of total household expenditures equals the present value of after-tax income plus initial asset holdings.

Notice that in calculating post-tax income in the above budget constraints, we consider a capital tax allowance $\theta\tau_K\overline{\delta}k_t$ for a fraction θ of depreciation costs. This formulation of the depreciation allowance reflects two assumptions about how the allowance works in actual tax codes: First, depreciation allowances are usually set in terms of fixed depreciation rates applied to the book or tax value of capital, instead of the true physical depreciation rate that varies with utilization. Hence, we set the depreciation rate for the capital tax allowance at a constant rate $\overline{\delta}$ that differs from the actual physical depreciation rate $\delta(m)$. The second assumption is that the depreciation allowance only applies to a fraction θ of the capital stock, because in practice it generally applies only to the capital income of businesses and self-employed, and not to residential capital.[u]

We assume that capital income is taxed according to the residence principle, in line with features of the tax systems in the United States and Europe, but countries are allowed to tax capital income at different rates.[v] This also implies, however, that in order

[u] Using the standard 100% depreciation allowance also has two unrealistic implications. First, it renders m independent of τ_K in the long-run. Second, in the short-run τ_K affects the utilization decision margin only to the extent that it reduces the marginal benefit of utilization when traded off against the marginal cost due to changes in the marginal cost of investment.

[v] In principle, the choice of residence vs source based taxation can be viewed as part of the choices made along with the values of tax rates. Indeed, Huizinga (1995) shows that generally optimal taxation would call for a mix of source- and residence-based taxation. In practice, however, most tax systems are effectively residence-based, because widespread bilateral tax treaties provide for source-based-determined tax payments of residents of one country to claim credits for taxes paid to foreign governments.

to support a competitive equilibrium with different capital taxes across countries we must assume that physical capital is owned entirely by domestic residents. Without this assumption, cross-country arbitrage of returns across capital and bonds at common world prices implies equalization of pre- and post-tax returns on capital, which therefore requires identical capital income taxes across countries. For the same reason, we must assume that international bond payments are taxed at a common world rate, which we set to zero for simplicity. For more details, see Mendoza and Tesar (1998). Other forms of financial-market segmentation, such as trading costs or short-selling constraints, could be introduced for the same purpose, but make the model less tractable.

3.1.1.2 Firms

Firms hire labor and effective capital services to maximize profits, given by $y_t - w_t l_t - r_t \tilde{k}_t$, taking factor rental rates as given. The production function is assumed to be Cobb–Douglas:

$$y_t = F(\tilde{k}_t, l_t) = \tilde{k}_t^{1-\alpha} l_t^{\alpha}$$

where α is labor's share of income and $0 < \alpha < 1$. Firms behave competitively and thus choose $\tilde{k}_t$ and l_t according to standard conditions:

$$(1-\alpha)\tilde{k}_t^{-\alpha} l_t^{\alpha} = r_t,$$

$$\alpha \tilde{k}_t l_t^{\alpha-1} = w_t.$$

Because of the linear homogeneity of the production technology, these factor demand conditions imply that at equilibrium $y_t = w_t l_t + r_t \tilde{k}_t$.

3.1.1.3 Government

Fiscal policy has three components. First, government outlays, which include predetermined sequences of government purchases of goods, g_t, and transfer/entitlement payments, e_t, for $t = 0, \ldots, \infty$. In our baseline results, we assume that $g_t = \bar{g}$ and $e_t = \bar{e}$ where $\bar{g}$ and $\bar{e}$ are the steady state levels of government purchases and transfers before the post-2008 surge in public debt. Because entitlements are lump-sum transfer payments, they are always nondistortionary in this representative agent setup, but still a calibrated value of $\bar{e}$ creates the need for the government to raise distortionary tax revenue, since we do not allow for lump-sum taxation. Government purchases do not enter in household utility or the production function, and hence it would follow trivially that a strategy to restore fiscal solvency after an increase in debt should include setting $g_t = 0$. We rule out this possibility because it is unrealistic, and also because if the model is modified to allow government purchases to provide utility or production benefits, cuts in these purchases would be distortionary in a way analogous to raising taxes.

The second component of fiscal policy is the tax structure. This includes time invariant tax rates on consumption τ_C, labor income τ_L, capital income τ_K, and the depreciation allowance limited to a fraction θ of depreciation expenses.

The third component is government debt, d_t. We assume the government is committed to repay its debt, and thus it must satisfy the following sequence of budget constraints for $t=0,\ldots,\infty$:

$$d_t-(1+\gamma)q_t^g d_{t+1}=\tau_C c_t+\tau_L w_t l_t+\tau_K(r_t m_t-\theta\overline{\delta})k_t-(g_t+e_t).$$

The right-hand-side of this equation is the primary fiscal balance, which is financed with the change in debt net of debt service in the left-hand-side of the constraint.

Public debt is sustainable in this setup in the same sense as we defined it in Section 2. The IGBC must hold (or equivalently, the government must also satisfy a no-Ponzi-game condition): The present value of the primary fiscal balance equals the initial public debt d_0. Since we calibrate the model using shares of GDP, it is useful to rewrite the IGBC also in shares of GDP. Defining the primary balance as $pb_t\equiv\tau_C c_t+\tau_L w_t l_t+\tau_K(r_t m_t-\theta{-}\delta)k_t-(g_t+e_t)$, the IGBC in shares of GDP is:

$$\frac{d_0}{y_{-1}}=\psi_0\left[\frac{pb_0}{y_0}+\sum_{t=1}^{\infty}\left(\left[\prod_{i=0}^{t-1}v_i\right]\frac{pb_t}{y_t}\right)\right], \tag{7}$$

where $v_i\equiv(1+\gamma)\psi_i q_i^g$ and $\psi_i\equiv y_{i+1}/y_i$. In this expression, primary balances are discounted to account for long-run growth at rate γ, transitional growth ψ_i as the economy converges to the long-run, and the equilibrium price of public debt q_i^g. Since y_0 is endogenous (ie, it responds to increases in d_0 and the fiscal policy adjustments needed to offset them), we write the debt ratio in the left-hand-side as a share of pre-debt-shock output y_{-1}, which is predetermined.

Combining the budget constraints of the household and the government, and the firm's zero-profit condition, we obtain the home resource constraint:

$$F(m_t k_t,l_t)-c_t-g_t-x_t=(1+\gamma)q_t b_{t+1}-b_t.$$

3.1.2 Equilibrium, Tax Distortions, and International Externalities

A competitive equilibrium for the model is a sequence of prices $\{r_t,\ r_t^*,\ q_t,\ q_t^g,\ q_t^{g*},\ w_t,\ w_t^*\}$ and allocations $\{k_{t+1},\ k_{t+1}^*,\ m_{t+1},\ m_{t+1}^*,\ b_{t+1},\ b_{t+1}^*,\ x_t,\ x_t^*,\ l_t,\ l_t^*,\ c_t,\ c_t^*,\ d_{t+1},\ d_{t+1}^*\}$ for $t=0,\ldots,\infty$ such that: (a) households in each region maximize utility subject to their corresponding budget constraints and no-Ponzi game constraints, taking as given all fiscal policy variables, pretax prices, and factor rental rates; (b) firms maximize profits subject to the Cobb–Douglas technology taking as given pretax factor rental rates; (c) the government budget constraints hold for given tax rates and exogenous sequences of government purchases and entitlements; and (d) the following market-clearing conditions hold in the global markets of goods and bonds:

$$\omega(y_t - c_t - x_t - g_t) + (1-\omega)\left(y_t^* - c_t^* - x_t^* - g_t^*\right) = 0,$$

$$\omega b_t + (1-\omega) b_t^* = 0,$$

where ω denotes the initial relative size of the two regions.

The model's optimality conditions are useful for characterizing the model's tax distortions and their international externalities. Consider first the Euler equations for capital (excluding adjustment costs for simplicity), international bonds and domestic government bonds. These equations yield the following arbitrage conditions:

$$\frac{(1+\gamma)u_1(c_t, 1-l_t)}{\tilde{\beta}\, u_1(c_{t+1}, 1-l_{t+1})} = (1-\tau_K)F_1(m_{t+1}k_{t+1}, l_{t+1})m_{t+1} + 1 - \delta(m_{t+1}) + \tau_K\theta\overline{\delta} = \frac{1}{q_t} = \frac{1}{q_t^g},$$
$$\frac{(1+\gamma)u_1(c_t^*, 1-l_t^*)}{\tilde{\beta}\, u_1(c_{t+1}^*, 1-l_{t+1}^*)} = (1-\tau_K^*)F_1(m_{t+1}^*k_{t+1}^*, l_{t+1}^*)m_{t+1}^* + 1 - \delta(m_{t+1}^*) + \tau_K^*\theta\overline{\delta} = \frac{1}{q_t} = \frac{1}{q_t^{g*}}. \tag{8}$$

Fully integrated financial markets imply that intertemporal marginal rates of substitution in consumption are equalized across regions, and are also equal to the rate of return on international bonds. Since physical capital is immobile across countries, and capital income taxes are residence-based, households in each region face their own region's tax on capital income. Arbitrage equalizes the after-tax returns on capital across regions, but pre-tax returns differ, and hence differences in tax rates are reflected in differences in capital stocks and output across regions. Arbitrage in asset markets also implies that bond prices are equalized. Hence, at equilibrium: $q_t = q_t^g = q_t^{g*}$.

As shown in Mendoza and Tesar (1998), unilateral changes in the capital income tax result in a permanent reallocation of physical capital, and ultimately a permanent shift in wealth, from a high-tax to a low-tax region. Thus, even though physical capital is immobile across countries, perfect mobility of financial capital and arbitrage of asset returns induces movements akin to international mobility of physical capital. In the stationary state with balanced growth, however, the global interest rate R (the inverse of the bond price, $R \equiv 1/q$) is a function of β, γ and σ:

$$R = \frac{(1+\gamma)^\sigma}{\beta},$$

and thus is independent of tax rates. The interest rate does change along the transition path and alters the paths of consumption, output and international asset holdings. In particular, as is standard in the international tax competition literature, each country would have an incentive to behave strategically by tilting the path of the world interest rate in its favor to attract more capital. When both countries attempt this, the outcome is lower capital taxes but also lower welfare for both (which is the well-known race-to-the-bottom result of the tax competition literature).

Consider next the optimality condition for labor:

$$\frac{u_2(c_t, 1-l_t)}{u_1(c_t, 1-l_t)} = \frac{1-\tau_L}{1+\tau_C} F_2(k_t, l_t).$$

Labor and consumption taxes drive the standard wedge $(1 - \tau_W) \equiv (1 - \tau_L)/(1 + \tau_C)$ between the leisure-consumption marginal rate of substitution and the pre-tax real wage (which is equal to the marginal product of labor). Since government outlays are kept constant and the consumption tax is constant, consumption taxation does not distort saving plans, and hence any $(\tau_C,\ \tau_L)$ pair consistent with the same τ_W yields identical allocations, prices, and welfare.

Many Neoclassical and Neokeynesian dynamic equilibrium models feature tax distortions like the ones discussed above, but they also tend to underestimate the elasticity of the capital tax base to changes in capital taxes, because k is predetermined at the beginning of each period, and changes gradually as it converges to steady state. In the model we described, the elasticity of the capital tax base can be adjusted to match the data because capital income taxes have an additional distortion absent from the other models: They distort capacity utilization decisions. In particular, the optimality condition for the choice of m_t is:

$$F_1(m_t k_t, l_t) = \frac{1+\Phi_t}{1-\tau_K} \delta'(m_t), \tag{9}$$

where $\Phi_t = \eta\left(\frac{(1+\gamma)k_{t+1} - (1-\delta(m_t))k_t}{k_t} - z\right)$ is the marginal adjustment cost of investment. The capital tax creates a wedge between the marginal benefit of utilization on the left-hand-side of this condition and the marginal cost of utilization on the right-hand-side. An increase in τ_K, everything else constant, reduces the utilization rate.[w] Intuitively, a higher capital tax reduces the after-tax marginal benefit of utilization, and thus reduces the rate of utilization. Note also that the magnitude of this distortion depends on where the capital stock is relative to its steady state, because the sign of Φ_t depends on Tobin's Q, which is given by $Q_t = 1 + \Phi_t$. If $Q_t > 1$ ($\Phi_t > 0$), the desired investment rate is higher than the steady-state investment rate. In this case, $Q_t > 1$ increases the marginal cost of utilization (because higher utilization means faster depreciation, which makes it harder to attain the higher target capital stock). The opposite happens if $Q_t < 1$ ($\Phi_t < 0$). In this case, the faster depreciation at higher utilization rates makes it easier to

[w] This follows from the concavity of the production function and the fact that $\delta(m_t)$ is increasing and convex.

run down the capital stock to reach its lower target level. Thus, an increase in τ_K induces a larger decline in the utilization rate when the desired investment rate is higher than its long-run target (ie, $\Phi_t > 0$).

The interaction of endogenous utilization and the limited depreciation allowance plays an important role in this setup. Endogenous utilization means that the government cannot treat the existing (predetermined) k as an inelastic source of taxation, because effective capital services decline with the capital tax rate even when the capital stock is already installed. This weakens the revenue-generating capacity of capital taxation, and it also makes capital taxes more distorting, since it gives agents an additional margin of adjustment in response to capital tax hikes (ie, capital taxes increase the post-tax marginal cost of utilization, as shown in eq. 9). The limited depreciation allowance widens the base of the capital tax, but it also strengthens the distortionary effect of τ_K by reducing the post-tax marginal return on capital (see eq. 8). As we show in the quantitative results, the two mechanisms result in a dynamic Laffer curve with a standard bell shape and consistent with empirical estimates of the capital tax base elasticity, while removing them results in a Laffer curve that is nearly-linearly increasing for a wide range of capital taxes.

The cross-country externalities from tax changes work through three distinct transmission channels that result from the tax distortions discussed in the previous paragraphs. First, relative prices, because national tax changes alter the prices of financial assets (including internationally traded assets and public debt instruments) as well as the rental prices of effective capital units and labor. Second, the distribution of wealth across the regions, because efficiency effects of tax changes by one region affect the allocations of capital and net foreign assets across regions (even when physical capital is not directly mobile). Third, the erosion of tax revenues, because via the first two channels the tax policies of one region affect the ability of the other region to raise tax revenue. When one region responds to a debt shock by altering its tax rates, it generates external effects on the other region via these three channels. Given the high degree of financial and trade integration in the world economy today, abstracting from these considerations in quantitative estimates of the effects of fiscal policy is a significant shortcoming.

3.2 Calibration to Europe and the United States

We use data from the United States and the 15 largest European countries to calibrate the model at a quarterly frequency.[x] We calibrate the home region (US) to the United States, and the foreign region (EU15) to the aggregate of the 15 European countries. The EU15 aggregates are GDP-weighted averages. Table 5 presents key macroeconomic statistics and fiscal variables for the all the countries and the two region aggregates in 2008.

[x] The European countries include Austria, Belgium, Denmark, Finland, Greece, France, Germany, Ireland, Italy, the Netherlands, Poland, Portugal, Spain, Sweden, and the United Kingdom. These countries account for over 94% of the European Union's GDP.

Table 5 Macroeconomic stance as of 2008

	EU15											GDP-weighted ave.		
	AUT	**BEL**	**DEU**	**ESP**	**FRA**	**GBR**	**ITA**	**NLD**	**POL**	**SWE**	**Other**	**EU15**	**US**	**All**
(a) Macro aggregates														
τ_C	0.19	0.17	0.17	0.12	0.17	0.14	0.13	0.20	0.21	0.26	0.23	0.17	0.04	0.11
τ_L	0.51	0.47	0.41	0.35	0.45	0.30	0.48	0.47	0.38	0.55	0.39	0.41	0.27	0.35
τ_K	0.25	0.45	0.24	0.25	0.38	0.40	0.38	0.26	0.16	0.37	0.31	0.32	0.37	0.34
c/y	0.53	0.52	0.56	0.57	0.57	0.64	0.59	0.45	0.62	0.47	0.58	0.57	0.68	0.62
x/y	0.22	0.24	0.19	0.29	0.22	0.17	0.21	0.20	0.24	0.20	0.23	0.21	0.21	0.21
g/y	0.19	0.23	0.18	0.19	0.23	0.22	0.20	0.26	0.19	0.26	0.21	0.21	0.16	0.19
tb/y	0.06	0.01	0.06	−0.06	−0.02	−0.02	−0.01	0.08	−0.04	0.07	−0.02	0.00	−0.05	−0.02
Rev/y	0.48	0.49	0.44	0.37	0.50	0.42	0.46	0.47	0.40	0.54	0.45	0.45	0.32	0.39
$\text{Total Exp}/y$	0.49	0.50	0.44	0.41	0.53	0.47	0.49	0.46	0.43	0.52	0.48	0.47	0.39	0.43
(b) Debt shocks														
d_{2007}/y_{2007}	0.31	0.73	0.43	0.18	0.36	0.28	0.87	0.28	0.17	−0.23	0.13	0.38	0.43	0.40
d_{2011}/y_{2011}	0.45	0.80	0.51	0.46	0.63	0.62	1.00	0.38	0.32	−0.25	0.45	0.58	0.74	0.65
$\Delta d/y$	0.14	0.07	0.09	0.28	0.27	0.33	0.14	0.10	0.15	−0.02	0.32	0.20	0.31	0.25

Other is a GDP-weighted average of Denmark, Finland, Greece, Ireland, and Portugal.

Source: OECD Revenue Statistics, OECD National income Accounts, and EuroStat. Tax rates are author’s calculations based on Mendoza, E.G., Razin, A., Tesar, L.L. 1994. Effective tax rates in macroeconomics: cross-country estimates of tax rates on factor incomes and consumption. J. Monet. Econ. 34 (3), 297323. “Total Exp” is total noninterest government outlays.

The first three rows of Table 5 show estimates of effective tax rates on consumption, labor, and capital calculated from revenue and national income accounts statistics using the methodology originally introduced by Mendoza et al. (1994) (MRT). The United States and EU15 have significantly different tax structures. Consumption and labor tax rates are much higher in EU15 than in the United States (0.17 vs 0.04 for τ_C and 0.41 vs 0.27 for τ_L), while capital taxes are higher in the United States (0.37 vs 0.32). The labor and consumption tax rates imply a consumption-leisure tax wedge τ_W of 0.298 for the United States vs 0.496 in EU15. Thus, EU15 has much higher effective tax distortion on labor supply. Notice also that inside of EU15 there is also some tax heterogeneity, particularly with respect to Great Britain, which has higher capital tax and lower labor tax than most of the other EU15 countries.

With regard to aggregate expenditure–GDP ratios, the United States has a much higher consumption share than EU15, by 11 percentage points. EU15 has a larger government expenditure share (current purchases of goods and services, excluding transfers) than the United States by 5 percentage points. Their investment shares are about the same, at 0.21. For net exports, the United States has a deficit of 5% while EU15 has a balanced trade (with the caveat that the latter includes all trade the individual EU15 countries conduct with each other and with the rest of the world). In light of this, we set the trade balance to zero in both countries for simplicity. In terms of fiscal flows, both total tax revenues and government outlays (including expenditures and transfer payments) as shares of GDP are higher in EU15 than in the United States, by 13 and 8 percentage points, respectively. Thus, the two regions differ sharply in all three fiscal instruments (taxes, current government expenditures, and transfer payments).

The bottom panel of Table 5 reports government debt to GDP ratios and their change between end–2007 (beginning of 2008) and end–2011. These changes are our estimate of the increases in debt (or “debt shocks”) that each country and region experienced, and hence they are the key exogenous impulse used in the quantitative experiments. These debt ratios correspond to general government net financial liabilities as a share of GDP as reported in *Eurostat*. As the table shows, debt ratios between end–2007 and 2011 rose sharply for all countries except Sweden, where the general government actually has a net asset position (ie, negative net liabilities) that changed very little. The size of the debt shocks differs substantially across the two regions. The United States entered the Great Recession with a higher government debt to output ratio than EU15 (0.43 vs 0.38) and experienced a larger increase in the debt ratio (0.31 vs 0.20).

Table 6 lists the calibrated parameter values and the main source for each value. The calibration is set so as to represent the balanced-growth steady state that prevailed before the debt shocks occurred using 2008 empirical observations for the corresponding allocations. The value of ω is set at 0.46 so as to match the observation that the United States accounts for about 46% of the combined GDP of the United States and EU15 in 2008. Tax rates, government expenditure shares and debt ratios are calibrated to the values in the United States and EU15 columns of Table 5, respectively. The limit on the

Table 6 Parameter values

Preferences:		**US**	**EU15**	**Sources**
β	Discount factor	0.998		Steady state Euler equation for capital
σ	Risk aversion	2.000		Standard DSGE value
a	Labor supply elasticity	2.675		$\bar{l}=0.18$ (Prescott, 2004)
Technology:				
α	Labor income share	0.61		Trabandt and Uhlig (2011)
γ	Growth rate	0.0038		Real GDP p.c. growth of sample countries (Eurostat 1995–2011)
η	Capital adjustment cost	2		Elasticity of capital tax base (Gruber and Rauh, 2007; Dwenger and Steiner, 2012)
$\bar{m}$	Capacity utilization	1		Steady state normalization
$\delta(\bar{m})$	Depreciation rate	0.0163		Capital law of motion, $x/y=0.19$, $k/y=2.62$ (OECD, AMECO)
χ_0	$\delta(m)$ coefficient	0.023	0.024	Optimality condition for utilization given $\delta(\bar{m})$, $\bar{m}$
χ_1	$\delta(m)$ exponent	1.44	1.45	Set to yield $\delta(\bar{m})=0.0164$
ω	Country size	0.46	0.54	GDP share in all sample countries
Fiscal policy:				
g/y	Gov't exp share in GDP	0.16	0.21	OECD National Income Accounts
τ_C	Consumption tax	0.04	0.17	MRT modified
τ_L	Labor income tax	0.27	0.41	MRT modified
τ_K	Capital income tax	0.37	0.32	MRT modified
θ	Depreciation allowance limitation	0.20		$(REV_K^{corp}/REV_K)(K^{NR}/K)$, OECD Revenue Statistics and EU KLEMS

Note: The implied growth adjusted discount factor $\tilde{\beta}$ is 0.995, and the implied precrisis annual interest rate is 3.8%. REV_K^{corp}/REV_K is the ratio of corporate tax revenue to total capital tax revenue. K^{NR}/K is the ratio of nonresidential fixed capital to total fixed capital.

depreciation allowance, θ, is set to capture the facts that tax allowances for depreciation costs apply only to capital income taxation levied on businesses and self-employed, and do not apply to residential capital (which *is* included in k). Hence, the value of θ is set as $\theta=(REV_K^{corp}/REV_K)(K^{NR}/K)$, where (REV_K^{corp}/REV_K) is the ratio of revenue from corporate capital income taxes to total capital income tax revenue, and (K^{NR}/K) is the ratio of nonresidential fixed capital to total fixed capital. Using 2007 data from OECD *Revenue Statistics* for revenues, and from the European Union's *EU KLEMS* database for capital stocks for the ten countries with sufficient data coverage,[y] these ratios range from 0.32% to 0.5% for (REV_K^{corp}/REV_K) and from 27% to 52% for (K^{NR}/K). Weighting by

[y] These countries are Austria, Denmark, Finland, Germany, Italy, Netherlands, Spain, Sweden, the United Kingdom, and the United States.

GDP, the aggregate value of θ is 0.20. Also the value for the United States is close to the weighted value for the European countries.

The technology and preference parameters are set the same across the United States and EU15, except the parameters χ_0 and χ_1 in the depreciation function. The common parameters are calibrated to target the weighted average statistics for all sample countries. The labor share of income, α, is set to 0.61, following Trabandt and Uhlig (2011). The quarterly rate of labor-augmenting technological change, γ, is 0.0038, which corresponds to the 1.51% weighted average annual growth rate in real GDP per capita of all the countries in our sample between 1995 and 2011, based on Eurostat data. We normalize the long-run capacity utilization rate to $\bar{m}=1$. Given γ at 0.0038, x/γ at 0.19 and k/γ at 2.62 from the data, we solve for the long-run depreciation rate from the steady-state law of motion of the capital stock, $x/\gamma=(\gamma+\delta(\bar{m}))k/\gamma$.[z] This yields $\delta(\bar{m})=0.0163$ per quarter. The constant depreciation rate for claiming the depreciation tax allowance, $\overline{\delta}$, is set equal to the steady state depreciation rate of 0.0163.

The value of χ_0 follows then from the optimality condition for utilization at steady state, which yields $\chi_0=\delta(\bar{m})+\dfrac{1+\gamma-\beta}{\beta}-\tau_K\overline{\delta}$. Given this, the value of χ_1 follows from evaluating the depreciation rate function at steady state, which implies $\chi_0\bar{m}^{\chi_1}/\chi_1=\delta(\bar{m})$. Given the different capital tax rates in the United States and EU15, the implied values for χ_0 and χ_1 are slightly different across countries: χ_0 is 0.0233 in the United States and 0.0235 in EU15, and χ_1 is 1.435 in the United States and 1.445 in EU15.

The preference parameter, σ, is set at a commonly used value of 2. The exponent of leisure in utility is set at $a=2.675$, which is taken from Mendoza and Tesar (1998). This value supports a labor allocation of 18.2 h, which is in the range of the 1993–96 averages of hours worked per person aged 15–64 reported by Prescott (2004). The value of β follows from the steady-state Euler equation for capital accumulation, using the values set above for the other parameters that appear in this equation:

$$\frac{\gamma}{\tilde{\beta}}=1+(1-\tau_K)(1-\alpha)\frac{\gamma}{k}-\delta(\bar{m})+\tau_K\theta\overline{\delta}.$$

This yields $\tilde{\beta}=0.995$, and then since $\tilde{\beta}=\beta(1+\gamma)^{1-\sigma}$ it follows that $\beta=0.998$. The values of β, γ and σ pin down the steady-state gross real interest rate, $R=\beta^{-1}(1+\gamma)^{\sigma}=1.0093$. This is equivalent to a net annual real interest rate of about 3.8%.

Once R is determined, the steady-state ratio of net foreign assets to GDP is pinned down by the net exports-GDP ratio. Since we set $tb/\gamma=0$, $b/\gamma=(tb/\gamma)/\left[(1+\gamma)R^{-1}-1\right]=0$. In addition, the steady-state government budget constraint yields

[z] Investment rates are from the OECD National Income Accounts and capital-output ratios are from the AMECO database of the European Commission.

an implied ratio of government entitlement payments to GDP $e/y = Rev/y - g/y - (d/y)\left[1-(1+\gamma)R^{-1}\right] = 0.196$. Under this calibration approach, both b/y and e/y are obtained as residuals, given that the values of all the terms in the right-hand-side of the equations that determine them have already been set. Hence, they generally will not match their empirical counterparts. In particular, for entitlement payments the model underestimates the 2008 observed ratio of entitlement payments to GDP (0.196 in the model vs 0.26 in the data for All EU). Notice, however, that when the model is used to evaluate tax policies to restore fiscal solvency, the fact that entitlement payments are lower than in the data strengthens our results, because lower entitlements means a lower required amount of revenue than what would be needed to support observed transfer payments, thus making it easier to restore solvency. We show below that restoring fiscal solvency is difficult and implies nontrivial tax adjustments with sizable welfare costs and cross-country spillovers, all of which would be larger with higher government revenue requirements due to higher entitlement payments.

The value of the investment-adjustment-cost parameter, η, cannot be set using steady-state conditions, because adjustment costs wash out at steady state. Hence, we set the value of η so that the model is consistent with the mid-point of the empirical estimates of the short-run elasticity of the capital tax base to changes in capital tax rates. The range of empirical estimates is 0.1–0.5, so the target midpoint is 0.3.[aa] Under the baseline symmetric calibration, the model matches this short-run elasticity with $\eta = 2.0$. This is also in line with estimates in House and Shapiro (2008) of the response of investment in long-lived capital goods to relatively temporary changes in the cost of capital goods.[ab]

Table 7 reports the 2008 GDP ratios of key macro-aggregates in the data and the model's corresponding steady-state allocations for the US–EU15 calibration. As noted earlier, this calibration captures the observed differences in the size of the regions, their fiscal policy parameters, and their public debt-GDP ratios. Notice in particular that the consumption-output ratios and the fiscal revenue-output ratios from the data were not directly targeted in the calibration, but the two are closely matched by the model. Hence, the model's initial stationary equilibrium before the increases in public debt is a reasonably good match to the observed initial conditions in the data.

[aa] The main estimate of the elasticity of the *corporate* tax base relative to corporate taxes in the United States obtained by Gruber and Rauh (2007) is 0.2. Dwenger and Steiner (2012) obtained around 0.5 for Germany. Grubler and Rauh also reviewed the large literature estimating the elasticity of *individual* tax bases (which include both labor and capital income taxes collected from individuals) to individual tax rates and noted this: "The broad consensus…is that the elasticity of taxable income with respect to the tax rate is roughly 0.4. Moreover, the elasticity of actual income generation through labor supply/savings, as opposed to reported income, is much lower. And most of the response of taxable income to taxation appears to arise from higher income groups."

[ab] They estimated an elasticity of substitution between capital and consumption goods in the 6–14 range. In the variant of our model without utilization choice, this elasticity is equal to $1/(\eta\delta)$. Hence, for $\delta(\bar{m}) = 0.0164$, elasticities in that range imply values of η in the 1–2.5 range.

Table 7 Balanced growth allocations (GDP ratios) of 2008

	United States		EU15	
	Data	**Model**	**Data**	**Model**
c/y	0.68	0.63	0.57	0.56
i/y	0.21	0.21	0.21	0.23
g/y^*	0.16	0.16	0.21	0.21
tb/y	− 0.05	0.00	0.00	0.00
Rev/y	0.32	0.32	0.45	0.46
d/y^*	0.76	0.76	0.60	0.60

3.3 Quantitative Results

The goal of the quantitative experiments is to use the numerical solutions of the model to study whether alternative fiscal policies can restore fiscal solvency, which requires increasing the present discounted value of the primary balance in the right-hand-side of (7) by as much as the observed increases in debt.[ac] Notice that the change in this present value reflects changes in the endogenous equilibrium dynamics of the primary balance-GDP ratio in response to the changes in fiscal policy variables. In turn, the changes in primary balance dynamics reflect the effects of these policy changes on equilibrium allocations and prices that determine tax bases, and the computation of the present value reflects also the response of the equilibrium interest rates (ie, debt prices).

We conduct a set of experiments in which we assume that the United States or EU15 implement unilateral increases in either capital or labor tax rates, so we can quantify the effects on equilibrium allocations and prices, sustainable debt (ie, primary balance dynamics), and social welfare in both regions. We also compare these results with those obtained if the same tax changes are implemented assuming the countries are closed economies, so we can highlight the cross-country externalities of unilateral tax changes.

The model is solved numerically using a modified version of the algorithm developed by Mendoza and Tesar (1998, 2005), which is based on a first-order approximation to the equilibrium conditions around the steady state. Standard perturbation methods cannot be applied directly, because trade in bonds implies that, when the model's pre-debt-crisis steady state is perturbed, the equilibrium transition paths of allocations and prices, and the new steady-state equilibrium need to be solved for simultaneously.[ad] This is because

[ac] The observed increases in debt between end-2007 (beginning of 2008) and end-2011 can be viewed as exogenous increases in d_0/y_{-1} in the left-hand-side of the IGBC (7). As reported in Table 5, the US debt ratio rose by 31 percentage points from 41%, and that of the EU15 rose by 20 percentage points from 38%.

[ad] Alternative solution methods that make the interest rate or the discount factor ad-hoc functions of net foreign assets (NFA), or that assume that holding these assets is costly, are also not useful, because they impose calibrated NFA positions that cannot be affected by tax changes, whereas the "true" model without these modifications can yield substantial world redistribution of wealth as a result of tax policy changes.

in models of this class stationary equilibria depend on initial conditions, and thus cannot be determined separately from the models' dynamics. Mendoza and Tesar dealt with this problem by developing a solution method that nests a perturbation routine for solving transitional dynamics within a shooting algorithm. This method iterates on candidate values of the new long-run net foreign asset positions to which the model converges after being perturbed by debt and tax changes, until the candidate values match the positions the model converges to when simulated forward to its new steady state starting from the calibrated pre-debt-crisis initial conditions.

3.3.1 Dynamic Laffer Curves

We start the analysis of the quantitative results by constructing "Dynamic Laffer Curves" (DLC) that show how unilateral changes in capital or labor taxes in one region affect that region's sustainable public debt. These curves map values of τ_K or τ_L into the equilibrium present discounted value of the primary fiscal balance. For each value that a given tax rate in the horizontal axis takes, we solve the model to compute the intertemporal sequence of total tax revenue, which varies as equilibrium allocations and prices vary, while government purchases and entitlement payments are kept constant. Then we compute the present value of the primary balance, which therefore captures the effect of changes in the equilibrium sequence of interest rates. We take the ratio of this present value to the initial output y_{-1} (ie, GDP in the steady state calibrated to pre-2008 data) so that it corresponds to the term in the right-hand-side of the IGBC (7), and plot the result as a *change* relative to the 2007 public debt ratio. Hence, the values along the vertical axis of the DLCs show the change in d_0/y_{-1} that particular values of τ_K or τ_L can support as sustainable debt at equilibrium (ie, debt that satisfies the IGBC with equality). By construction, the curves cross the zero line at the calibrated tax rates of the initial stationary equilibrium, because those tax rates yield exactly the same present discounted value of the primary balance as the initial calibration. To make the observed debt increases sustainable, there needs to be a value of the tax rate in the horizontal axis such that the DLC returns a value in the vertical axis that matches the observed change in debt.

Since the "passive" region whose taxes are not being changed unilaterally is affected by spillovers of the other region's tax changes, there needs be an adjustment in the passive region so that its IGBC is unchanged (ie, it maintains the same present discounted value of primary fiscal balances). We refer to this adjustment as maintaining "revenue neutrality" in the passive region. In principle this can be done by changing transfers, taxes or government purchases. However, since we have assumed already that government purchases are kept constant in both regions, reducing distortionary tax rates in response to favorable tax spillovers would be more desirable than increasing transfer payments, which are nondistortionary. Hence, we maintain revenue neutrality in the passive region by adjusting the labor tax rate.

3.3.1.1 Dynamic Laffer Curves for Capital Taxes

The DLCs for capital taxes are plotted in Fig. 8. The panel (A) is for the US region, and the panel (B) is for EU15. The solid lines show the open-economy curves and the dotted lines are for when the countries are in autarky. As explained above, the DLCs intersect the zero line at the initial tax rates of $\tau_K = 0.37$ and $\tau_K^* = 0.32$ by construction. We also show in the plots the increases in debt observed in each region, as shown in Table 5: The

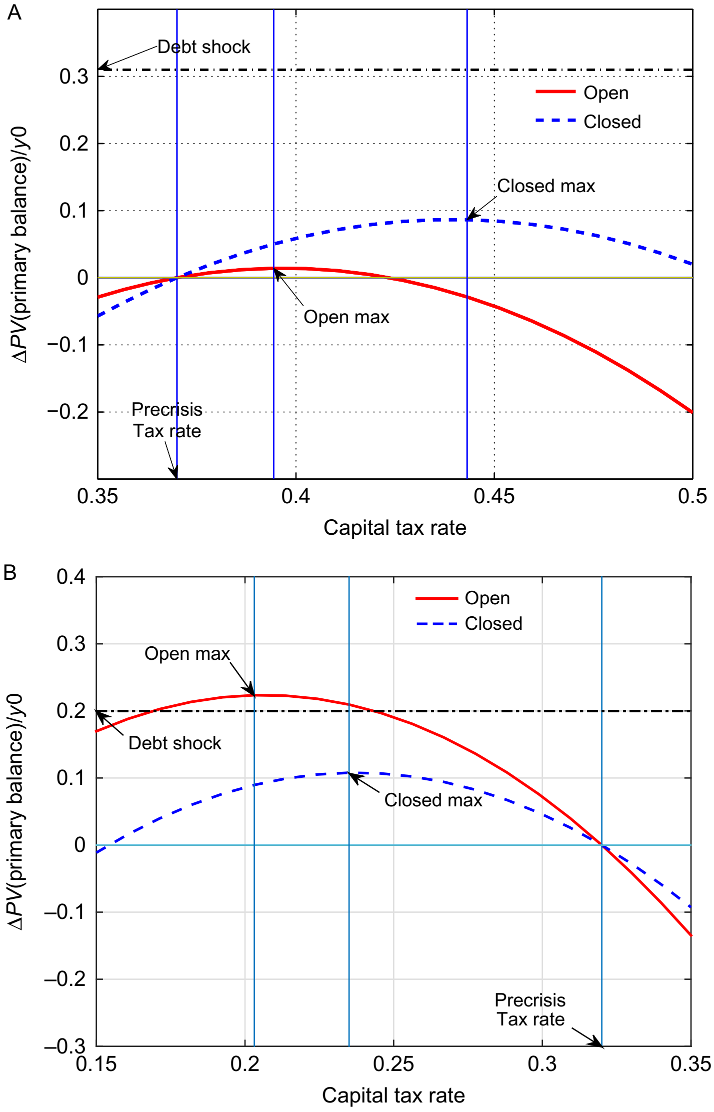


Fig. 8 Dynamic Laffer curves of capital tax rates. (A) United States. (B) EU15.

US net public debt ratio rose 31 percentage points and that of EU15 rose 20 percentage points. These increases are marked with the "Debt Shock" line in Fig. 8.

Fig. 8 shows that the DLCs of the United States and EU15 are very different, with those for EU15 seating higher, shifted to the left, and showing more curvature than those for the United States. Hence, unilateral changes in capital tax rates show a capacity to sustain larger debt increases in EU15 than in the United States, and can do so at lower tax rates. These marked differences are the result of the heterogeneity in fiscal policies present in the data and captured in the calibration, and in the open-economy scenario they are also partly explained by the international externalities of the unilateral tax changes assumed in constructing the DLCs. EU15 has higher revenue-generating capacity because of higher labor and consumption taxes at identical labor income shares and similar consumption shares, although in terms of primary balance the higher revenue is partly offset by higher government purchases. On the other hand, the United States has a lower capital tax rate and by enough to make a significant difference in the inefficiencies created by capital taxes across the two regions, as we illustrate in more detail below. Moreover, the magnitude of heterogeneity in the capital tax DLCs that results from a given magnitude of heterogeneity in fiscal variables depends on the model's modifications made to match the observed elasticity of the capital base. We illustrate below that DLCs are very different if we remove capacity utilization and the limited depreciation allowance.

Beyond the difference in position and shape of the capital tax DLCs across the United States and EU15, these DLCs deliver three striking results: First, unilateral changes in the US capital tax cannot restore fiscal solvency and make the observed increase in debt unsustainable (the peaks of the DLCs of the US region either as a closed or an open economy are significantly below the debt shock line). The maximum point of the open-economy DLC is attained at $\tau_K = 0.402$, which produces an increase in the present value of the primary balance of only 2 percentage points of GDP, far short of the required 31. In contrast, the maximum point of the open-economy DLC for EU15 is attained around $\tau_K^* = 0.21$, which rises the present value of the primary balance by 22 percentage points of GDP, slightly more than the required 20. Under autarky, however, the EU15 DLC also peaks below the required level, and hence capital taxes also cannot restore fiscal solvency for EU15 as a closed economy. This result also reflects the strong cross-country externalities that we discuss in more detail below (ie, unilateral capital tax *cuts* yield significantly more sustainable debt for EU15 as an open economy than under autarky).

Second, capital income taxes in EU15 are highly inefficient. The current capital tax rate is on the increasing segment of the DLC for the United States but on the decreasing segment for EU15. This has two important implications. One is that EU15 could have sustained the calibrated initial debt ratio of 38% at capital taxes below 15%, instead of the 32% tax rate obtained from the data. The second is that to make the observed 20 percentage points increase in debt sustainable, EU15 can *reduce* its capital tax almost in half to about 17% in the open-economy DLC. In both cases, the sharply lower capital taxes would be much less distortionary and thus would increase efficiency significantly.

Third, cross-country externalities of capital income taxes are very strong, and under our baseline calibration, they hurt (favor) the capacity to sustain debt of the United States (EU15). For the United States, the DLC under autarky is steeper than in the open-economy case, and it peaks at a higher tax rate of 43% and with a higher increase in the present value of the primary balance of about 10 percentage points. Thus, the United States can always sustain more debt, or support higher debt increases relative to the calibrated baseline, for a given increase in τ_K under autarky than as an open economy. This occurs because by increasing its capital tax *unilaterally* as an open economy the United States not only suffers the efficiency losses in capital accumulation and utilization, but it also triggers reallocation of physical capital from the United States to EU15, which results in reductions (increases) in the United States (EU15) factor payments and consumption, and thus lower (higher) tax bases in the United States (EU15). The same mechanism explains why reducing the capital tax in EU15 unilaterally generates much less revenue under autarky than in the open-economy case. In the latter, cutting the EU15 capital tax unilaterally triggers the same forces as a unilateral increase in the US capital tax.

This quantitative evidence of strong externalities of capital taxes across financially integrated economies demonstrates that evaluating “fiscal space,” or the capacity to sustain debt, using closed-economy models leads to seriously flawed estimates of the effectiveness of capital taxes as a tool to restore debt sustainability. The results also suggest that incentives for strategic interaction leading to capital income tax competition are strong, and get stronger as higher debts need to be reconciled with fiscal solvency (as evidenced by the history of corporate tax competition inside the EU since the 1980s). Mendoza et al. (2014) study this issue using a calibration that splits the European Union into two regions, one including the countries most affected by the European debt crisis (Greece, Ireland, Italy, Portugal, and Spain) and the second including the rest of the Eurozone members.

3.3.1.2 Dynamic Laffer Curves of Labor Tax Rates

Fig. 9 shows the DLCs for the labor tax rate. Notice that the open-economy and autarky DLCs are similar within each region (although more similar for EU15 than for the United States), which indicates that international externalities are much weaker in this case. This is natural, because labor is an immobile factor, and although it can still trigger cross-country spillovers via general-equilibrium effects, these are much weaker than the first-order effects created by unilateral changes of capital taxes via the condition that arbitrages after-tax returns on all assets across countries.

The main result of the DLCs for labor taxes is that the DLCs for the United States are much higher than those for EU15. Since the international externalities are weak for the

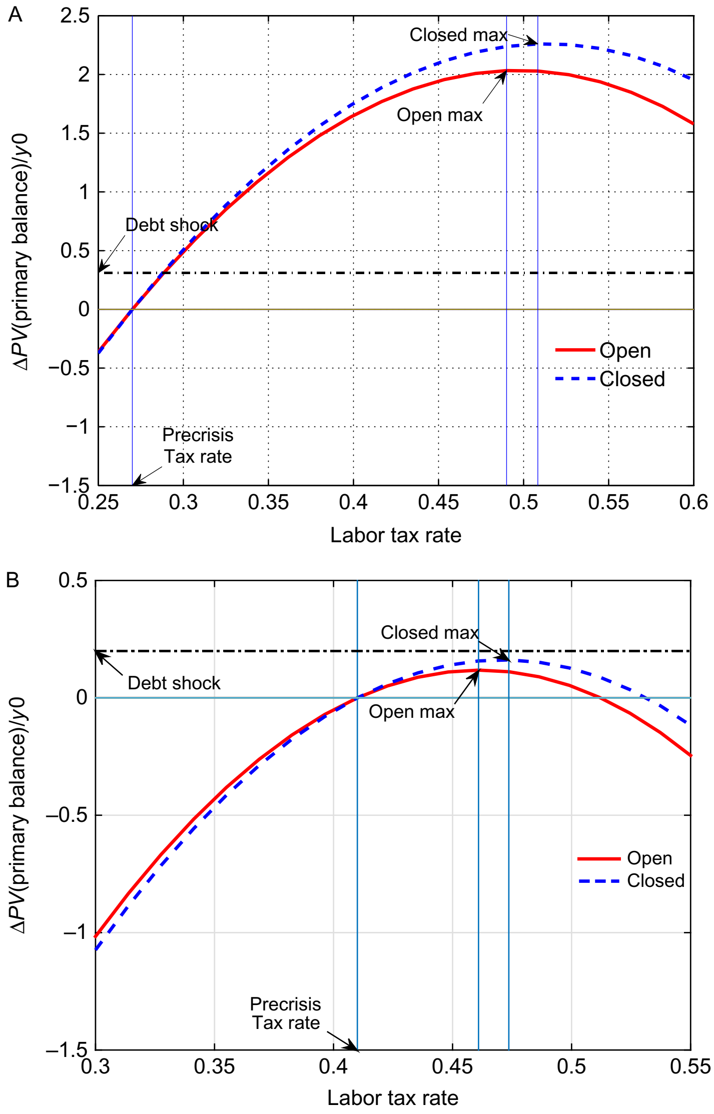


Fig. 9 Dynamic Laffer curves of labor tax rates. (A) United States. (B) EU15.

labor tax, this result is only due to the different initial conditions resulting from the fiscal heterogeneity captured in our calibration, and in particular to the large differences in initial labor and consumption taxes (41 vs 27% for labor and 17 vs 4% for consumption in EU15 vs the United States, respectively). Increasing the calibrated τ_L for the US region to the EU15 rate of 41%, keeping all other US parameters unchanged, shifts down its labor tax DLC almost uniformly by about 200 percentage points in the 0.25–0.55 interval of labor tax rates. This happens because, for an increase in the labor tax of a given size, the difference in initial conditions implies that the US region generates a larger increase in the present value of total tax revenue than EU15, and since the present value of government outlays is nearly unchanged in both, the larger present value of revenue is amplified into a significantly larger increase in the present value of the primary balance.[ae]

The US open-economy DLC for τ_L is considerably steeper than for τ_K, and it peaks at a tax rate of 0.48, which would make sustainable an initial debt ratio larger than in the initial baseline by 200 percentage points of GDP, much more than the 31 percentage points required by the data. The labor tax rate that the United States as an open or closed economy needs to make the observed debt increase sustainable is about 29%, which is just a two-percentage-point increase relative to the initial tax rate. Hence, these results show that, from the perspective of macroeconomic efficiency that representative-agent models of financially integrated economies like the one we are using emphasize, labor taxes are a significantly more effective tool for restoring fiscal solvency in the United States than capital taxes.

The DLC of EU15 yields much less positive results. Since the initial consumption-labor wedge is already much higher in this region than in the United States, the fiscal space of the labor tax rate is very limited. In either the closed- or open-economy cases, the DLC peaks at a labor tax rate of 46% and yields an increase of only about 10 percentage points in the present value of the primary balance, which is half of the 20-percentage-points increase EU15 needs make the observed debt increase sustainable.

It is interesting to note that the debt increase in the United States was about 10 percentage points larger than in Europe, yet the model predicts that given the initial conditions in tax rates and government outlays before the increases in debt, unilateral tax adjustments in Europe cannot generate a sufficient increase in the present value of the primary balance to make their higher debt sustainable. The exception is the capital tax in the open-economy scenario, in which this is possible only because EU15 would benefit significantly from a negative externality on the US region. In contrast, the results show that a modest increase in labor taxes (or consumption taxes since they are equivalent in this model) can restore fiscal solvency in the United States.

It is useful to compare the results we reported here with those of similar exercises in other existing studies based on Neoclassical models, particularly those by Trabandt and

[ae] The percent change in the present value of the primary balance after a tax change of a given magnitude relative to before (assuming that the present value of government outlays does not change) can be expressed as $z[1 + PDV(g + e)/PDV(pb)]$, where $z$is the percent change in the present value of tax revenues after the tax change relative to before, and $PDV(g + e)$ and $PDV(pb)$ are the pretax-change present values of total government outlays and the primary balance, respectively. Hence, for $z > 0$ and since total outlays are much larger than the primary balance $[PDV(g + e)/PDV(pb)] >> 1$, a given difference in z across the United States and EU15 translates into a much larger percent difference in the present value of the primary balance.

Uhlig (2011, 2012) and Auray et al. (2013). Trabandt and Uhlig (2011, 2012) used a closed-economy model without endogenous capacity utilization and focused mainly on steady-state Laffer curves (ie, Laffer curves that map tax rates into steady state tax revenues), while the DLCs studied here are for present values taking into account both transitional dynamics and steady-state changes caused by tax changes relative to the calibrated tax rates. Qualitatively, the results in Trabandt and Uhlig (2011) are similar to the ones in this chapter because they find that capital tax hikes generate much smaller increases in revenue than labor taxes. They find that the maximum increases in steady-state tax revenue obtained with capital (labor) taxes are 6 (30)% for the United States and 1 (8)% for Europe. Quantitatively; however, the results reported here differ not only because both transitional dynamics and steady-states are included, but also because the two-country model with capacity utilization captures the cross-country externalities of tax policy and the observed elasticity of the capital tax base, and these two features undermine the revenue-generating capacity of tax hikes.

Trabandt and Uhlig (2012) extend their analysis to gauge the sustainability of observed debt levels in response to hypothetical permanent increases in interest rates. Keeping government transfers, total outlays and debt constant at observed levels, they calculate the maximum real interest rate at which the revenue generated at the peak of steady-state Laffer curves would satisfy the steady-state government budget constraint. That is, effectively they compute the interest rate at which the Blanchard ratio of the previous section holds with debt and spending set at observed levels and tax revenue set at the maxima of steady-state Laffer curves. They find that the maximum real interest rate for the United States is larger than for European countries if labor taxes are moved to the peak of the Laffer curves. These calculations, however, inherit the limitations of the Blanchard ratios as measures of sustainable debt discussed in the previous section, and imply unusually large primary fiscal surpluses. For instance, depending on the debt measure used, Trabandt and Uhlig estimate the maximum interest rate for the United States in the 12–15.5% range. With a 92% debt ratio, a 1.5% annualized output growth rate and the 12% interest rate, the US economy requires a 9.6% steady-state primary surplus. The largest primary surplus observed in US history using Bohn's historical dataset starting in 1790 was 6.3%, and the average was just 0.4%. Moreover, moving the labor tax to the peak of the Laffer curve reduces steady-state output by 27%, which suggests that the welfare cost of the tax hike is quite large.

Auray et al. (2013) use a Neoclassical model of a small open economy to conduct a quantitative comparison of tax policies aimed at lowering European debt ratios. They introduce a FRF in the class of the ones examined in the previous section: Increases of the debt ratio at date t above its date-t target induce increases in the date-t primary surplus above its date-t target. The primary balance adjustment is obtained by adjusting one of the tax rates as needed to satisfy the FRF. In this environment, lowering the debt ratio requires higher tax rates in the short term in exchange for lower rates in the long term as steady-state debt service falls. They find that a cut of 10 percentage points in the debt ratio can be attained with an increase in welfare using the capital income tax, roughly no change in welfare using the consumption tax, and a welfare loss using the labor income tax. Qualitatively, the model studied here would produce similar results if applied to a similar debt-reduction experiment. Since the capital income tax is highly distorting, using the benefit of the lower debt service burden to cut the capital income tax would be best for welfare and efficiency. Their setup, however, is not calibrated to match the capital tax base elasticity and abstracts from cross-country externalities because of the small-open-economy assumption.

3.3.2 Macroeconomic Effects of Tax Rate Changes

We analyze next the macroeconomic effects of unilateral changes in capital and labor tax rates. In the first experiment, the United States increases its capital tax rate from the initial value of 0.37 to 0.402, which is the maximum point of the open-economy DLC for the United States. Table 8 shows the effects of this change on both regions in the open-economy model and on the US region as a closed economy. EU15 reduces its labor tax rate from 0.41 to 0.40 to maintain revenue neutrality, which is the result of favorable externalities from the tax hike in the United States.

The capital tax hike in the United States as an open economy leads to an overall welfare cost of 2.19% vs 2.22% as a closed economy, while EU15 obtains a welfare gain of 0.74%.[af] Comparing the US outcomes as an open economy relative to the closed economy under the same 40.2% capital tax rate, we find that the sustainable debt (ie, the present value of the primary balance) rises by a factor of 4.5 (from 1.37% to 6.16%). The welfare loss is nearly the same (2.2%), but normalizing by the amount of revenue generated, the United States is much better off in autarky. Thus, seen from this perspective, the United States would have strong incentives for either engaging in strategic interaction (ie, tax competition) or for considering measures to limit international capital mobility.

The 0.74% welfare gain that EU15 obtains from the US unilateral capital tax hike is a measure of the normative effect of the cross-country externalities of capital tax changes. The United States can raise more revenue by increasing τ_K along the upward-sloping region of its DLC, but its ability to do so is significantly hampered by the adverse externality it faces due to the erosion of its tax bases. In EU15, the same externality indirectly improves government finances, or reduces the distortions associated with tax collection, and provides it with an unintended welfare gain.

[af] Welfare effects are computed as in Lucas (1987), in terms of a percent change in consumption constant across all periods that equates lifetime utility under a given tax rate change with that attained in the initial steady state. The overall effect includes transitional dynamics across the pre- and post-tax-change steady states, as well as changes across steady states. The steady-state effect only includes the latter.

Table 8 Macroeconomic effects of an increase in US capital tax rate (the EU15 maintains revenue neutrality with labor tax)

	Open economy				**Closed economy**	
	United States		**EU15**		**United States**	
Tax rates	**Old**	**New**	**Old**	**New**	**Old**	**New**
τ_K	0.37	0.40	0.32	0.32	0.37	0.40
τ_C	0.04	0.04	0.17	0.17	0.04	0.04
τ_L	0.27	0.27	0.41	0.40	0.27	0.27
PV of fiscal deficit over precrisis GDP as percentage point change from original ss		1.37		0.00		6.16
Welfare effects (percent)						
Steady-state gain		−2.27		0.59		−2.55
Overall gain		−2.19		0.74		−2.22
Percentage changes	**Impact effect**	**Long-run effect**	**Impact effect**	**Long-run effect**	**Impact effect**	**Long-run effect**
y	−1.23	−3.87	−0.15	1.25	−2.35	−3.57
c	−1.87	−2.83	1.44	1.28	−1.53	−2.91
k	0.00	−7.61	0.00	1.25	0.00	−7.32
Percentage point changes						
tb/y	3.21	−0.30	−2.70	0.24		
i/y	−3.01	−1.02	1.77	0.00	−0.91	−1.02
r	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
l	0.11	−0.17	−0.01	0.21	−0.13	−0.11
m	−4.23	−0.866	−0.315	−0.000	−5.277	−0.866

The impact and long-run effects on key macro-aggregates in both regions are shown in the bottom half of Table 8. The corresponding transition paths of macroeconomic variables as the economies move from the precrisis steady state to the new steady state are illustrated in Fig. 10. The increase in τ_K causes US capital to fall over time to a level 7.6% below the precrisis level, while EU15's capital rises to a level 1.25% above the pre-tax-change level. Capacity utilization falls at home in both the short run and the long run, which is a key component of the model capturing the reduced revenue-generating capacity of capital tax hikes when the endogeneity of capacity utilization is considered. We show later in this section that this mechanism indeed drives the elasticity of the capital tax base in the model, which matches that of the data and is higher than what standard representative-agent models of taxation show.

On impact when the United States increases its capital tax, labor increases in the United States and falls slightly in EU15, but this pattern reverses during the transition

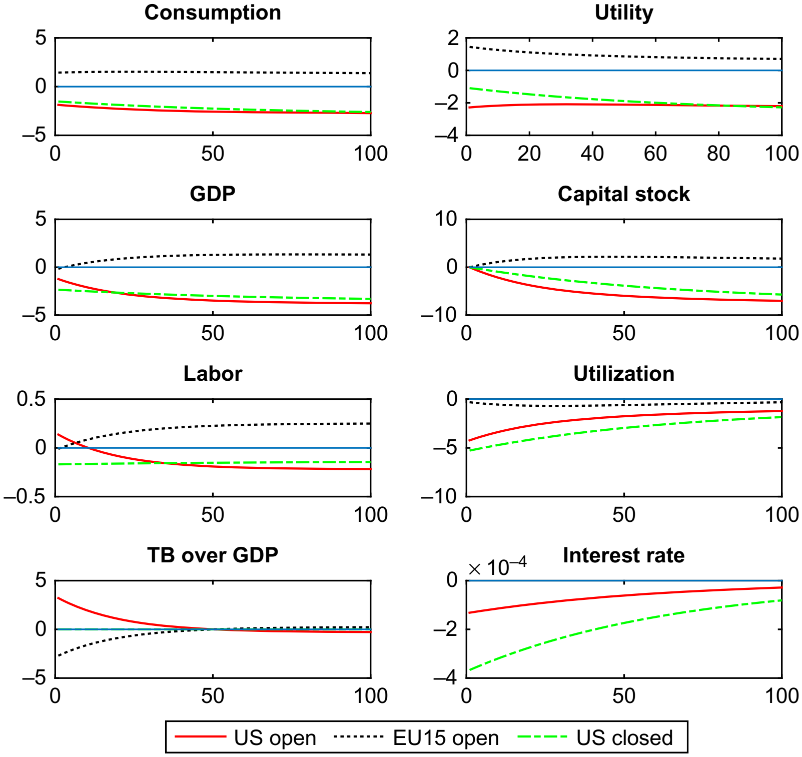


Fig. 10 Responses of macro variables to a US capital tax rate increase.

to steady state because of the lower (higher) capital stock in US (EU15) region in the new steady state. Consequently, US output contracts by almost 4% in the long-run, underscoring efficiency losses due to the capital tax increase and the costs of the fiscal adjustment. The United States increases its net foreign asset position (NFA) by running trade surpluses (tb/y) in the early stages of transition, while EU15 decreases its NFA position by running trade deficits. The US trade surpluses reflect saving to smooth out the cost of the efficiency losses, as output follows a monotonically decreasing path. Still, utility levels are lower than when the United States implements the same capital tax under autarky, because of the negative cross-country spillovers.

We next look at the responses of fiscal variables when the United States increases its capital tax, plotted in Fig. 11. In the United States, tax revenue from capital income increases almost immediately to a higher constant level when τ_K rises, while the revenues from labor and consumption taxes decline both on impact and in the long run. Labor and consumption tax rates are not changing, but both tax bases fall on impact and then decline monotonically to their new, lower steady states. The primary fiscal balance and total revenue both rise initially but then converge to about the same levels as in the precrisis stationary equilibrium. For the primary balance, this pattern is implied by the pattern of the

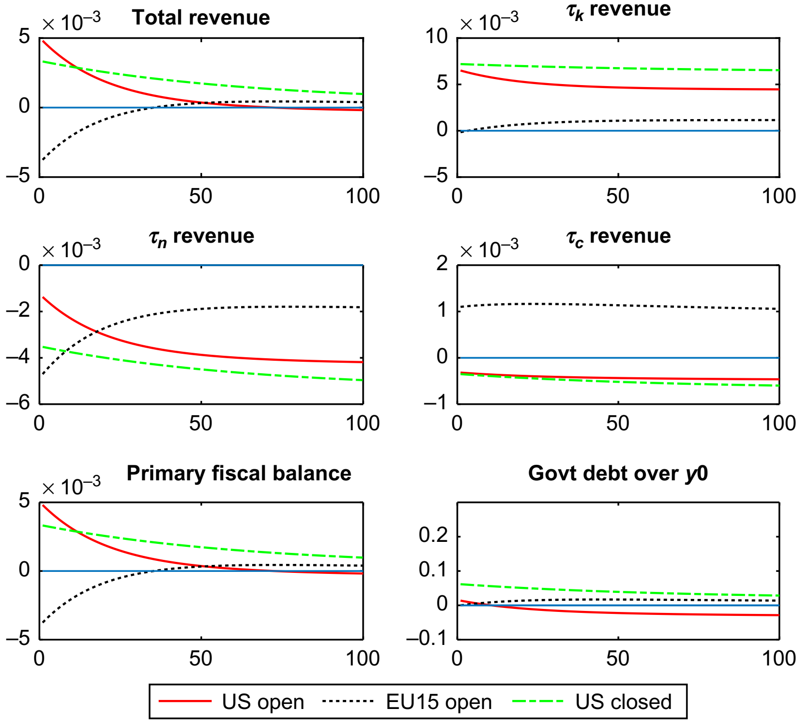


Fig. 11 Responses of fiscal variables to a US capital tax rate increase.

total revenue, since government expenditures and entitlements are held constant. For total revenue, the transitional increase indicates that the rise in capital tax revenue more than offsets the decline in the revenue from the other taxes in the transition, while in the long-run they almost offset each other exactly. This is possible because the change in τ_K to 0.4 is on the increasing side of the Laffer curve, and in fact it is the maximum point of the curve. Hence, this capital tax hike does not reduce capital tax revenues.

The public debt dynamics in the bottom-right panel of Fig. 11 shows that on impact, government debt in the United States responds to the 40% tax rate by increasing 5 percentage points, reflecting the extra initial debt that can be supported at the higher capital tax rate. Since the primary fiscal balance rises on impact and then declines monotonically, the debt ratio also falls monotonically during the transition, and converges to a ratio that is actually about 4 percentage points below the precrisis level. Hence, the initial debt increase allowed by the capital tax hike is followed by a protracted decline in debt converging to a debt ratio even lower that in the precrisis steady state. If the United States implements the same tax hike under autarky, it generates significantly larger revenues and primary balances, and hence the debt ratio increases more initially and converges to a higher steady state of 1 percentage points above the precrisis level. This is again a reflection of the cross-country externalities faced by the United States as an open economy, since equally sized tax hikes produce significantly higher revenues under autarky.

The cross-country externalities are also reflected in the fiscal dynamics of EU15 shown in Fig. 11. Maintaining revenue neutrality (in present value) still allows both its revenue and primary balance to fall initially, while in the long run both converge to very similar levels as in the precrisis steady state. Removing the labor tax adjustment in EU15 that maintains revenue neutrality, the present value of its primary balance as a share of GDP would increase by 10.1 percentage points relative to the precrisis ratio, and both its revenue and primary balances would be higher than in the plots shown in Fig. 11. The welfare gain, however, would be negligible instead of 0.74% in lifetime consumption.

The next experiment examines the effects of lowering the EU15 capital tax rate so as to move it out of the decreasing segment of the DLC. To make this change analogous to the one in the previous experiment, we change the EU15 capital tax to the value at the maximum point of the DLC for EU15, which is about 21%. Table 9 summarizes the results. The cut in the EU15 capital tax rate generates an increase of about 22 percentage points in sustainable debt (just a notch above what is required to make the observed debt increase sustainable), and a large welfare gain of 6.9% for this region. Its capital stock rises over time to a level 26% higher than in the pre-tax-change steady state. Output, consumption, labor supply, and utilization all rise in both the short-run and the long-run in EU15, while the trade balance moves initially into a large trade deficit and then converges to a small surplus. The same tax cut in EU15 as a closed economy yields a much smaller rise in sustainable debt, of just under 10 percentage points, though the welfare gain is about the same as in the open economy. This result indicates that in this case the welfare gain largely reflects the reduction of the large inefficiencies due to the initial capital tax being in the decreasing side of the DLC. In the US region, the tax cut in EU15 causes a welfare loss of 0.2%, with capital declining 1.5 percent from the pre-tax-change level.

The next two experiments focus on changes in labor tax rates. The DLCs for the labor tax rate (Fig. 9) show that the US region has substantial capacity to raise tax revenues and sustain higher debt ratios by raising labor taxes. We examine in particular an increase of the labor tax rate that completely offsets the observed debt increase, which as we noted earlier is only about 2 percentage points higher than in the initial calibration (ie, the labor tax in the United States rises from 27% to 29%). The results are reported in Table 10. The declines in US output, consumption, capital, and welfare are much smaller than with the capital tax hike. Since the international spillovers are small, this tax change produces a welfare gain of just 0.18% in EU15. For the same reason, comparing the United States results as a closed vs open economy, the change in the present value of the primary balance is almost the same, in contrast with the large difference obtained for the capital tax. Also, keep in mind that the capital tax hike, even though it was set at the maximum point of the capital tax DLC of the United States as open economy, cannot generate enough revenue to offset the observed debt increase, whereas the labor tax hike does.

Table 9 Macroeconomic effects of a decrease in EU15 capital tax rate (the United States maintains revenue neutrality with labor tax)

	Open economy				Closed economy	
	United States		EU15		EU15	
Tax rates	**Old**	**New**	**Old**	**New**	**Old**	**New**
τ_K	0.37	0.37	0.32	0.20	0.37	0.37
τ_C	0.04	0.04	0.17	0.17	0.04	0.17
τ_L	0.27	0.28	0.41	0.41	0.27	0.41
PV of fiscal deficit over precrisis GDP as percentage point change from original ss		−0.00		22.34		9.62
Welfare effects (percent)						
Steady-state gain		0.36		7.35		7.93
Overall gain		−0.23		6.86		6.99
Percentage changes	**Impact effect**	**Long-run effect**	**Impact effect**	**Long-run effect**	**Impact effect**	**Long-run effect**
y	2.30	−1.40	6.05	12.77	8.38	11.99
c	−1.59	−0.64	5.82	9.03	5.14	9.19
k	0.00	−1.50	0.00	26.10	0.00	25.23
Percentage point changes						
tb/y	8.92	−0.75	−6.57	0.56		
i/y	−5.64	0.00	8.18	3.66	3.31	3.66
r	0.00	−0.00	0.00	−0.00	0.00	−0.00
l	0.47	−0.31	0.05	0.48	0.43	0.36
m	2.34	0.00	12.93	3.31	14.94	3.31

Now consider the case of increasing the EU15 labor tax. As explained earlier in discussing the labor DLCs, the EU15 initial consumption/labor wedge is already high, so the capacity for raising tax revenues using labor taxes is limited. In this experiment, we increase the labor tax in EU15 to the rate at the maximum point of the labor tax DLC of EU15 as an open economy, which implies a labor tax rate of 0.465. The results are summarized in Table 11. The higher EU15 labor tax increases the present value of the primary balance-GDP ratio by only 0.118, falling well short of the observed debt increase of 0.2. The welfare loss is large, at nearly 5%, with output, consumption, capital, and labor falling. EU15 can produce a higher present value of the primary balance (0.16) in the closed economy at a similar welfare loss. Again the international spillover for the labor tax rate is small, so the US region makes a negligible welfare gain.

Taken together these findings are consistent with two familiar results from tax analysis in representative-agent models, which emphasize the efficiency costs of tax distortions. First, the capital tax rate is the most distorting tax. Second, in open-economy models,

Table 10 Macroeconomic effects of an increase in the US labor tax rate (the EU15 maintains revenue neutrality with labor tax)

	United States		**EU15**		**United States**	
Tax rates	**Old**	**New**	**Old**	**New**	**Old**	**New**
τ_K	0.37	0.37	0.32	0.32	0.37	0.37
τ_C	0.04	0.04	0.17	0.17	0.04	0.04
τ_L	0.27	0.29	0.41	0.41	0.27	0.29
PV of fiscal deficit over precrisis GDP as percentage point change from original ss		31.00		0.00		31.95
Welfare effects (percent)						
Steady-state gain		−0.92		0.15		−0.98
Overall gain		−0.90		0.18		−0.91
Percentage changes	**Impact effect**	**Long-run effect**	**Impact effect**	**Long-run effect**	**Impact effect**	**Long-run effect**
y	−1.16	−1.75	−0.02	0.30	−1.41	−1.68
c	−1.88	−2.09	0.34	0.31	−1.80	−2.10
k	0.00	−1.75	0.00	0.30	0.00	−1.68
Percentage point changes						
tb/y	0.72	−0.07	−0.61	0.06		
i/y	−0.46	0.00	0.40	0.00	0.02	−0.00
r	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
l	−0.29	−0.35	0.00	0.05	−0.35	−0.34
m	−0.73	0.00	−0.06	−0.00	−0.96	0.00

taxation of a mobile factor (ie, capital) yields less revenue at greater welfare loss than taxation of the immobile factor (ie, labor). This is in line with our results showing that the cross-country tax externalities are strong for capital taxes but weak for labor taxes.

The sharp differences we found between the United States and EU15 also have important policy implications in terms of debates about debt-sustainability and the effects of fiscal adjustment via capital and labor taxes in Europe and the United States. With capital taxes, the model suggests that the United States is on the increasing side of the Laffer curve, though it cannot restore fiscal solvency for the observed debt shock of 31 percentage points (neither as an open economy nor as a closed economy). In contrast, the model suggests that Europe is on the decreasing side of the Laffer curve, and can make its observed debt increase of 20 percentage point sustainable by reducing its capital taxes and moving away from the decreasing side of the Laffer curve, and in the process make a substantial welfare gain. This is only possible, however, because the United States is assumed to maintain its capital tax rate unchanged as Europe's drops, which results in

Table 11 Macroeconomic effects of an increase in the EU15 labor tax rate (the United States maintains revenue neutrality with labor tax)

	United States		EU15		EU15	
Tax rates	**Old**	**New**	**Old**	**New**	**Old**	**New**
τ_K	0.37	0.37	0.32	0.32	0.37	0.37
τ_C	0.04	0.04	0.17	0.17	0.04	0.17
τ_L	0.27	0.27	0.41	0.47	0.27	0.47
PV of fiscal deficit over precrisis GDP as percentage point change from original ss		0.00		11.75		16.02
Welfare effects (percent)						
Steady-state gain		−0.12		−5.04		−5.19
Overall gain		0.07		−4.91		−4.92
Percentage changes	**Impact effect**	**Long-run effect**	**Impact effect**	**Long-run effect**	**Impact effect**	**Long-run effect**
y	−0.68	0.41	−4.28	−6.20	−5.06	−5.99
c	0.45	0.16	−7.35	−8.18	−7.13	−8.22
k	0.00	0.41	0.00	−6.20	0.00	−5.99
Percentage point changes						
tb/y	−2.47	0.22	2.16	−0.20		
i/y	1.64	−0.00	−1.29	−0.00	0.11	−0.00
r	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
l	−0.14	0.08	−0.90	−1.05	−1.04	−1.01
m	−0.67	−0.00	−2.87	0.00	−3.59	−0.00

large externalities that benefit Europe at the expense of the United States. Capital tax hikes under autarky cannot restore fiscal solvency for Europe either.

With labor taxes, although the model indicates that both the United States and Europe are on the increasing side of their DLCs, the US pre-2008 started with a much smaller consumption/labor distortion than Europe. As a result, the United States has substantial fiscal space to easily offset the debt increase with a small labor tax hike and a small welfare cost of 0.9%. In contrast, the model suggests that Europe cannot restore fiscal solvency after the observed increase in debt using labor taxes.

3.3.3 Why Are Utilization and Limited Depreciation Allowance Important?

As explained earlier, we borrowed from Mendoza et al. (2014) the idea of using endogenous capacity utilization and a limited tax allowance for depreciation expenses to build into the model a mechanism that produces capital tax base elasticities in line with empirical estimates. In contrast, standard dynamic equilibrium models without these features tend to have unrealistically low responses of the capital base to increases in capital taxes. To illustrate this point, we follow again Mendoza et al. in comparing DLCs for capital taxes in three scenarios (see Fig. 12): (i) a standard Neoclassical model with exogenous utilization and a full depreciation allowance ($\theta = 1$), shown as a dashed-dotted line; (ii) the same model but with a limited depreciation allowance ($\theta = 0.2$), shown as a dotted line; and (iii) the baseline calibration of our model with both endogenous utilization and a limited depreciation allowance (using again $\theta = 0.2$), shown as a solid line. All other

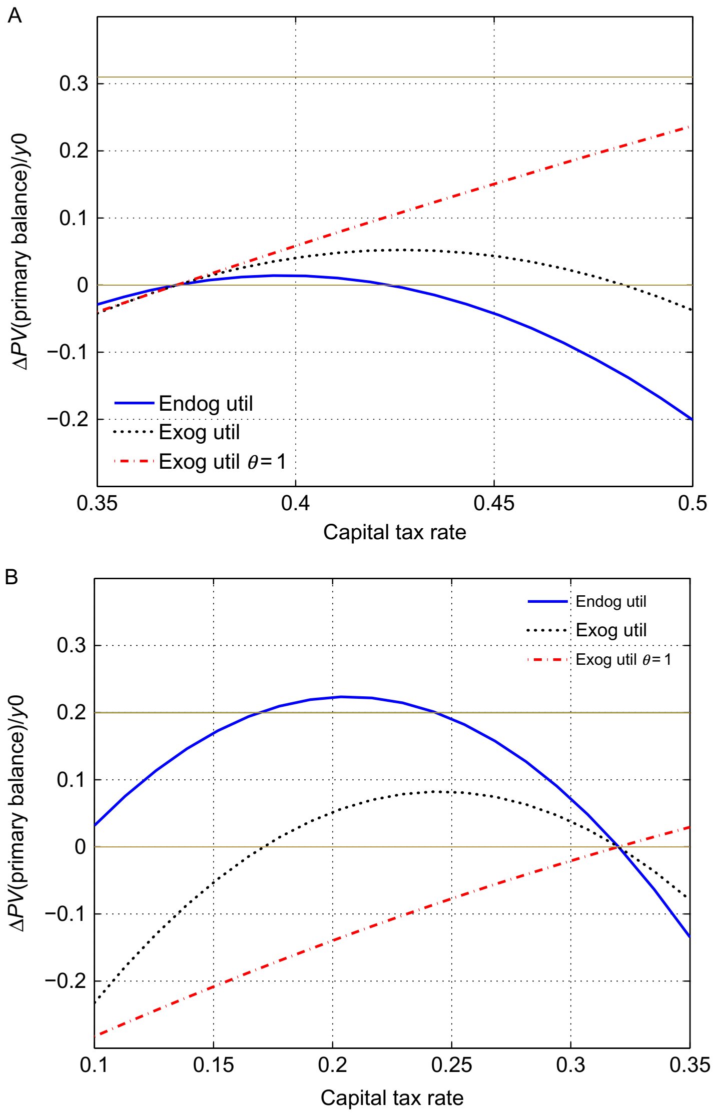


Fig. 12 Comparing dynamic Laffer curves for the capital tax rate. (A) United States. (B) EU15.

parameter values are kept the same. We show the three cases for the United States and EU15 region in panels (A) and (B) of the figure, respectively.

The DLCs for the three cases intersect at the initial calibrated tax rates of 0.37 and 0.32 for the United States and EU15 by construction. To the right of this point, the curves for case (i) are always above the other two, and the ones for case (ii) are always above the ones for case (iii). The opposite occurs to the left of the intersection points.

Consider the US plots. In case (i), the DLC has a positive, approximately linear slope in the 0.35–0.5 domain of capital tax rates. This curve continues to be increasing even when we extend the capital tax rate to 0.9, which is in line with the results obtained by Trabandt and Uhlig (2011).[ag] This behavior of the DLC for the capital tax follows from the fact that at any given date the capital stock is predetermined and has a low short-run elasticity. As a result, the government can raise substantial revenue over the transition period because the capital stock declines only gradually. The increased tax revenue during the transition dominates the fall in the steady-state, resulting in a nondecreasing DLC (recall the DLC is based on present value calculations).

Introducing limited depreciation allowance without endogenizing the utilization choice (case (ii)) has two effects that induce concavity in the DLC. First, it increases the effective rate of taxation on capital income, and thus weakens the incentive to accumulate capital and lowers the steady-state capital-output ratio and tax bases. On the other hand, it has a positive impact on revenue by widening the capital tax base. The first effect dominates the latter when the capital tax rate rises relative to the initial tax of 0.37, resulting in sharply lower DLC curve values than in case (i).

In case (iii) the tax allowance is again limited but now capacity utilization is endogenous. This introduces additional effects that operate via the distortions on efficiency and the ability to raise revenue discussed earlier: On the side of tax distortions, equation (9) implies that endogenous utilization adds to the efficiency costs of capital income taxation by introducing a wedge between the marginal cost and benefits of capital utilization. On the revenue side, endogenous utilization allows agents to make adjustments in effective capital (reducing it when taxes rise and increasing it when it falls), and thus alters the amount of taxable capital income. Hence, when utilization falls in response to increases in capital tax rates, it also weakens the government's ability to raise capital tax revenue. These effects lead to a bell-shaped DLC that has more curvature and is significantly below those in cases (i) and (ii). Thus, endogenous utilization makes capital taxes more distorting and weakens significantly the revenue-generating capacity of capital taxes.[ah]

[ag] They find that present-value Laffer curves of capital tax revenue peak at very high tax rates (discounting with the constant steady state interest rate) or have a positive slope over the full range (discounting with equilibrium interest rates).

[ah] Mendoza et al. also found that removing the limited depreciation allowance from case (iii) still results in a DLC below those of cases (i) and (ii), but it is also flatter and increasing for a wider range of capital taxes than case (iii).

Panel (B) of Fig. 12 shows DLCs for the three cases in the EU15 region. The results are analogous to Panel (A) but emphasizing now the region to the left of the intersection point, which is at the initial tax of 32%. In case (i), again the DLC has an increasing positive slope over a large range of the capital tax rate. Case (ii) shows that limiting the depreciation allowance again induces concavity in the DLC, with the EU15 initial capital tax already in the decreasing segment of the curve. Comparing with case (iii), the exogenous utilization case generates much less revenue. As in the US results, this occurs because with endogenous utilization, reductions in capital taxes lead to higher utilization rates that result in higher levels of capital income and higher wages, thus widening the two income tax bases.

The effects of endogenous utilization and limited depreciation have significant implications for the elasticity of the capital income tax base with respect to the capital tax. In particular, as Mendoza et al. (2014) showed, the model can be calibrated to match a short-run elasticity consistent with empirical estimates because of the combined effects of those two features. As documented earlier, the empirical literature finds estimates of the short-run elasticity of the capital tax base in the 0.1–0.5 range. Table 12 reports the model's comparable elasticity estimates and the effects on output, labor, and utilization 1 year after a 1% increase in the capital tax (relative to the calibrated baseline values), again for cases (i), (ii), and (iii) and in both the United States and EU15 regions.

The United States and EU15 results differ somewhat quantitatively, but qualitatively they make identical points: The neoclassical model with or without limited depreciation allowance (cases (i) and (ii)) yields short-run elasticities with the wrong sign (ie, the capital tax base *rises* in the short run in response to capital tax rate increases). The reason is that capital does not change much, since capital is predetermined in the period of the tax hike and changes little in the first period after because of investment adjustment costs, and

Table 12 Short-run elasticity of US capital tax base

	Elasticity	y_1	l_1	m_1
Empirical estimates	[0.1, 0.5]			
Model implications for the United States				
Exog. utilization and $\theta = 1$	−0.09	0.04%	0.011	
Exog. utilization and $\theta = 0.2$	−0.09	0.08%	0.028	
Endog. utilization and $\theta = 0.2$	0.29	−0.15%	0.010	−0.471
Model implications for the EU15				
Exog. utilization and $\theta = 1$	−0.04	0.01%	0.004	
Exog. utilization and $\theta = 0.2$	−0.02	0.03%	0.008	
Endog. utilization and $\theta = 0.2$	0.32	−0.14%	0.004	−0.393

Note: Elasticity is measured as the percentage decrease of capital tax base in the first year after a 1% increase in the capital tax rate is introduced. For empirical estimates, see Gruber and Rauh (2007) and Dwenger and Steiner (2012). y_1 and m_1 provides the percent deviation from the initial steady state in the impact year. l_1 denotes the percentage points change from the initial steady state.

labor supply rises due to a negative income shock from the tax hike. Since capital does not fall much and labor rises, output rises on impact, and thus taxable labor and capital income both rise, producing an elasticity of the opposite sign than that found in the data. In contrast, the model with endogenous utilization (case (iii)), generates a decline in output on impact due to a substantial drop in the utilization rate, despite the rise in labor supply. With the calibrated values of η, the model generates short-run elasticities of 0.29 and 0.32 for the United States and EU15, respectively, which are both well inside the range of empirical estimates.

It is also worth noting that with exogenous utilization, the model can produce a capital tax base elasticity in line with empirical evidence only if we set η to an unrealistically low value. The short-run elasticity of the capital tax base is negative for any $\eta > 1$, and it becomes positive and higher than 0.1 only for $\eta < 0.1$.[ai] This is significantly below the empirically relevant range of 1–2.5 documented in the calibration section. Moreover, at the value of $\eta = 2$ determined in our baseline calibration, the model without utilization choice yields a capital tax base elasticity of -0.09.

3.3.4 Further Considerations

We close this section with some important considerations and caveats of the structural analysis. In particular, we discuss the predictions of the structural framework for the case of Japan, which is challenging because of its high debt ratio, and the implications of considering the possibility of taxes on wealth or the capital stock.

Japan had a very high public debt to GDP ratio already before the global financial crisis, at about 82% by the end of 2007. By the end of 2011, its debt ratio had increased 46 percentage points to 128%. Hence the level and the change of Japan's debt ratio are both larger than what we saw in the United States and Europe.

What does the structural approach to debt sustainability tell us about the Japanese case? To answer this question, we reset the model so that the foreign region is now a proxy for Japan instead of EU15 and recompute the DLCs. In particular, we calibrate the foreign tax rates to match Japan's precrisis tax structure, using the same Mendoza-Razin-Tesar method we used for the United States and Europe. In 2007, Japan's capital tax rate was 39%, the labor tax rate was 31% and the consumption tax was 6%. This tax structure is similar to that of the United States. In fact, Japan's consumption-leisure tax wedge τ_W is 0.35, which is much closer to the 0.3 estimate for the United States than 0.5 for Europe. We also reset the relative country size to match the fact that Japan's GDP per capita is about 78% that of the United States. The rest of the structural parameters are kept the same as in our baseline analysis. The DLCs for Japan are shown in Fig. 13. The panel (A) is the DLC for the capital tax and the panel (B) is for the labor tax.

[ai] The intuition is simple. As η approaches zero the marginal adjustment cost of investment approaches zero, and hence the capital stock 1 year after the tax hike can respond with large declines.

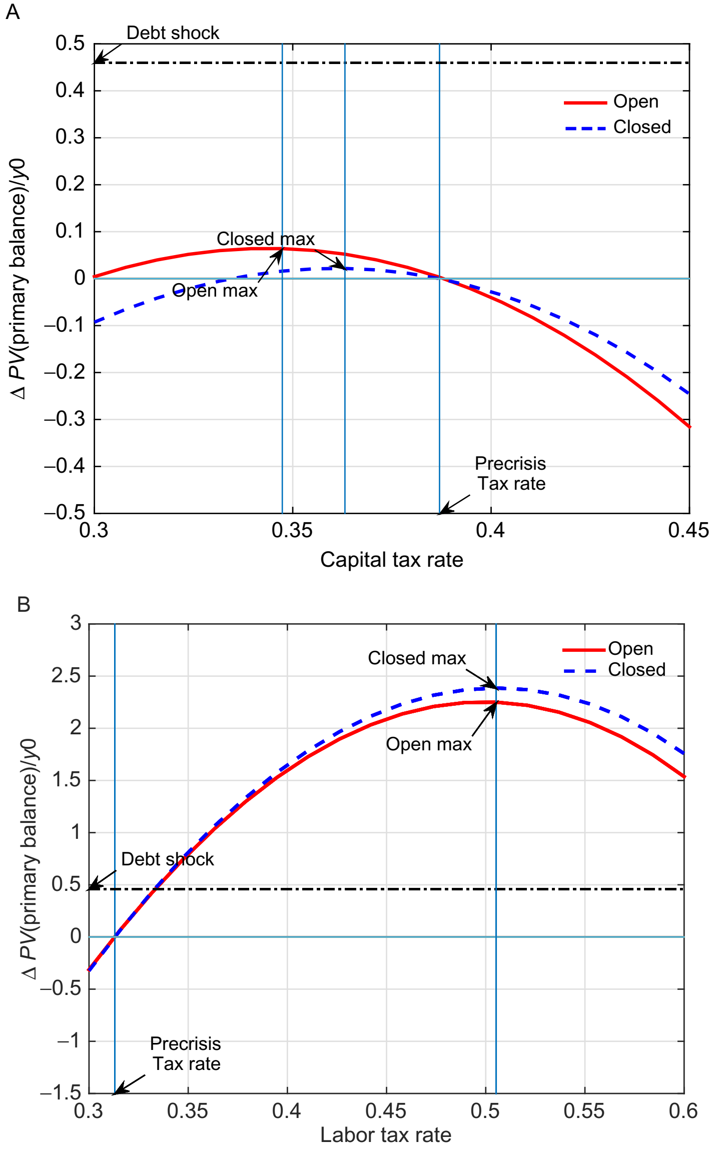


Fig. 13 Dynamic Laffer curves for Japan. (A) Capital tax. (B) Labor tax.

In general, the DLC results for Japan are a more extreme version of those for the United States: The capital tax cannot restore fiscal solvency because Japan's DLC for this tax peaks well below the required increase, while there is a lot of room for labor (or consumption) taxes to do it. One important difference is that the precrisis high capital tax rate in Japan is inefficient (ie, in the decreasing segment of the DLC). Because of this, the tax externalities work in the opposite direction to those observed for the US DLC, and so cutting the capital tax in Japan relative to the precrisis rate as a closed economy yields a smaller increase in the present value of the primary balance than as an open economy. One important caveat to the above results is that Japan has been stuck with slow growth and deflation for about two decades. Although raising consumption and labor taxes helps balance government budgets, higher taxes still cause efficiency and welfare losses. Japan did increase its consumption tax from 5% to 8% in April 2014, but after that the economy tipped back into recession and a further hike of the consumption tax to 10% was postponed. Moreover, if we reduce the long-run growth rate in the model to the 0.8% per-capita GDP growth rate observed on average in Japan between 2001 and 2014, the two DLCs shift downward sharply. The capital tax becomes effectively useless as it yields negligible amounts of extra revenue. The labor tax needed to make the debt sustainable is significantly higher, and thus the associated efficiency and welfare losses are larger as well.

Another caveat is that our analysis abstracts from Japan's aging demographics, rising pressures on government finance from public pensions and medical expenses, etc. These considerations place heavy burdens on the sustainability of public debt. Imrohoroglu and Sudo (2011) and Hansen and Imrohoroglu (2013) use a Neoclassical growth model to quantify the implications of the projected low population growth rate and permanent increase in total government outlays on fiscal sustainability. Imrohoroglu and Sudo find that even an increase in the consumption tax to 15% and an annual GDP growth of 3% over the next 20 years is not sufficient to restore fiscal balance unless expenditures are also contained. Hansen and Imrohoroglu find that fiscal sustainability requires the consumption tax rate be set to unprecedentedly high levels of 40–60%. Moreover, Imrohoroglu et al. (2016) and Braun and Joines (2015) use overlapping generation models and also find that current fiscal policies are not sustainable and large fiscal adjustments are needed.[aj]

Another important consideration in assessing the results of the structural analysis is that we abstracted from the possibility of taxing wealth, in particular taxing the initial capital stock. The optimal taxation literature has made the well-known argument that from an efficiency standpoint taxing the initial, predetermined capital stock is optimal. However, the argument hinges on the assumption of government commitment, which sets aside key issues of time consistency and the implications of lack of commitment.

In our model, a wealth tax would be equivalent to confiscation of a fraction of k_0 unexpectedly. Since utilization is endogenous, this tax would also affect utilization as of date 0: The marginal product of utilization declines with lower capital, utilization falls, and thus capital income and capital income tax revenue fall. But more importantly, three arguments raise serious questions about the possibility of taxing wealth in this way.

[aj] In the next section we discuss the implications of unfunded pension and entitlement liabilities for debt sustainability when the government is not committed to repay and responds to distributional incentives to default.

First, the government would have to sell confiscated capital to raise revenue (in the realistic scenario in which confiscated capital and government outlays involve different goods and services), which would lower the price at which capital goods can be sold. Second, the expectation of future confiscation of capital would not be zero, and to the extent that is positive it would act as a tax on future capital accumulation and capital income. Third, as an implication of the first two arguments, the wealth tax actually looks more like a government default that would seem to necessitate modeling government behavior without commitment (in fact, in a setup without utilization and capital as the only productive factor, the government confiscating some of k_0 is equivalent to defaulting on a fraction of the date-0 debt repayment).

Perhaps because of the above arguments, the history of wealth taxes has not been a happy one. Wealth taxes were discarded by Austria, Denmark, and Germany in 1997, by Finland, Iceland, and Luxembourg in 2006 and by Sweden in 2007. Interestingly, these countries claimed to ditch the wealth tax in efforts to get more revenue, not less. Moreover, implementing wealth taxation faces serious hurdles, particularly for the valuation of assets and for preventing tax evasion. Global financial integration also makes taxing wealth more difficult, because the expectation of potential future confiscation via wealth taxes mentioned above discourages investment and encourages capital flight (see the discussion in Eichengreen, 1989 and the recent experience with "tax inversions" in the United States).

To summarize where the chapter is at this point, we first explored the question of public debt sustainability from the viewpoint of an empirical approach based on the estimation and analysis of fiscal reaction functions. We found that the sufficiency condition for public debt to be sustainable (ie, for IGBC to hold), reflected in a positive conditional response of the primary balance to public debt, cannot be rejected by the data. At the same time, however, there is clear evidence that the fiscal dynamics observed in the aftermath of the recent surge in debt in advanced economies represent a significant structural break in the reaction functions. In plain terms, primary deficits have been too large, and are projected to remain too large, to be in line with the path projected by the reaction functions, and also relative to the fiscal adjustment process observed in previous episodes of large surges in debt.

The main limitation of the empirical approach is that it cannot say much about the macroeconomic effects of multiple fiscal adjustment paths that can restore debt sustainability. To address this issue, this section explored a structural approach that takes a variation of the workhorse two-country Neoclassical dynamic equilibrium model with an explicit fiscal sector. Capacity utilization and a limited tax allowance for depreciation expenses were used to match the observed elasticity of the capital tax base to capital tax changes. Then we calibrated this model to the United States and European data and used it to quantify the effects of unilateral changes in capital and labor taxes aimed at altering the ability of countries to sustain debt. The results suggest striking differences across Europe and the United States. For the United States, the results suggest that changes in capital taxes cannot make the observed increase in debt sustainable, while small increases in labor taxes could. For Europe, the model predicts that the ability of the tax system to make higher debt ratios sustainable is nearly fully exhausted. Capital taxation is highly inefficient and in the decreasing segment of DLCs, so cuts in capital taxes would be needed to restore fiscal solvency. Labor taxes are near the peak of the DLC, and even if increased to the maximum point they fail to increase the present value of the primary balance to make the observed surge in debt sustainable. Moreover, international externalities of capital income taxes are quantitatively large, suggesting that incentives for strategic interaction, and the classic race-to-the-bottom in capital income taxation are nontrivial.

In short, the results from the empirical and the structural approaches to evaluate debt sustainability cast doubt on the presumption that the high debt ratios reached by many advanced economies in the years since 2008 will be fully repaid. To examine debt sustainability allowing for the possibility of nonrepayment, however, we must consider a third approach that relaxes the assumption that the government is committed to repay domestic debt, which is central to the two approaches we have covered. In the next section of this chapter we turn our attention to this issue.

4. DOMESTIC DEFAULT APPROACH

We now examine debt sustainability from the perspective of a framework that abandons the assumption of a government committed to repay domestic debt. The emphasis is on the risk of de-jure, or outright, default on domestic public debt, not the far more studied issues of external sovereign default, which is the subject of another chapter in this Handbook, or de-facto default on domestic debt via inflation. Interest on domestic sovereign default is motivated by the seminal empirical study of Reinhart and Rogoff (2011), which documents episodes of outright default on domestic public debt in a cross-country historical dataset going back to 1750.[ak] Hall and Sargent (2014) describe in detail a similar episode in the process by which the US government handled the management of its debt in the aftermath of the Revolutionary War.

Reinhart and Rogoff noted that the literature has paid little attention to domestic sovereign default, and thus chose to title their paper *The Forgotten History of Domestic Debt*. As we document below, the situation has changed somewhat recently, but relatively speaking the study of domestic government defaults remains largely uncharted territory.

[ak] Reinhart and Rogoff identified 68 outright domestic default episodes, which occurred via mechanisms such as forcible conversions, lower coupon rates, unilateral reductions of principal, and suspensions of payments.

The ongoing European debt crisis also highlights the importance of studying domestic sovereign default, because four features of the crisis (thinking of Europe as a whole) make it resemble more a domestic default than an external default. First, countries in the Eurozone are highly integrated, with the majority of their public debt denominated in their common currency and held by European residents. Hence, a default means, to a large extent, a suspension of payments to "domestic" (ie, European) agents instead of external creditors. Second, domestic public-debt-GDP ratios are high in the Eurozone in general, and very large in the countries at the epicenter of the crisis (Greece, Ireland, Italy, Spain, and Portugal). Third, the Eurozones common currency and common central bank rule out the possibility of individual governments resorting to inflation as a means to lighten their debt burden without an outright default. Fourth, and perhaps most important from the standpoint of the theory proposed in this section, European-wide institutions such as the European Central Bank (ECB) and the European Commission are weighting the interests of both creditors and debtors in assessing the pros and cons of sovereign defaults by individual countries, and creditors and debtors are aware of these institutions concern and of their key role in influencing expectations and default risk.

Table 13 shows that the Eurozone's fiscal crisis has been characterized by rapid increases in public debt ratios and sovereign spreads that coincided with rising government expenditure ratios. The table also shows that debt ownership, as proxied by Gini coefficients of wealth distributions, is unevenly distributed in the seven countries listed, with mean and median Gini coefficients of around two-thirds. The degree of concentration in the ownership of public debt plays a key role in the framework of optimal

Table 13 Euro area: Key fiscal statistics and wealth inequality

	Gov. debt		Gov. exp.		Spreads		
Moment (%)	**Avg.**	**2011**	**Avg.**	**"Crisis peak"**	**Avg.**	**"Crisis peak"**	**Gini Wealth**
France	34.87	62.72	23.40	24.90	0.08	1.04	0.73
Germany	33.34	52.16	18.80	20.00	–	–	0.67
Greece	84.25	133.09	18.40	23.60	0.37	21.00	0.65
Ireland	14.07	64.97	16.10	20.50	0.11	6.99	0.58
Italy	95.46	100.22	19.40	21.40	0.27	3.99	0.61
Portugal	35.21	75.83	20.00	22.10	0.20	9.05	0.67
Spain	39.97	45.60	17.60	21.40	0.13	4.35	0.57
Avg.	48.17	76.37	19.10	21.99	0.22	7.74	0.64
Median	35.21	64.97	18.80	21.40	0.17	5.67	0.65

Note: Author's calculations are based on OECD Statistics, Eurostat, ECSB, and Davies et al. (2009). "Gov. debt" refers to total general government net financial liabilities (avg 1990–2007); "Gov. Exp." corresponds to government purchases in national accounts (avg 2000–07); "Sov spreads" correspond to the difference between interest rates of the given country and Germany for bonds of similar maturity (avg 2000–07). For a given country i, they are computed as $(1 + r^i)/(1 + r^{Ger}) - 1$. "Crisis Peak" refers to the maximum value observed during 2008–12 using data from Eurostat. "Gini wealth" are Gini wealth coefficients for 2000 from Davies, J., Sandstr´om, S., Shorrocks, A., Wolff, E. 2009. The level and distribution of global household wealth. NBER Working Paper 15508, appendix V.

domestic default examined in this section. The framework also predicts that spreads and the probability of default at higher when government outlays are higher.

The model on which this section is based follows the work of D'Erasmo and Mendoza (2013) and D'Erasmo and Mendoza (2014). The goal is to analyze the optimal default and borrowing decisions of a government unable to commit to repay debt placed with domestic creditors in an environment with incomplete markets. The key difference with standard external default models is in that the payoff of the government includes the utility of agents who are government bondholders, as well as nonbondholders. As a result, the main incentive to default is to redistribute resources across these two groups of agents.[al] Default is assumed to be nondiscriminatory (ie, the government cannot discriminate across any of its creditors when it defaults). There is explicit aggregate risk in the form of shocks to government outlays, and also implicit in the form of default risk.

Government bondholders and nonbondholders are modeled with identical CRRA preferences. Default is useful as a vehicle for redistribution across the two, but it also has costs. We explore the case in which there is an exogenous cost in terms of disposable income, similar to the exogenous income costs typical of the external default literature. But there can also be endogenous costs related to the reduced ability to smooth taxation and provide liquidity, and, in long-horizon environments, to the loss of access to government bonds as the asset used for self-insurance.

In this framework, public debt is sustainable when it is supported as part of the equilibrium without commitment. This implies that a particular price and stock of defaultable government bonds are sustainable only if they are consistent with the optimal debt-issuance and default plans of the government, the optimal savings plans of private agents, and the bond market-clearing condition. Sustainable debt thus factors in the risk of default, which implies paying positive risk premia on current debt issuance when future default is possible. Debt becomes unsustainable when default becomes the optimal choice ex post, or is unsustainable ex ante for debt levels that cannot be issued at a positive price (ie, when a given debt issued at t entails a 100% probability of default at t+1).

This model is not necessarily limited to a situation in which private agents hold directly government debt. It is also applicable to situations in which pension funds hold government bonds and retirement accounts are structured as individual accounts, or where the financial sector holds domestic sovereign debt and households hold claims

[al] The model should not be viewed as focusing necessarily on redistribution across the poor and rich, but across agents that hold public debt and those who do not. The two are correlated but need not be the same. For instance, Hall and Sargent (2014) describe how the domestic default after the US Revolutionary War implied redistribution from bondholders in the South to nonbondholders in the North, with both groups generally wealthy. Similarly, in the European debt crisis, a Greek default can be viewed as redistributing from German tax payers to Greek households and not according to their overall wealth.

on the financial sector. Moreover, the general principle that domestic default is driven by government's distributional incentives traded off against exogenous or endogenous default costs applies to more complex environments that include implicit (or contingent) government liabilities due, for example, to expected funding shortfalls in entitlement programs. Default in these cases can take the form of reforms like increasing retirement eligibility ages or imposing income ceilings in eligibility for programs like medicare. For simplicity, however, the quantitative analysis conducted later in this section is calibrated to data that includes only explicit government debt (total general government net financial liabilities as defined in Eurostat).

We develop the argument using the two-period model proposed by D'Erasmo and Mendoza (2013), which highlights the importance of the distributional incentives of default at the expense of setting aside endogenous default costs due to the loss of access to self-insurance assets. D'Erasmo and Mendoza (2014) and Dovis et al. (2014) study the role of distributional incentives to default on domestic debt, and the use of public debt in infinite horizon models with domestic agent heterogeneity. The two differ in that Dovis et al. (2014) assume complete domestic asset markets, which removes the role of public debt as providing social insurance for domestic agents. In addition, they focus on the solution to the Ramsey problem, in which default is not observed along the equilibrium path. D'Erasmo and Mendoza study an economy with incomplete markets, which turns the loss of the vehicle for self-insurance, and the severity of the associated liquidity constraints, into an endogenous cost of default that plays a central role in their results. They also solve for Markov-perfect equilibria in which default is possible as an equilibrium outcome.

The model discussed here is also related to the literature that analyzes the role of public debt as a self-insurance mechanism and a tool for altering consumption dispersion in heterogeneous-agents models without default (eg, Aiyagari and McGrattan (1998), Golosov and Sargent (2012), Azzimonti et al. (2014), Floden (2001) , Heathcote (2005), and Aiyagari et al. (2002)). A recent article by Pouzo and Presno (2014) introduces the possibility of default into models in this class. They study optimal taxation and public debt dynamics in a representative-agent setup similar to Aiyagari et al. (2002) but allowing for default and renegotiation.

The recent interest in domestic sovereign default also includes a strand of literature focusing on the consequences of default on domestic agents, its relation with secondary markets, discriminatory vs nondiscriminatory default, and the role of domestic debt in providing liquidity to the corporate sector (see Guembel and Sussman, 2009, Broner et al., 2010, Broner and Ventura, 2011, Gennaioli et al., 2014, Basu, 2009, Brutti, 2011, Mengus, 2014, and Di Casola and Sichlimiris, 2014). There are also some recent studies motivated by the 2008 financial crisis that focus on the interaction between sovereign debt and domestic financial institutions such as Sosa-Padilla (2012), Bocola (2014), Boz et al. (2014), and Perez (2015).

4.1 Model Structure

Consider a two-period economy $t=0, 1$ inhabited by a continuum of agents with aggregate unit measure. All agents have the same preferences, which are given by:

$$u(c_0)+\beta E[u(c_1)],\ u(c)=\frac{c^{1-\sigma}}{1-\sigma}$$

where $\beta \in (0, 1)$ is the discount factor and c_t for $t=0, 1$ is individual consumption. The utility function $u(\cdot)$ takes the standard CRRA form.

All agents receive a nonstochastic endowment y each period and pay lump-sum taxes τ_t, which are uniform across agents. Taxes and newly issued government debt are used to pay for government consumption g_t and repayment of outstanding government debt. The (exogenous) initial supply of outstanding government bonds at $t = 0$ is denoted B_0. Agents differ in their initial wealth position, which is characterized by their holdings of government debt at the beginning of the first period.[am] Given B_0, the initial wealth distribution is defined by a fraction γ of households who are the L-type individuals with initial bond holdings b_0^L, and a fraction $(1-\gamma)$ who are the H-types and hold b_0^H, where $b_0^H=\frac{B_0-\gamma b_0^L}{1-\gamma}\geq b_0^L\geq 0$. This value of b_0^H is the amount consistent with market-clearing in the government bond market at $t=0$, since we are assuming that the debt is entirely held by domestic agents. The initial distribution of wealth is exogenous, but the distribution at the beginning of the second period is endogenously determined by the agents' savings choices of the first period.

The budget constraints of the two types of households in the first period are given by:

$$c_0^i+q_0 b_1^i=y+b_0^i-\tau_0 \quad \text{for } i=L,H. \tag{10}$$

Agents collect the payout on their initial holdings of government debt (b_0^i), receive endowment income y, and pay lump-sum taxes τ_0. These net-of-tax resources are used to pay for consumption and purchases of new government bonds b_1^i. Agents are not allowed to take short positions in government bonds, which is equivalent to assuming that bond purchases must satisfy the familiar no-borrowing condition often used in heterogeneous-agents models: $b_1^i\geq 0$.

The budget constraints in the second period differ depending on whether the government defaults or not. If the government repays, the budget constraints take the standard form:

$$c_1^i=y+b_1^i-\tau_1 \quad \text{for } i=L,H. \tag{11}$$

[am] Andreasen et al. (2011), Ferriere (2014), and Jeon and Kabukcuoglu (2014) study environments in which domestic income heterogeneity plays a central role in the determination of external defaults.

If the government defaults, there is no repayment on the outstanding debt, and the agents' budget constraints are:

$$c_1^i = (1-\phi(g_1))y - \tau_1 \quad \text{for } i = L, H. \tag{12}$$

As is standard in the external sovereign default literature, we allow for default to impose an exogenous cost that reduces income by a fraction ϕ. This cost is often modeled as a function of the realization of a stochastic endowment income, but since income is constant in this setup, we model it as a function of the realization of government expenditures in the second period g_1. In particular, the cost is a nonincreasing, step-wise function: $\phi(g_1) \geq 0$, with $\phi'(g_1) \leq 0$ for $g_1 \leq \bar{g}_1$, $\phi'(g_1) = 0$ otherwise, and $\phi''(g_1) = 0$. Hence, $\bar{g}_1$ is a threshold high value of g_1 above which the marginal cost of default is zero. This formulation is analogous to the step-wise default cost as a function of income proposed by Arellano (2008) and now widely used in the external default literature, and it also captures the idea of asymmetric costs of tax collection (see Barro, 1979 and Calvo, 1988). Note, however, that for the model to support equilibria with debt under a utilitarian government all we need is $\phi(g_1) > 0$. The additional structure is useful for the quantitative analysis and for making it easier to compare the model with the standard external default models.[an]

At the beginning of $t = 0$, the government has outstanding debt B_0 and can issue one-period, nonstate contingent discount bonds $B_1 \in \mathcal{B} \equiv [0, \infty)$ at the price $q_0 \geq 0$. Each period it collects lump-sum revenues τ_t and pays for outlays g_t. Since g_0 is known at the beginning of the first period, the relevant uncertainty with respect to government expenditures is for g_1, which follows a log-normal distribution $N((1-\rho_g)\mu_g + \rho_g \ln(g_0), \frac{\sigma_g^2}{(1-\rho_g^2)})$.[ao] We do not restrict the sign of τ_t, so $\tau_t < 0$ represents lump-sum transfers.[ap]

[an] In external default models, the nonlinear cost makes default more costly in "good" states, which alters default incentives to make default more frequent in "bad" states, and it also contributes to support higher debt levels.

[ao] This is similar to an AR(1) process and allows us to control the correlation between g_0 and g_1 via ρ_g, the mean of the shock via μ_g and the variance of the unpredicted portion via σ_g^2. Note that if $\ln(g_0) = \mu_g$, $g_1 \sim N(\mu_g, \frac{\sigma_g^2}{(1-\rho_g^2)})$.

[ap] Some studies in the sovereign debt literature have examined models that include tax and expenditure policies, as well as settings with foreign and domestic lenders, but always maintaining the representative agent assumption (eg, Cuadra et al., 2010; Vasishtha, 2010). More recently Dias et al. (2012) examined the benefits of debt relief from the perspective of a *global* social planner with utilitarian preferences. Also in this literature, Aguiar and Amador (2013) analyze the interaction between public debt, taxes and default risk and Lorenzoni and Werning (2013) study the dynamics of debt and interest rates in a model where default is driven by insolvency and debt issuance driven by a fiscal reaction function.

At equilibrium, the price of debt issued in the first period must be such that the government bond market clears:

$$B_t = \gamma b_t^L + (1-\gamma) b_t^H \quad \text{for } t = 0, 1. \tag{13}$$

This condition is satisfied by construction in period 0. In period 1, however, the price moves endogenously to clear the market.

The government has the option to default at $t = 1$. The default decision is denoted by $d_1 \in \{0, 1\}$ where $d_1 = 0$ implies repayment. The government evaluates the values of repayment and default using welfare weight ω for $L-$type agents and $1-\omega$ for $H-$type agents. This specification encompasses cases in which, for political reasons for example, the welfare weights are biased toward a particular type so $\omega \neq \gamma$ or the case in which the government acts as a utilitarian social planner in which $\omega = \gamma$.[aq] At the moment of default, the government evaluates welfare using the following function:

$$\omega u(c_1^L) + (1-\omega) u(c_1^H).$$

At $t = 0$, the government budget constraint is

$$\tau_0 = g_0 + B_0 - q_0 B_1. \tag{14}$$

The level of taxes in period 1 is determined after the default decision. If the government repays, taxes are set to satisfy the following government budget constraint:

$$\tau_1^{d_1=0} = g_1 + B_1. \tag{15}$$

Notice that, since this is a two-period model, equilibrium requires that there are no outstanding assets at the end of period 1 (ie, $b_2^i = B_2 = 0$ and $q_1 = 0$). If the government defaults, taxes are simply set to pay for government purchases:

$$\tau_1^{d_1=1} = g_1. \tag{16}$$

The analysis of the model's equilibrium proceeds in three stages. First, characterize the households' optimal savings problem and determine their payoff (or value) functions, taking as given the government debt, taxes and default decision. Second, study how optimal government taxes and the default decision are determined. Third, examine the optimal choice of debt issuance that internalizes the outcomes of the first two stages. We characterize these problems as functions of B_1, g_1, γ and ω, keeping the initial conditions (g_0, B_0, b_0^L) as exogenous parameters. Hence, for given γ and ω, we can index the value

[aq] This relates to the literature on political economy and sovereign default, which largely focuses on external default (eg, Amador, 2003, Dixit and Londregan, 2000, D'Erasmo, 2011, Guembel and Sussman, 2009, Hatchondo et al., 2009, and Tabellini, 1991), but includes studies like those of Alesina and Tabellini (1990) and Aghion and Bolton (1990) that focus on political economy aspects of government debt in a closed economy, and the work of Aguiar et al. (2013) on optimal policy in a monetary union subject to self-fulfilling debt crises.

of a household as of $t = 0$, before g_1 is realized, as a function of $\{B_1\}$. Given this, the level of taxes τ_0 is determined by the government budget constraint once the equilibrium bond price q_0 is set. Bond prices are forward looking and depend on the default decision of the government in period 1, which will be given by the decision rule $d(B_1, g_1, \gamma, \omega)$.

4.2 Optimization Problems and Equilibrium

Given B_1, γ, and ω a household with initial debt holdings b_0^i for $i = L, H$ chooses b_1^i by solving this maximization problem:

$$v^i(B_1,\gamma,\omega) = \max_{b_1^i} \left\{ u(\gamma + b_0^i - q_0 b_1^i - \tau_0) + \beta E_{g_1}\left[(1-d_1)u(\gamma + b_1^i - \tau_1^{d_1=0}) \right.\right. \\ \left.\left. + d_1 u(\gamma(1-\phi(g_1)) - \tau_1^{d_1=1})\right]\right\}, \tag{17}$$

subject to $b_1^i \geq 0$. The term $E_{g_1}[.]$ represents the expected payoff across the repayment and default states in period 1. Notice in particular that the payoff in case of default does not depend on the level of individual debt holdings (b_1^i), reflecting the fact that the government cannot discriminate across households when it defaults.

A key feature of the above problem is that agents take into account the possibility of default in choosing their optimal bond holdings. The first-order condition, evaluated at the equilibrium level of taxes, yields this Euler equation:

$$u'(c_0^i) \geq \beta(1/q_0)E_{g_1}\left[u'(\gamma - g_1 + b_1^i - B_1)(1 - d_1(B_1,g_1,\gamma))\right], = \text{ if } b_1^i > 0 \tag{18}$$

In states in which, given (B_1, γ, ω), the value of g_1 is such that the government chooses to default $(d_1(B_1, g_1, \gamma, \omega) = 1)$, the marginal benefit of an extra unit of debt is zero.[ar] Thus, conditional on B_1, a larger default set (ie, a larger set of values of g_1 such that the government defaults), implies that the expected marginal benefit of an extra unit of savings decreases. As a result, everything else equal, a higher default probability results in a lower demand for government bonds, a lower equilibrium bond price, and higher taxes. This has important redistributive implications, because when choosing the optimal debt issuance, the government will internalize how, by altering the bond supply, it affects the expected probability of default and the equilibrium bond prices. Note also that from the agents' perspective, the default choice $d_1(B_1, g_1, \gamma, \omega)$ is independent of b_1^i.

The above Euler equation is useful for highlighting some important properties of the equilibrium pricing function of bonds:

1. The premium over a world risk-free rate (defined as q_0/β, where $1/\beta$ can be viewed as a hypothetical opportunity cost of funds for an investor, analogous to the role played by the world interest rate in the standard external default model) generally differs from the default probability for two reasons: (a) agents are risk averse, and (b) in the repayment state, agents face higher taxes, whereas in the standard model investors are not

[ar] Utility in the case of default equals $u(\gamma(1 - \phi(g_1)) - g_1)$, which is independent of b_1^i.

taxed to repay the debt. For agents with positive bond holdings, the above optimality condition implies that the premium over the risk-free rate is $E_{g_1}\left[u'(y-g_1+b_1^i-B_1)(1-d_1)/u'(c_0^i)\right]$.

2. If the Euler equation for H −type agents holds with equality (ie, $b_1^H>0$) and L −type agents are *credit constrained* (ie, $b_1^L=0$), the H −type agents are the marginal investor and their Euler equation can be used to derive the equilibrium price.
3. For sufficiently high values of B_1, γ or $1-\omega$ the government chooses $d_1(B_1,g_1,\gamma,\omega)=1$ for all g_1. In these cases, the expected marginal benefit of purchasing government bonds vanishes from the agents' Euler equation, and hence the equilibrium for that B_1 does not exist, since agents would not be willing to buy debt at any finite price.[as] These values of B_1 are therefore unsustainable ex ante (ie, these debt levels cannot be sold at a positive price).

The equilibrium bond pricing functions $q_0(B_1,\ \gamma,\ \omega)$, which returns bond prices for which, as long as consumption for all agents is nonnegative and the default probability of the government is less than 1, the following market-clearing condition holds:

$$B_1=\gamma b_1^L(B_1,\ \gamma,\ \omega)+(1-\gamma)b_1^H(B_1,\ \gamma,\ \omega), \tag{19}$$

where B_1 in the left-hand-side of this expression represents the public bonds supply, and the right-hand-side is the aggregate government bond demand.

As explained earlier, we analyze the government's problem following a backward induction strategy by studying first the default decision problem in the final period $t=1$, followed by the optimal debt issuance choice at $t=0$.

4.2.1 Government Default Decision at t = 1

At $t=1$, the government chooses to default or not by solving this optimization problem:

$$\max_{d\in\{0,1\}}\left\{W_1^{d=0}(B_1,\ g_1,\ \gamma,\ \omega),\ W_1^{d=1}(g_1,\ \gamma,\ \omega)\right\}, \tag{20}$$

where $W_1^{d=0}(B_1,\ g_1,\ \gamma,\ \omega)$ and $W_1^{d=1}(B_1,\ g_1,\ \gamma,\ \omega)$ denote the values of the social welfare function at the beginning of period 1 in the case of repayment and default, respectively. Using the government budget constraint to substitute for $\tau_1^{d=0}$ and $\tau_1^{d=1}$, the government's payoffs can be expressed as:

$$W_1^{d=0}(B_1,\ g_1,\gamma,\omega)=\omega u(y-g_1+b_1^L-B_1)+(1-\omega)u(y-g_1+b_1^H-B_1) \tag{21}$$

and

$$W_1^{d=1}(g_1,\gamma,\omega)=u(y(1-\phi(g_1))-g_1). \tag{22}$$

[as] This result is similar to the result in standard models of external default showing that rationing emerges at t for debt levels so high that the government would choose default at all possible income realizations in $t+1$.

Combining these payoff functions, if follows that the government defaults if this condition holds:

$$\begin{aligned} &\omega\left[u(y-g_1+\overbrace{(b_1^L-B_1)}^{\leq 0})-u(y(1-\phi(g_1))-g_1)\right] \\ &+(1-\omega)\left[u(y-g_1+\overbrace{(b_1^H-B_1)}^{\geq 0})-u(y(1-\phi(g_1))-g_1)\right]\leq 0 \end{aligned} \tag{23}$$

Notice that all agents forego g_1 of their income to government absorption regardless of the default choice. Moreover, debt repayment reduces consumption and welfare of L types and rises them for H types, whereas default implies the same consumption and utility for both types of agents.

The distributional effects of a default are implicit in condition (23). Given that debt repayment affects the cash-in-hand for consumption of L and H types according to $(b_1^L-B_1)\leq 0$ and $(b_1^H-B_1)\geq 0$, respectively, it follows that, for a given B_1, the payoff under repayment allocates (weakly) lower welfare to L agents and higher to H agents, and that the gap between the two is larger the larger is B_1. Moreover, since the default payoffs are the same for both types of agents, this is also true of the *difference* in welfare under repayment vs default: It is higher for H agents than for L agents and it gets larger as B_1 rises. To induce default, however, it is necessary not only that L agents have a smaller difference in the payoffs of repayment vs default, but that the difference is negative (ie, they must attain lower welfare under repayment than under default), which requires $B_1 > b_1^L + y\phi(g_1)$. This also implies that taxes under repayment need to be necessarily larger than under default, since $\tau_1^{d=0}-\tau_1^{d=1}=B_1$.

We can illustrate the distributional mechanism driving the default decision by comparing the utility levels associated with the consumption allocations of the default and repayment states with those that would be socially efficient. To this end, it is helpful to express the values of hypothetical optimal private debt holdings in period 1 as $b_1^L=B_1-\epsilon$ and $b_1^H(\gamma)=B_1+\dfrac{\gamma}{1-\gamma}\epsilon$, for some $\epsilon\in[0,\ B_1]$. That is, ϵ represents a given hypothetical decentralized allocation of debt holdings across agents.[at] Consumption allocations under repayment would therefore be $c_1^L(\epsilon)=y-g_1-\epsilon$ and $c_1^H(\gamma,\epsilon)=y-g_1+\dfrac{\gamma}{1-\gamma}\epsilon$, so ϵ also determines the decentralized consumption dispersion.

[at] We take ϵ as given at this point because it helps us explain the intuition behind the distributional default incentives of the government, but ϵ is an equilibrium outcome solved for later on. Also, ϵ must be nonnegative, otherwise H types would be the nonbondholders.

The government payoff under repayment can be rewritten as:

$$W^{d=0}(\epsilon, g_1, \gamma, \omega) = \omega u(y - g_1 + \epsilon) + (1 - \omega)u\left(y - g_1 + \frac{\gamma}{1-\gamma}\epsilon\right).$$

The efficient dispersion of consumption that the social planner would choose is characterized by the value of ϵ^{SP} that maximizes social welfare under repayment. In the particular case of $\omega = \gamma$ (ie, when the government is utilitarian and uses welfare weights that match the wealth distribution), ϵ^{SP} satisfies this first-order condition:

$$u'\left(y - g_1 + \frac{\gamma}{1-\gamma}\epsilon^{SP}\right) = u'\left(y - g_1 - \epsilon^{SP}\right). \tag{24}$$

Hence, the efficient allocations are characterized by zero consumption dispersion, because equal marginal utilities imply $c^{L,SP} = c^{H,SP} = y - g_1$, which is attained with $\epsilon^{SP} = 0$.

Continuing under the utilitarian government assumption ($\omega = \gamma$), consider now the government's default decision when default is costless ($\phi(g_1) = 0$). Given that the only policy instruments the government can use, other than the default decision, are nonstate contingent debt and lump-sum taxes, it is straightforward to show that default will always be optimal. This is because default supports the socially efficient allocations in the decentralized equilibrium (ie, it yields zero consumption dispersion with consumption levels $c^L = c^H = y - g_1$). This outcome is invariant to the values of B_1, g_1, γ and ϵ (over their relevant ranges). This result also implies, however, that in this model a utilitarian government without default costs can never sustain debt.

The above scenario is depicted in Fig. 14, which plots the social welfare function under repayment as a function of ϵ as the bell-shaped curve, and the social welfare under default (which is independent of ϵ), as the black dashed line. Clearly, the maximum welfare under repayment is attained when $\epsilon = 0$ which is also the efficient amount of consumption dispersion ϵ^{SP}. Moreover, since the relevant range of consumption dispersion is $\epsilon > 0$, welfare under repayment is decreasing in ϵ over the relevant range.

These results can be summarized as follows:

Result 1. *If $\phi(g_1) = 0$ for all g_1 and $\omega = \gamma$, then for any $\gamma \in (0, 1)$ and any (B_1, g_1), the social value of repayment $W^{d=0}(B_1, g_1, \gamma)$ is decreasing in ϵ and attains its maximum at the socially efficient point $\epsilon^{SP} = 0$ (ie, when welfare equals $u(y - g_1)$). Hence, default is always optimal for any given decentralized consumption dispersion $\epsilon > 0$.*

The outcome is very different when default is costly. With $\phi(g_1) > 0$, default still yields zero consumption dispersion, but at lower levels of consumption and therefore utility,

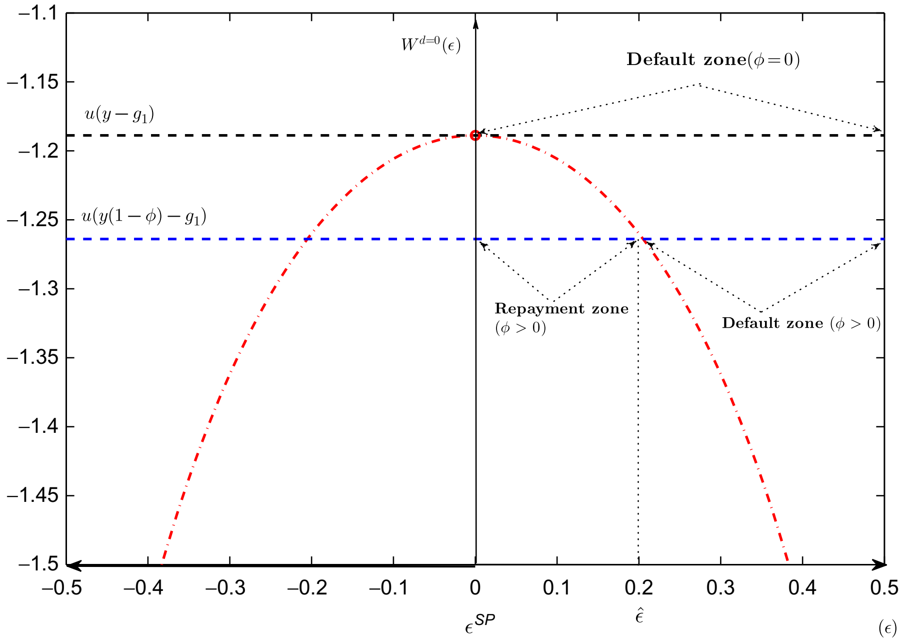


Fig. 14 Default decision and consumption dispersion.

since consumption allocations under default are $c^L = c^H = (1 - \phi(g_1))y - g_1$. This does not alter the result that the social optimum is $\epsilon^{SP} = 0$, but what changes is that default can no longer support the socially efficient consumption allocations. Instead, there is now a threshold amount of consumption dispersion in the decentralized equilibrium, $\hat{\epsilon}(\gamma)$, which varies with γ and such that for $\epsilon \geq \hat{\epsilon}(\gamma)$ default is again optimal, but for lower ϵ repayment is now optimal. This is because when ϵ is below the threshold, repayment produces a level of social welfare higher than under default.

Fig. 14 also illustrates this scenario. The default cost lowers the common level of utility of both types of agents, and hence of social welfare, in the default state (shown in the figure as the blue dashed line), and $\hat{\epsilon}(\gamma)$ is determined where social welfare under repayment and under default intersect. If the decentralized consumption dispersion with the debt market functioning (ϵ) is between 0 and less than $\hat{\epsilon}(\gamma)$ then it is optimal for the government to repay. Intuitively, if consumption dispersion is not too large, the government prefers to repay because the income cost imposed on agents to remove consumption dispersion under default is too large. Moreover, as γ rises the domain of $W_1^{d=0}$ narrows, and thus $\hat{\epsilon}(\gamma)$ falls and the interval of decentralized consumption dispersions that supports repayment narrows. This is natural because a higher γ causes the planner to weight more L-types in the social welfare function, which are agents with weakly lower utility in the repayment state.

These results can be summarized as follows:

Result 2. *If $\phi(g_1) > 0$, then for any $\gamma \in (0, 1)$ and any (B_1, g_1), there is a threshold value of consumption dispersion $\hat{\epsilon}(\gamma)$ such that the payoffs of repayment and default are equal: $W^{d=0}(B_1, g_1, \gamma) = u(y(1-\phi(g_1)) - g_1)$. The government repays if $\epsilon < \hat{\epsilon}(\gamma)$ and defaults otherwise. Moreover, $\hat{\epsilon}(\gamma)$ is decreasing in γ.*

Introducing a bias in the welfare function of the government (relative to utilitarian social welfare) can result in repayment being optimal even without default costs, which provides for an alternative way to sustain debt subject to default risk. Assuming $\phi(g_1) = 0$, there are two possible scenarios depending on the relative size of γ and ω. First, if $\omega > \gamma$, the planner again always chooses default as in the setup with $\omega = \gamma$. This is because for any $\epsilon > 0$, the decentralized consumption allocations feature $c^H > c^L$, while the planner's optimal consumption dispersion requires $c^H \leq c^L$, and hence ϵ^{SP} cannot be implemented. Default brings the planner the closest it can get to the payoff associated with ϵ^{SP} and hence it is always chosen.

In the second scenario $\omega < \gamma$, which means that the government's bias assigns more (less) weight to H (L) types than the fraction of each type of agents that actually exists. In this case, the model can support equilibria with debt even without default costs. In particular, there is a threshold consumption dispersion $\hat{\epsilon}$ such that default is optimal for $\epsilon \geq \hat{\epsilon}$, where $\hat{\epsilon}$ is the value of ϵ at which $W_1^{d=0}(\epsilon, g_1, \gamma, \omega)$ and $W_1^{d=1}(g_1)$ intersect. For $\epsilon < \hat{\epsilon}$, repayment is preferable because $W_1^{d=0}(\epsilon, g_1, \gamma, \omega) > W_1^{d=0}(g_1)$. Thus, without default costs, equilibria for which repayment is optimal require two conditions: (a) that the government's bias favors bond holders ($\omega < \gamma$), *and* (b) that the debt holdings chosen by private agents do not produce consumption dispersion in excess of $\hat{\epsilon}$.

Fig. 15 illustrates the outcomes just described. This figure plots $W_1^{d=0}(\epsilon, g_1, \gamma, \omega)$ for $\omega \gtreqless \gamma$. The planner's default payoff and the values of ϵ^{SP} for $\omega \gtreqless \gamma$ are also identified in the plot. The vertical intercept of $W_1^{d=0}(\epsilon, g_1, \gamma, \omega)$ is always $W^{d=1}(g_1)$ for any values of ω and γ, because when $\epsilon = 0$ there is zero consumption dispersion and that is also the outcome under default. In addition, the bell-shaped form of $W_1^{d=0}(\epsilon, g_1, \gamma, \omega)$ follows from $u'(.) > 0, u''(.) < 0$.[au]

Take first the case with $\omega > \gamma$. In this case, the planner's payoff under repayment is the dotted bell curve. Here, $\epsilon^{SP} < 0$, because the optimality condition implies that the planner's optimal choice features $c^L > c^H$. Since default is the only instrument available to the government, however, these consumption allocations are not feasible, and by choosing

[au] Note in particular that $\frac{\partial W_1^{d=0}(\epsilon, g_1, \gamma, \omega)}{\partial \epsilon} \gtreqless 0 \iff \frac{u'(c^H(\epsilon))}{u'(c^L(\epsilon)}) \gtreqless \left(\frac{\omega}{\gamma}\right)\left(\frac{1-\gamma}{1-\omega}\right)$. Hence, the planner's payoff is increasing (decreasing) at values of ϵ that support sufficiently low (high) consumption dispersion so that $\frac{u'(c^H(\epsilon))}{u'(c^L(\epsilon))}$ is above (below) $\left(\frac{\omega}{\gamma}\right)\left(\frac{1-\gamma}{1-\omega}\right)$.

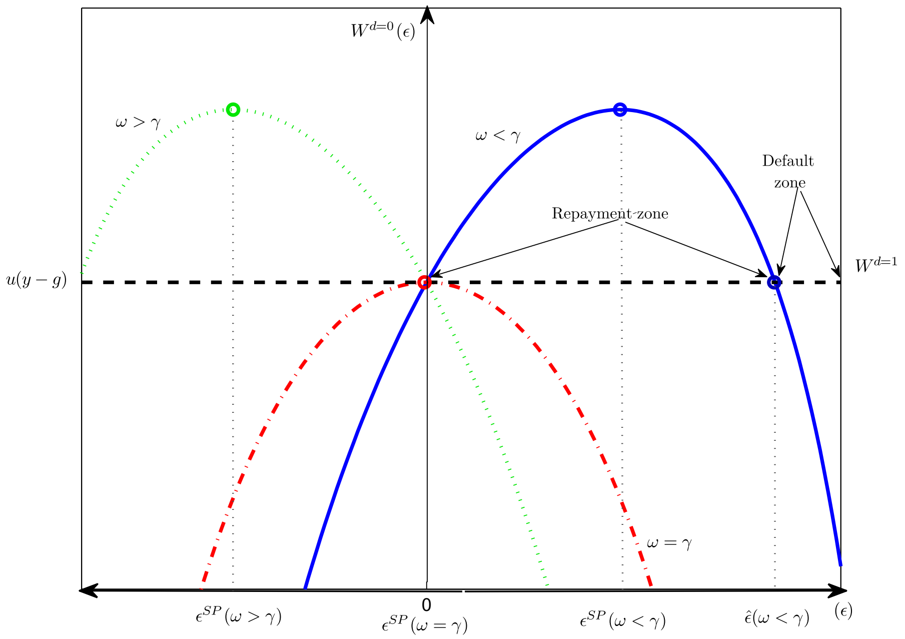


Fig. 15 Default decision with nonutilitarian planner ($\phi = 0$).

default the government attains $W^{d=1}$, which is the highest feasible government payoff for any $\epsilon \geq 0$. In contrast, in the case with $\omega = \gamma$, for which the planner's payoff function is the dashed bell curve, the planner chooses $\epsilon^{SP} = 0$, and default attains exactly the same payoff, so default is chosen. In short, if $\omega \geq \gamma$, the government always defaults for any decentralized distribution of debt holdings represented by $\epsilon > 0$, and thus equilibria with debt cannot be supported.

When $\omega < \gamma$, the planner's payoff is the continuous curve. The intersection of the downward-sloping segment of $W_1^{d=0}(\epsilon, g_1, \gamma, \omega)$ with $W^{d=1}$ determines the default threshold $\hat{\epsilon}$ such that default is optimal only in the *default zone* where $\epsilon \geq \hat{\epsilon}$. Default is still a second-best policy for the planner, because with it the planner cannot attain $W^{d=0}(\epsilon^{SP})$, it just gets the closest it can get. In contrast, the choice of repayment is preferable in the *repayment zone* where $\epsilon < \hat{\epsilon}$,, because in this zone $W_1^{d=0}(\epsilon, g_1, \gamma, \omega) > W^{d=1}(g_1)$.

Adding default costs to this political bias setup ($\phi(g_1) > 0$) makes it possible to support repayment equilibria even when $\omega \geq \gamma$. As Fig. 16 shows, with default costs there are threshold values of consumption dispersion, $\hat{\epsilon}$, separating repayment from default zones for $\omega \lesseqgtr \gamma$.

It is also evident in Fig. 16 that the range of values of ϵ for which repayment is chosen widens asγ rises relative to ω. Thus, when default is costly, equilibria with repayment require only the condition that the debt holdings chosen by private agents, which are implicit in ϵ, do not produce consumption dispersion larger than the value of$\hat{\epsilon}$ associated

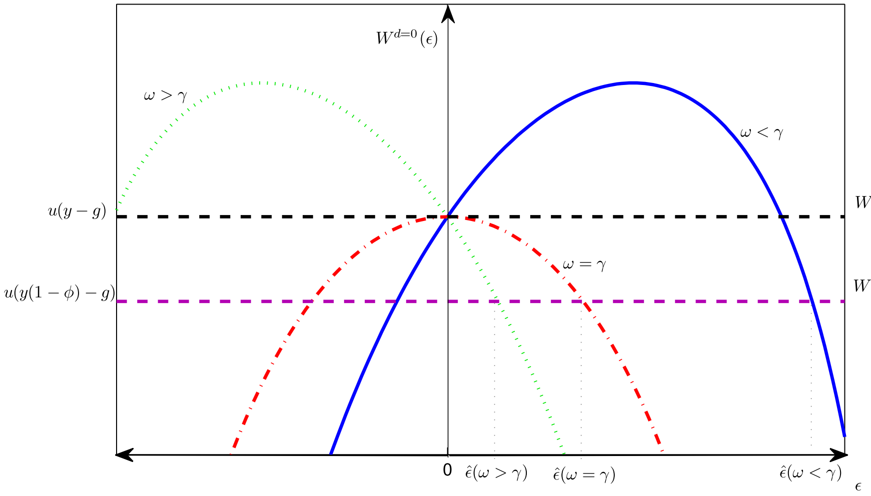


Fig. 16 Default decision with nonutilitarian planner when $\phi(g_1) > 0$.

with a given (ω, γ) pair. Intuitively, the consumption of H-type agents must not exceed that of L-type agents by more than what $\hat{\epsilon}$ allows, because otherwise default is optimal.

The fact that a government biased in favor of bond holders can find it optimally to repay may seem unsurprising. As we argue later, however, in fact governments with this bias can be an endogenous outcome of majority voting if the fraction of agents that are nonbondholders is sufficiently large. This occurs when these agents are liquidity constrained (ie, hitting the no-borrowing constraint), because in this case they prefer that the government favors bondholders so that it can sustain higher debt levels because public debt provides them with liquidity.

4.2.2 Government Debt Issuance Decision at $t = 0$

We are now in a position to study how the government chooses the optimal amount of debt to issue in the initial period. These are the model's predicted sustainable debt levels ex ante, some of which will be optimally defaulted on ex post, depending on the realization of $g_1$in the second period. Both the government and the private sector are aware of this, so the debt levels that can be issued at equilibrium in the first period are traded at prices that can carry a default risk premium, which will be the case if for a given debt stock there are some values of g_1 for which default is the optimal choice in the second period.

The government's optimization problem is easier to understand if we first illustrate how public debt serves as a tool for altering consumption dispersion across agents both within a period and across periods. In particular, consumption dispersion in each period and repayment state is given by these conditions:

$$c_0^H - c_0^L = \frac{1}{1-\gamma}[B_0 - q_0(B_1,\gamma,\omega)B_1],$$

$$c_1^{H,d=0} - c_1^{L,d=0} = \frac{1}{1-\gamma}B_1,$$

$$c_1^{H,d=1} - c_1^{L,d=1} = 0.$$

These expressions make it clear that, given B_0, issuing at least some debt ($B_1 > 0$) reduces consumption dispersion at $t=0$ compared with no debt ($B_1 = 0$), but increases it at $t=1$ if the government repays (ie, $d=0$). Moreover, the use of debt as tool for redistribution of consumption at $t=0$ is hampered by a Laffer curve relationship just like the distortionary taxes of the previous section. In this case, it takes the form of the debt Laffer curve familiar from the external default literature, which is defined by the mapping from an amount of debt issued B_1 to the resources the government acquires with that amount of borrowing, $q_0(B_1, \gamma, \omega)B_1$. This mapping behaves like a Laffer curve because higher debt issuance carries a higher default risk, which reduces the price of the debt. Near zero debt the default risk is also zero so higher debt increases resources for the government, at very high debt near the region at which debt is unsustainable ex ante, higher debt reduces resources because the price falls proportionally much more than the debt rises, and in between we obtain the bell-shaped Laffer curve relationship. It follows then from this Laffer curve that, starting from $B_1 = 0$, consumption dispersion at $t = 0$ first falls as B_1 increases, but there is a critical positive value of B_1 beyond which it becomes an increasing function of debt.

At $t=0$, the government chooses its debt policy internalizing the above consumption dispersion effects, including the debt Laffer curve affecting date-0 dispersion, and their implications for social welfare. Formally, the government chooses B_1 so as to maximize the "indirect" social welfare function:

$$W_0(\gamma,\omega) = \max_{B_1}\left\{\omega v^L(B_1,\gamma,\omega) + (1-\omega)v^H(B_1,\gamma,\omega)\right\}. \tag{25}$$

where v^L and v^H are the private agents' value functions obtained from solving the problems defined in the Bellman equation (17) taking into account the government budget constraints and the equilibrium pricing function of bonds.

Focusing on the case with utilitarian government ($\omega = \gamma$), we can gain some intuition about the solution of this maximization problem from its first-order condition (assuming that the relevant functions are differentiable):

$$u'(c_0^H) = u'(c_0^L) + \frac{\eta}{q_0(B_1,\gamma,\omega)\gamma}\left\{\beta E_{g_1}[\Delta d \Delta W_1] + \gamma\mu^L\right\}$$

where

$$\eta \equiv q_0(B_1, \gamma, \omega)/\left(q_0'(B_1, \gamma, \omega)B_1\right) < 0,$$
$$\Delta d \equiv d(B_1 + \delta, g_1, \gamma) - d(B_1, g_1, \gamma) \geq 0, \quad \text{for } \delta > 0 \text{ small},$$
$$\Delta W_1 \equiv W_1^{d=1}(g_1, \gamma) - W_1^{d=0}(B_1, g_1, \gamma) \geq 0,$$
$$\mu^L \equiv q_0(B_1, \gamma, \omega)u'(c_0^L) - \beta E_{g_1}\left[(1-d^1)u'(c_1^L)\right] > 0.$$

In these expressions, η is the price elasticity of the demand for government bonds, $\Delta d \Delta W_1$ represents the marginal distributional benefit of a default, and μ^L is the shadow value of the borrowing constraint when it binds for L-type agents.

If both types of agents could be unconstrained in their savings decisions, so that $\mu_L = 0$, and if there is no change in the risk of default (or assuming commitment to remove default risk entirely), so that $E_{g_1}[\Delta d \Delta W_1] = 0$, then the optimality condition simplifies to:

$$u'(c_0^H) = u'(c_0^L).$$

Hence, in this case the social planner would want to issue debt so as to equalize marginal utilities of consumption across agents at date 0, which requires simply setting B_1 to satisfy $q_0(B_1, \gamma, \omega)B_1 = B_0$. If it is the case that L-types are constrained (ie, $\mu_L > 0$), and still assuming no change in default risk or a government committed to repay, the optimality condition becomes:

$$u'(c_0^H) = u'(c_0^L) + \frac{\eta\mu^L}{q_0(B_1, \gamma, \omega)}.$$

Since $\eta < 0$, this result implies $c_0^L < c_0^H$, because $u'(c_0^L) > u'(c_0^H)$. Thus, even with unchanged default risk or no default risk at all, the government's debt choice sets B_1 as needed to maintain an optimal, positive level of consumption dispersion, which is the one that supports an excess in marginal utility of L-type agents relative to H-type agents equal to $\frac{\eta\mu^L}{q_0(B_1, \gamma, \omega)}$. Moreover, since optimal consumption dispersion is positive, we can also ascertain that $B_0 > q_0(B_1, \gamma, \omega)B_1$, which using the government budget constraint implies that the government runs a primary surplus at $t = 0$. The government borrows resources, but less than it would need in order to eliminate all consumption dispersion (which requires zero primary balance).

The intuition for the optimality of issuing debt can be presented in terms of tax smoothing and savings: Date-0 consumption dispersion without debt issuance would be $B_0/(1 - \gamma)$, but this is more dispersion than what the government finds optimal, because by choosing $B_1 > 0$ the government provides tax smoothing (ie, reduces date-0 taxes) for everyone, which in particular eases the L-type agents credit constraint, and provides also a desired vehicle of savings for H types. Thus, positive debt increases consumption of L types (since $c_0^L = \gamma - g_0 - B_0 + q_0(B_1, \gamma, \omega)B_1$), and reduces consumption of H types (since $c_0^H = \gamma - g_0 + \left(\frac{\gamma}{1-\gamma}\right)(B_0 - q_0(B_1, \gamma, \omega)B_1)$). But issuing debt (assuming repayment) also increases consumption dispersion a $t = 1$, since debt is then paid with higher taxes on all agents, while H agents collect also the debt repayment. Thus, the debt is being chosen optimally to trade off the social costs and benefits of reducing (increasing) date-0 consumption and increasing (reducing) date-1 consumption for agents who are bondholders (nonbondholders). In doing so, the government internalizes the debt Laffer curve and the fact that additional debt lowers the price of bonds and helps reduce μ^L, which in turn reduces the government's optimal consumption dispersion.[av]

In the presence of default risk and if default risk changes near the optimal debt choice, the term $E_{g_1}[\Delta d \Delta W_1]$ enters in the government's optimality condition with a positive sign, which means the optimal gap in the date-0 marginal utilities across agents widens even more. Hence, the government's optimal choice of consumption dispersion for $t = 0$ is greater than without default risk, and the expected dispersion for $t = 1$ is lower, because in some states of the world the government will choose to default and consumption dispersion would then drop to zero. This also suggests that the government chooses a lower value of B_1 than in the absence of default risk, since date-0 consumptions are further apart. Moreover, the debt Laffer curve now plays a central role in the government's weakened incentives to borrow, because as default risk rises the price of bonds drops to zero faster and the resources available to reduce date-0 consumption dispersion peak at lower debt levels. In short, default risk reduces the government's ability to use nonstate-contingent debt in order to reduce consumption dispersion.

In summary, the more constrained the L −types agents are (higher μ^L) or the higher the expected distributional benefit of a default (higher $E_{g_1}[\Delta d \Delta W_1]$), the larger the level of debt the government finds optimal to issue. Both of these mechanisms operate as pecuniary externalities: They matter only because the government debt choice can alter the equilibrium price of bonds which is taken as given by private agents.

For given values of γ and ω, a *Competitive Equilibrium with Optimal Debt and Default Policies* is a pair of value functions $v^i(B_1, \gamma, \omega)$ and decision rules $b^i(B_1, \gamma, \omega)$ for $i = L, H$, a government bond pricing function $q_0(B_1, \gamma, \omega)$ and a set of government policy functions $\tau_0(B_1, \gamma, \omega)$, $\tau_1^{d\in\{0,1\}}(B_1, g_1, \gamma, \omega)$, $d(B_1, g_1, \gamma, \omega)$, $B_1(\gamma, \omega)$ such that:

1. Given the pricing function and government policy functions, $v^i(B_1, \gamma, \omega)$ and $b_1^i(B_1, \gamma, \omega)$ solve the households' problem.
2. $q_0(B_1, \gamma, \omega)$ satisfies the market-clearing condition of the bond market (equation (19)).
3. The government default decision $d(B_1, g_1, \gamma, \omega)$ solves problem (20).

[av] Note, however, that without default risk the Laffer curve has less curvature than with default risk, because $q_0^{ND}(B_1, \gamma) \geqq q_0(B_1, \gamma)$.

4. Taxes $\tau_0(B_1, \gamma, \omega)$ and $\tau_1^d(B_1, g_1, \gamma, \omega)$ are consistent with the government budget constraints.
5. The government debt policy $B_1(\gamma, \omega)$ solves problem (25).

4.3 Quantitative Analysis

We study the quantitative predictions of the model using a calibration based on European data. Since the model is simple, the goal is not to match closely the observed dynamics of debt and risk premia in Europe, but to show that a reasonable set of parameter values can support an equilibrium in which sustainable debt subject to default risk exists.[aw] We also use this numerical analysis to study show the dispersion of initial wealth and the bias in government welfare affect sustainable debt.

4.3.1 Calibration

The model is calibrated to annual frequency, and most of the parameter values are set to match moments computed using European data. The parameter values that need to be set are the subjective discount factor β, the coefficient of relative risk aversion σ, the moments of the stochastic process of government expenditures $\{\mu_g, \rho_g, \sigma_g\}$, the initial levels of government debt and expenditures (B_0, g_0), the level of income y, the initial wealth of $L-$type agents b_0^L and the default cost function $\phi(g_1)$. The calibrated parameter values are summarized in Table 14. We evaluate equilibrium outcomes for values of γ and ω in the [0,1] interval. Data for the United States and Europe documented in D'Erasmo and Mendoza (2013) suggest that the empirically relevant range for γ is [0.55,0.85]. Hence, when taking a stance on a particular value of γ is useful we use $\gamma = 0.7$, which is the mid point of the plausible range.

The preference parameters are set to standard values: $\beta = 0.96$, $\sigma = 1$. We also assume for simplicity that $L-$types start with zero wealth, $b_0^L = 0$.[ax] This and the other calibration parameters result in savings plans such that L-type agents are credit constrained, and hence $b_1^L = 0$.

We estimate an AR(1) process for government expenditures-GDP ratio (in logs) for France, Germany, Greece, Ireland, Italy, Spain and Portugal and set $\{\mu_g, \rho_g, \sigma_g\}$ to the cross-country averages of the corresponding estimates. This results in the following

[aw] We solve the model following a backward-recursive strategy analogous to the one used in the theoretical analysis. First, for each pair $\{\gamma, \omega\}$ and taking as given B_1, we solve for the equilibrium price and default functions by iterating on $\{d_1, q_0, b_1^i\}$. Then, in the second stage we complete the solution of the equilibrium by finding the optimal choice of B_1 that solves the government's date-0 optimization problem (25). As explained earlier, for given values of B_1, γ and ω an equilibrium with debt will not exist if either the government finds it optimal to default on B_1 for all realizations of g_1 or if at the given B_1 the consumption of L types is nonpositive.

[ax] $\sigma = 1$ and $b_0^L = 0$ are also useful because under these assumptions we can obtain closed-form solutions and establish some results analytically.

Table 14 Model parameters

Parameter		Value
Discount factor	β	0.96
Risk aversion	σ	1.00
Avg. Income	y	0.79
Low household wealth	b_0^L	0.00
Avg. gov. consumption	μ_g	0.18
Autocorrel. G	ρ_g	0.88
Std. dev. error	σ_g	0.017
Initial gov. debt	B_0	0.79
Output cost default	ϕ_0	0.02

Note: Government expenditures, income, and debt values are derived using Eurostat data for France, Germany, Greece, Ireland, Italy, Spain, and Portugal.

values $\mu_g = 0.1812$, $\rho_g = 0.8802$ and $\sigma_e = 0.017$. We set $g_0 = \mu_g$ and use the quadrature method proposed by Tauchen (1986) with 45 nodes in $G_1 \equiv \{\underline{g}_1, \ldots, \bar{g}_1\}$ to generate the realizations and transition probabilities of g_1.

Average income y is calibrated such that the model's aggregate resource constraint is consistent with the data when GDP is normalized to one. This implies that the value of the agents' aggregate endowment must equal GDP net of fixed capital investment and net exports, since the latter two are not modeled. The average for the period 1970–2012 for the same set of countries used to estimate the g_1 process implies $y = 0.7883$.[ay]

We set the initial debt level $B_0 = 0.79$ so that at the maximum observed level of inequality in the data, $\gamma = 0.85$, there is at least one feasible level of B_1 when $\omega = \gamma$. We assume that the default cost takes the following form: $\phi(g_1) = \phi_0 + (\bar{g} - g_1)/y$, where $\bar{g}$ is calibrated to represent an "unusually large" realization of g_1 set equal to the largest realization in the Markov process of government expenditures, which is in turn set equal to 3 standard deviations from the mean (in logs).[az]

We calibrate ϕ_0 to match an estimate of the observed frequency of *domestic* defaults. According to Reinhart and Rogoff (2011), historically, domestic defaults are about 1/4 as frequent as external defaults (68 domestic vs 250 external in their data since 1750). Since the probability of an external default has been estimated in the range of 3–5% (see, for example, Arellano, 2008), the probability of a domestic default is about 1%. The model is close to this default frequency on average when solved over the empirically relevant

[ay] Note also that under this calibration of y and the Markov process of g_1, the gap $y - g_1$ is always positive, even for $g_1 = \bar{g}_1$, which in turn guarantees $c_1^H > 0$ in all repayment states.

[az] This cost function shares a key feature of the default cost functions widely used in the external default literature to align default incentives so as to support higher debt ratios and trigger default during recessions (see Arellano, 2008 and Mendoza and Yue, 2012): The default cost is an *increasing* function of disposable income $(y - g_1)$. In addition, this formulation ensures that the agents' consumption during a default never goes above a given threshold.

range of γ's ($\gamma \in [0.55, 0.85]$) if we set $\phi_0 = 0.02$. Note, however, that the calibration of ϕ_0 and B_0 to match their corresponding targets needs to be done jointly by repeatedly solving the model until both targets are well approximated.

4.3.2 Utilitarian Government ($\omega = \gamma$)

We study first a set of results obtained under the assumption $\omega = \gamma$, because the utilitarian government is a natural benchmark. Since the default decision of the government derives from the agents' utility under the repayment and default alternatives at $t = 1$, it is useful to map the ordinal utility measures into cardinal measures by computing "individual welfare gains of default," which are standard consumption-equivalent values that equalize utility under default and repayment. Given the CRRA functional form, the individual welfare gains of default reduce simply to the percent changes in consumption across the default and no-default states of each agent at $t = 1$:

$$\alpha^i(B_1, g_1, \gamma) = \frac{c_1^{i,d=1}(B_1, g_1, \gamma)}{c_1^{i,d=0}(B_1, g_1, \gamma)} - 1 = \frac{(1 - \phi(g_1))\gamma - g_1}{\gamma - g_1 + b_1^i - B_1} - 1$$

A positive (negative) value of $\alpha^i(B_1, g_1, \gamma)$ implies that agent i prefers government default (repayment) by an amount equivalent to an increase (cut) of $\alpha^i(\cdot)$ percent in consumption. The individual welfare gains of default are aggregated using γ to obtain the utilitarian representation of the social welfare gain of default:

$$\overline{\alpha}(B_1, g_1, \gamma) = \gamma\alpha^L(B_1, g_1, \gamma) + (1 - \gamma)\alpha^H(B_1, g_1, \gamma).$$

A positive value indicates that default induces a social welfare gain and a negative value a loss.

Fig. 17 shows two intensity plots of the social welfare gain of default for the ranges of values of B_1 and γ in the vertical and horizontal axes, respectively. Panel (A) is for a low value of government purchases, $\underline{g}_1$, set 3 standard deviations below μ_g, and panel (B) is for a high value $\bar{g}_1$ set 3 standard deviations above μ_g. "No Equilibrium Zone", represent values of (B_1, γ) for which the debt market collapses and no equilibrium exists.[ba]

The area in which the social welfare gains of default are well defined in these intensity plots illustrates two of the key mechanisms driving the government's distributional incentives to default: First, fixing γ, the welfare gain of default is higher at higher levels of debt, or conversely the gain of repayment is lower. Second, keeping B_1 constant, the welfare gain of default is also increasing in γ (ie, higher wealth concentration increases

[ba] Note that to determine if $c_0^L \leq 0$ at some (B_1, γ) we also need $q_0(B_1, \gamma)$, since combining the budget constraints of the L types and the government yields $c_0^L = \gamma - g_0 - B_0 + q_0 B_1$. Hence, to evaluate this condition we take the given B_1 and use the H-types Euler equation and the market clearing condition to solve for $q_0(B_1, \gamma, \omega)$, and then determine if $\gamma - g_0 - B_0 + q_0 B_1 \leq 0$, if this is true, then (B_1, γ) is in the lower no-equilibrium zone.

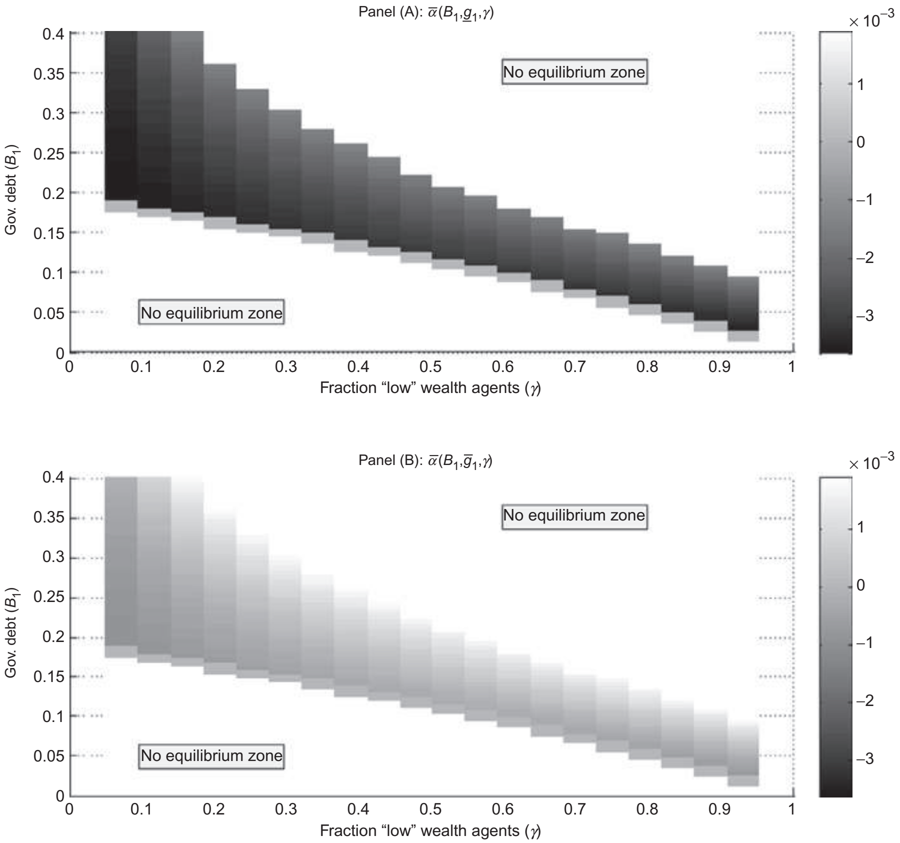


Fig. 17 Social welfare gains of default $\overline{\alpha}(B_1, g_1, \gamma)$. *Note:* The intensity of the color or shading in these plots indicates the magnitude of the welfare gain according to the legend shown to the right of the plots. The regions shown in white and marked as "no equilibrium zone," represent values of (B_1, γ) for which the debt market collapses and no equilibrium exists.

the welfare gain of default). This implies that lower levels of wealth dispersion are sufficient to trigger default at higher levels of debt.[bb] For example, for a debt ratio of 20% of GDP ($B_1 = 0.20$) and $g_1 = \bar{g}_1$, social welfare is higher under repayment if $0 \leq \gamma \leq 0.25$ but it becomes higher under default if $0.25 < \gamma \leq 0.6$, and for higher γ there is no equilibrium because the government prefers default not only for $g_1 = \bar{g}_1$ but for all possible g_1. If instead the debt is 40% of GDP, then social welfare is higher under default for all the values of γ for which an equilibrium exists.

[bb] Note that the cross-sectional variance of initial debt holdings is given by $Var(b) = B^2 \dfrac{\gamma}{1-\gamma}$ when $b_0^L = 0$. This implies that the cross-sectional coefficient of variation is equal to $CV(b) = \dfrac{\gamma}{1-\gamma}$, which is increasing in γ for $\gamma \leq 1/2$.

The two panels in Fig. 17 differ in that panel (B) displays a well-defined transition from a region in which repayment is socially optimal ($\overline{\alpha}(B_1,g_1,\gamma)<0$) to one where default is optimal ($\overline{\alpha}(B_1,g_1,\gamma)>0$) but in panel (A) the social welfare gain of default is never positive, so repayment is always optimal. This reflects the fact that higher g_1 also weakens the incentives to repay. In the "No Equilibrium Zone" in the upper right, there is no equilibrium because at the given γ the government chooses to default on the given B_1 for all values of g_1. In the "No Equilibrium Zone" in the lower left, there is no equilibrium because the given (B_1, γ) would yield $c_0^L \leq 0$, and so the government would not supply that particular B_1.

Consider next the government's default decision choice, which is driven by the sign of the social welfare gains of default. It is evident from Fig. 17 that the government defaults the higher g_1 for given B_1 and γ, the higher B_1 for a given γ and g_1, or at higher γ at given B_1 and g_1. It follows then that we can compute a threshold value of γ such that the government is indifferent between defaulting and repaying in period $t=1$ for a given (B_1,g_1). These indifference thresholds ($\hat{\gamma}(B_1,g_1)$) are plotted in Fig. 18 against debt levels ranging from 0 to 0.4 for three values of government expenditures $\{\underline{g}_1, \mu_g, \overline{g}_1\}$. For any given ($B_1$, g_1), the government chooses to default if $\gamma \geq \hat{\gamma}$.

Fig. 18 shows that the default threshold is decreasing in B_1. Hence, the government tolerates higher debt ratios without defaulting only if wealth concentration is sufficiently low. Also, default thresholds are decreasing in g_1, because the government has stronger

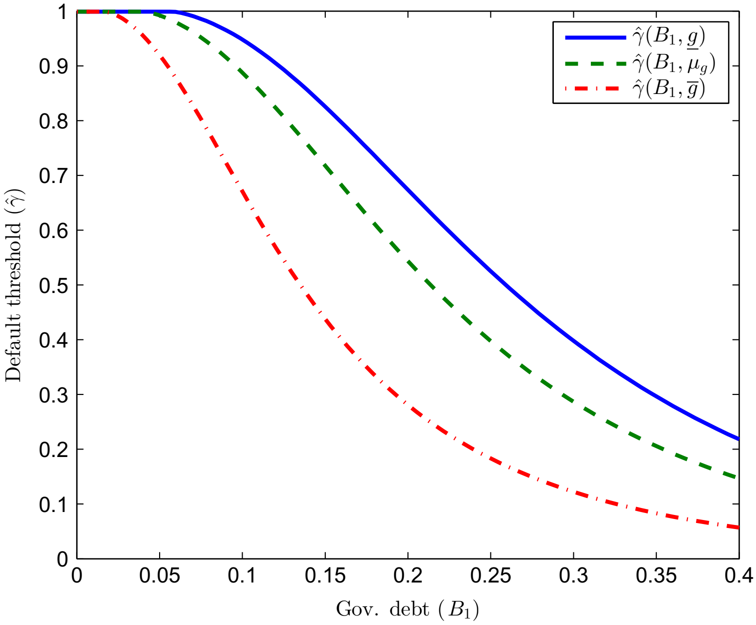


Fig. 18 Default threshold $\hat{\gamma}(B_1,g_1)$.

incentives to default when government expenditures are higher (ie, the threshold curves shift inward).[bc] This last feature of $\hat{\gamma}$ is very important to determine equilibria with sustainable debt subject to default risk. If, for a given value of B_1, γ is higher than the curve representing $\hat{\gamma}$ for the lowest realization in the Markov process of g_1 (which is also the value of $\underline{g_1}$), the government defaults for sure and, as explained earlier, there is no sustainable debt at equilibrium. Alternatively, if for a given value of B_1, γ is lower than the curve representing $\hat{\gamma}$ for the highest realization of g_1 (which is the value of $\overline{g}_1$), the government repays for sure and debt would be issued effectively without default risk. Thus, for the model to support equilibria with sustainable debt subject to default risk, the optimal debt chosen by the government in the first period for a given γ must lie between these two extreme threshold curves. We show below that this is the case in this quantitative experiment.

Before showing those results, it is important to highlight three key properties of the bond pricing function. The quantitative results for this function, the details of which we omit to save space, reflect the properties discussed in the model analysis:

1. *The equilibrium price is decreasing in $B_1$for given γ* (the pricing functions shift downward as B_1 rises). This follows from a standard demand-and-supply argument: For a given γ, as the government borrows more, the price at which the H types are willing to demand the additional debt falls and the interest rate rises.
2. *Default risk reduces the price of bonds below the risk-free price and thus induces a risk premium.* Intuitively, when there is no default risk (ie, for combinations of B_1 and γ such that the probability of default is zero) both prices are identical. However, as the probability of default rises, agents demand a premium in order to clear the bond market.
3. *Bond prices are a nonmonotonic function of wealth dispersion*: When default risk is sufficiently low, bond prices are increasing in γ, but eventually they become a steep decreasing function of γ. Higher γ implies a more dispersed wealth distribution, so that H-type agents become a smaller fraction of the population, and hence they must demand a larger amount of debt per capita in order to clear the bond market (ie, b_1^H increases with γ), which pushes bond prices up. While default risk is low this "demand composition effect" dominates and thus bond prices rise with γ, but as γ increases and default risk rises (since higher wealth dispersion strengthens default incentives), the growing risk premium becomes the dominating force (at about $\gamma > 0.5$) and produces bond prices that fall sharply as γ increases.

Finally we examine the numerical solutions of the model's full equilibrium with optimal debt and default policies. The key element of the solution is the sustainable debt, which is also the government's optimal choice of debt issuance in the first period at the equilibrium price (ie, the optimal B_1 that solves problem (25)). We show this sustainable debt as an equilibrium manifold (ie, as a plot of the sustainable debt obtained by

[bc] $\hat{\gamma}$ approaches zero for B_1 sufficiently large, but in Fig. 18 B_1 reaches 0.40 only for exposition purposes.

solving the model's equilibrium over a range of values of γ). Given this sustainable debt, we can then use the functions that describe optimal debt demand plans of private agents in both periods, the government's default choice in period 1, bond prices, and default risk for *any* value of B_1 to determine the corresponding *equilibrium* manifold values of all of the model's endogenous variables.

Fig. 19 shows the four main components of the equilibrium manifolds: Panel (A) plots the manifold of sustainable first-period debt issuance of the model with default risk, $B_1^*(\gamma)$, and also, for comparison, the debt in the case when the government is committed to repay so that debt is risk free, $B_1^{RF}(\gamma)$. Panel (B) shows equilibrium debt prices that correspond to the sustainable debt of the same two economies. Panel (C) shows the default spread (the difference in the inverses of the bond prices). Panel (D) shows the probability of default. Since in principle the government that has the option to default can still choose a debt level for which it could prefer to repay in all realizations of g_1,

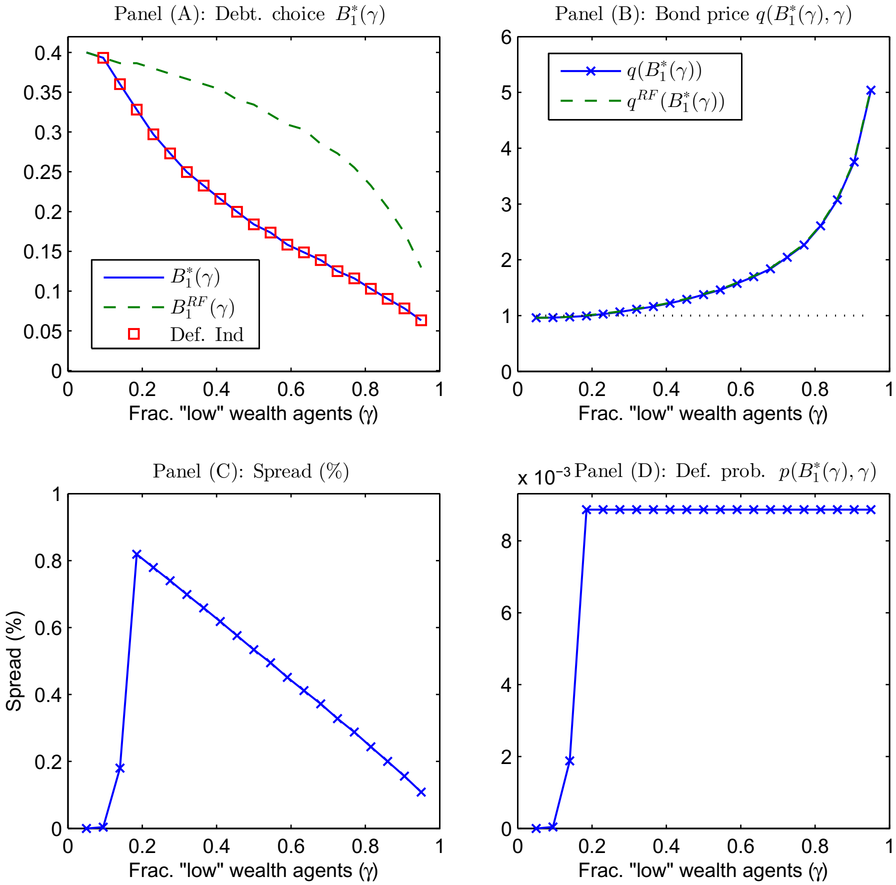


Fig. 19 Equilibrium manifolds.

we identify with a square in Panel (A) the equilibria in which $B_1^*(\gamma)$ has a positive default probability. This is the case for all but the smallest value of gamma considered ($\gamma = 0.05$), in which the government sets $B_1^*(\gamma)$ at 40% of GDP with zero default probability.

Panel (A) shows that sustainable debt falls as γ increases in both the economy with default risk and the economy with a government committed to repay. This occurs because in both cases the government seeks to reallocate consumption across agents and across periods by altering the product $q(B_1)B_1$ optimally, and in doing this it internalizes the response of bond prices to its debt choice. As γ rises, this response is influenced by stronger default incentives and a stronger demand composition effect. The latter dominates in this quantitative experiment, because panel (B) shows that the equilibrium bond prices always rise with γ. Hence, the government internalizes that as γ rises the demand composition effect strengthens demand for bonds, pushing bond prices higher, and as a result it can actually attain a higher $q(B_1)B_1$ by choosing lower B_1. This is a standard Laffer curve argument: In the upward slopping segment of this curve, increasing debt increases the amount of resources the government acquires by borrowing in the first period.

In the range of empirically relevant values of γ, sustainable debt ratios range from 20% to 32% of GDP without default risk and from 8% to 15% with default risk. Since the median in the European data is 35%, these ratios are relatively low, but still they are notable given the simplicity of the two-period setup. In particular, the model lacks the stronger income- and tax-smoothing effects and the self-insurance incentives of a longer life horizon (see Aiyagari and McGrattan, 1998), and it has an upper bound on the optimal debt choice for $\gamma = [0,1]$ lower than $B_0/(1+\beta)$ (which is the upper bound as $\gamma \rightarrow 0$ in the absence of default risk).

Panel (B) shows that bond prices of sustainable debt range from very low to very high as γ rises, including prices sharply above 1 that imply large negative real interest rates on public debt. In fact, as D'Erasmo and Mendoza (2013) explain, equilibrium bond prices are similar and increasing in γ with or without default risk, because at equilibrium the government chooses debt positions for which default risk is low (see panel (D)), and thus the demand composition effect that strengthens as γ rises dominates and yields bond prices increasing in γ and similar with or without default risk.[bd]

[bd] Everything else equal, our model predicts that higher income dispersion (either due to less progressive tax systems or underlying households' income or bond positions) results in higher spreads. In D'Erasmo and Mendoza (2013), we show that an economy with more progressive tax system results in lower spreads. The intuition is simple. The more the government can redistribute via means of taxation the lower the incentives to redistribute through a domestic default on the debt. The results in that paper show that incentives to default do not disappear but spreads decrease considerably. Also in D'Erasmo and Mendoza (2013), we present evidence of the nonlinear relationship between debt to income ratios and wealth inequality. Data limitations prevents us from extending this analysis to the relationship between spreads and income dispersion or the progressivity of the tax system.

Panels (C) and (D) show that, in contrast with standard models of external default, in this model the default spread is neither similar to the probability of default nor does it have a monotonic relationship with it.[be] Both the spread and the default probability start at zero for $\gamma = 0.05$ because $B_1^*(0.05)$ has zero default probability. As γ increases up to 0.2, both the spread and the default probability of the sustainable debt are similar in magnitude and increase together, but for $\gamma > 0.2$ the spread falls with γ while the default probability remains unchanged around 0.9%. For $\gamma = 0.95$ the probability of default is 9 times larger than the spread (0.9 vs 0.1%).

The role of the government's incentives to reallocate consumption across agents and across periods internalizing the response of bond prices when choosing debt can be illustrated further by examining the debt Laffer curve. Fig. 20 shows debt Laffer curves for five values of γ in the [0.05,0.95] range.

In all but one case, the sustainable debt $B_1^*(\gamma)$ (ie, the equilibrium debt chosen optimally by the government at the equilibrium price) is located at the maximum of the corresponding Laffer curve. In these cases, setting debt higher than at the maximum is

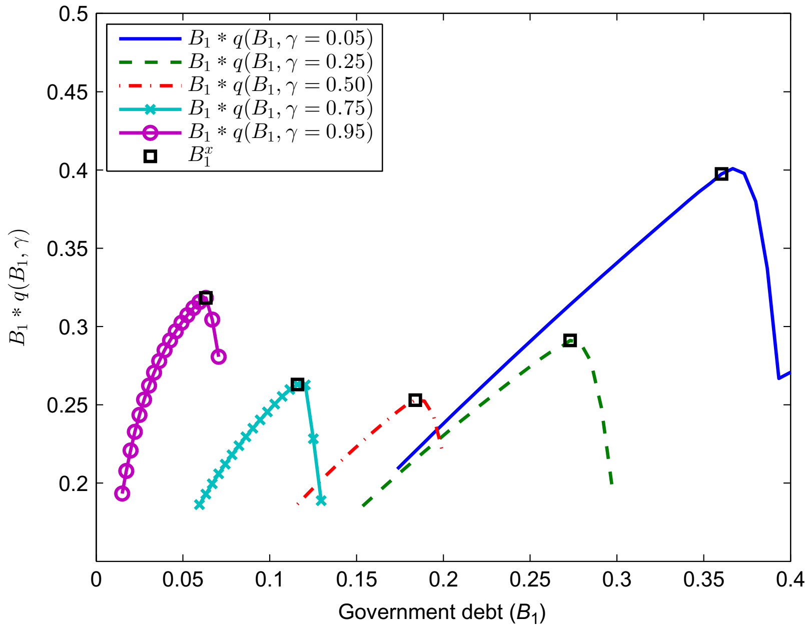


Fig. 20 Debt Laffer curve. *Note*: Each curve is truncated at values of B_1 in the horizontal axis that are either low enough for $c_0^L \leq 0$ or high enough for default to be chosen for all realizations of g_1, because as noted before in these cases there is no equilibrium.

[be] In the standard models, the two are similar and a monotonic function of each other because of the arbitrage condition of a representative risk-neutral investor.

suboptimal because default risk reduces bond prices sharply, moving the government to the downward-sloping segment of the Laffer curve. Setting debt lower than the maximum is also suboptimal, because then default risk is low and extra borrowing generates more resources since bond prices change little, leaving the government in the upward-sloping segment region of the Laffer curve. Thus, if the optimal debt has a nontrivial probability of default, the government's debt choice exhausts its ability to raise resources by borrowing. The exception is the case with $\gamma = 0.05$, in which $B_1^*(\gamma)$ has zero default probability. In this case, the government's optimal debt is to the left of the maximum of the Laffer curve, and thus the debt choice does not exhaust the government's ability to raise resources by borrowing. This also happens when the default probability is positive but negligible. For example, when $\gamma = 0.15$ the default probability is close to zero and the optimal debt choice is again slightly to the left of the maximum of the corresponding Laffer curve.

4.3.3 Biased Welfare Weights ($\omega \neq \gamma$)

The final experiment we conduct examines how the results change if we allow the weights of the government's payoff function to display a bias in favor of bondholders. Fig. 21 shows how the planner's welfare gain of default varies with ω and γ for two different levels of government debt ($B_{1,L} = 0.143$ and $B_{1,H} = 0.185$). The no-equilibrium region, which exists for the same reasons as before, is shown in white.

In line with the previous discussion, within the region where the equilibrium is well-defined, the planner's value of default increases monotonically as ω increases, keeping γ constant, and falls as actual wealth concentration (γ) rises, keeping ω constant. Because of this, the north-west and south-east corners in each of the panels present cases that are at very different positions on the preference-for-default spectrum. When ω is low, even for very high values of γ, the government prefers to repay (north-west corner), because the government puts relatively small weight on L-type agents. On the contrary, when ω is high, even for low levels of γ, a default is preferred. It is also interesting to note that as we move from Panel (A) to Panel (B), so that government debt raises, the set of γ's and ω's such that the equilibrium exists or repayment is preferred (ie, a negative $\overline{\alpha}(B_1, g_1, \gamma, \omega)$) expands. This is because as we increase the level of debt B_1, as long as the government does not choose to default for all g_1, the higher level of debt allows L-type agents to attain positive levels of consumption (since initial taxes are lower).

Panels (A)–(D) in Fig. 22 display the model's equilibrium outcomes for the sustainable debt chosen by the government in the first period and the associated equilibrium bond prices, spreads and default probabilities under three possible values of ω, all plotted as functions of γ. It is important to note that along the blue curve of the utilitarian case both ω and γ effectively vary together because they are always equal to each other, while in the other two plots ω is fixed and γ varies. For this reason, the line corresponding to the ω_L

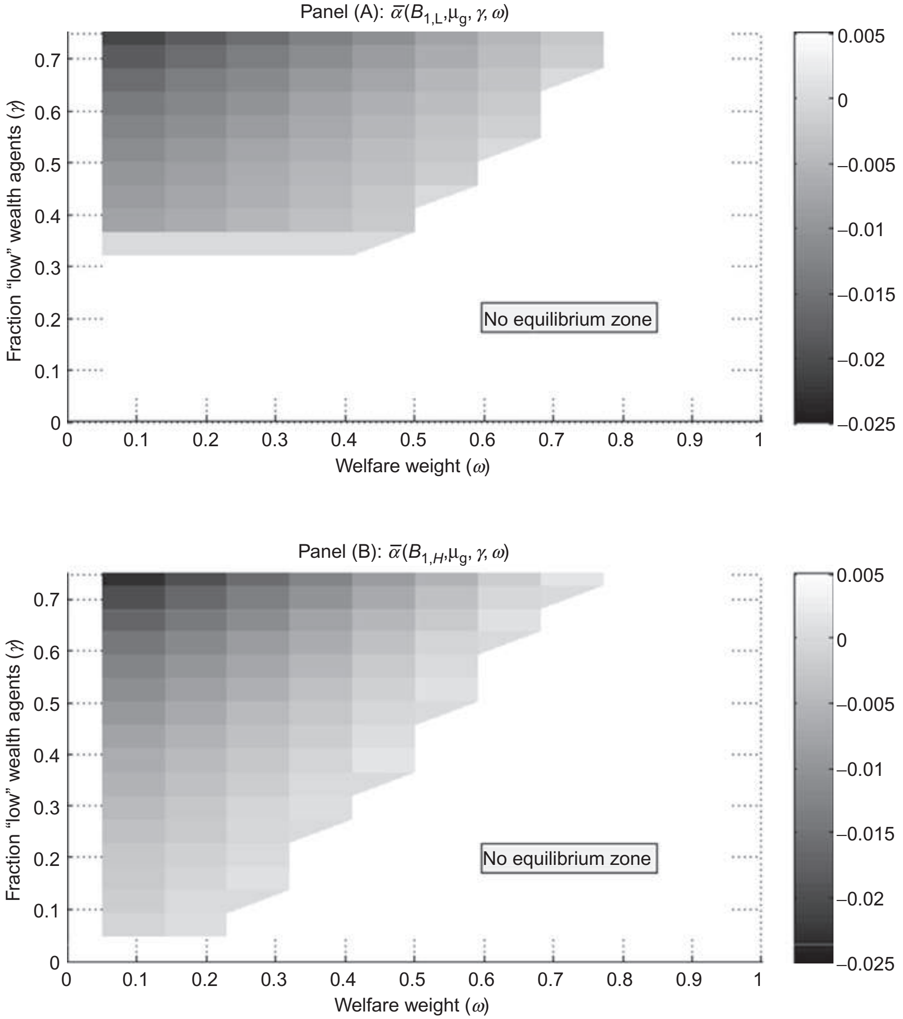


Fig. 21 Planner's welfare gain of default $\bar{\alpha}(B_1, g_1, \gamma, \omega)$.

case intersects the benchmark solution when $\gamma = 0.32$, and the one for ω_H intersects the benchmark when $\gamma = 0.50$.

Fig. 22 shows that the optimal debt level is increasing in γ. This is because the incentives to default grow weaker and the repayment zone widens as γ increases for a fixed value of ω. It is also interesting to note that in the ω_L and ω_H cases the equilibrium exists only for a small range of values of γ that are lower than ω. Without default costs each curve would be truncated exactly where γ equals either ω_H or ω_H, but since these simulations retain the default costs used in the utilitarian case, there can still be equilibria with debt for some lower values of γ (as explained earlier).

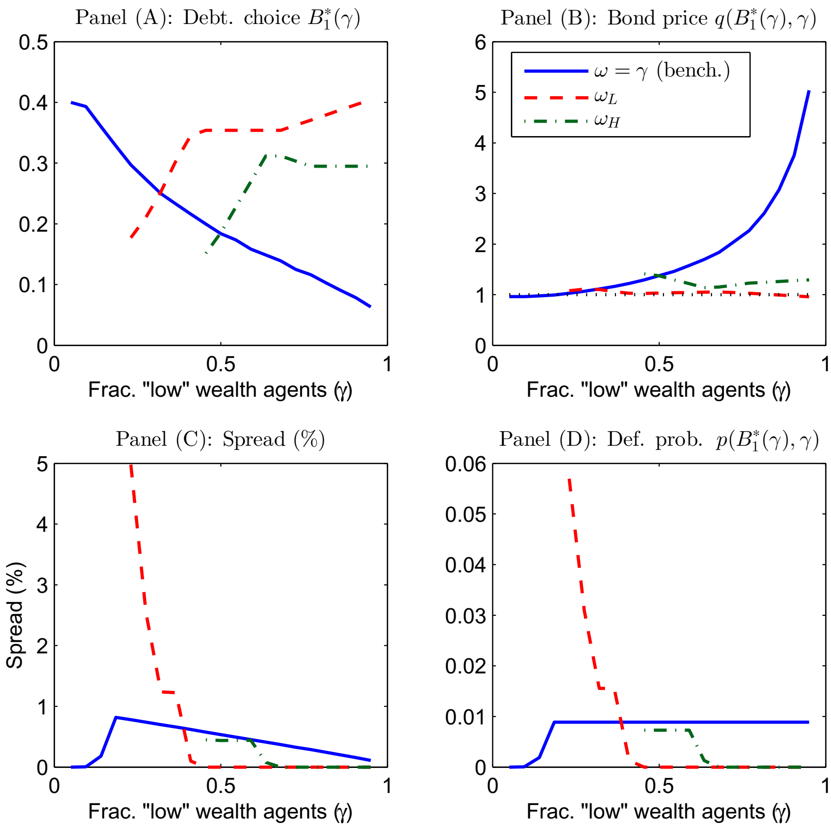


Fig. 22 Equilibrium manifolds with government bias at different values of ω.

With the bias in favor of bondholders, the government is still aiming to optimize debt by focusing on the resources it can reallocate across periods and agents, which are still determined by the debt Laffer curve $q_0(.)B_1$, and internalizing the response of bond prices to debt choices.[bf] This relationship, however, behaves very differently than in the benchmark model, because now *higher* sustainable debt is carried at increasing equilibrium bond prices, which leads the planner internalizing the price response to choose higher debt, whereas in the benchmark model *lower* debt was sustained at increasing equilibrium bond prices, which led the planner internalizing the price response to choose lower debt.[bg]

[bf] When choosing B_1, the government takes into account that higher debt increases disposable income for L-type agents in the initial period but it also implies higher taxes in the second period (as long as default is not optimal). Thus, the government is willing to take on more debt when ω is lower.

[bg] Fig. 22 makes clear that with the government bias, the level of sustainable debt changes with the preferences of the government. Even though we do not model how these preferences arise, it is evident that two countries with the same fundamentals (ie, distribution of wealth and income) could end up with very different levels of sustainable debt depending on how household preferences are aggregated by the government in power.

The behavior of equilibrium bond prices (panel (B)) with either $\omega_L = 0.32$ or $\omega_H = 0.50$ differs markedly from the utilitarian case. In particular, the prices no longer display an increasing, convex shape, instead they are a relatively flat and nonmonotonic function of γ. This occurs because the higher supply of bonds that the government finds optimal to provide offsets the demand composition effect that increases individual demand for bonds as γ rises.

The domestic default approach to study sustainable debt adds important insights to those obtained from the empirical and structural approaches, both of which assumed repayment commitment. In particular, panel (A) of Fig. 19 shows that sustainable debt falls sharply once risk of default is present, even when it is very small, and that (if the government is utilitarian) sustainable debt falls sharply with the concentration of bond ownership, because of the strengthened incentive to use default as a tool for redistribution. Hence, estimates of sustainable debt based on models in which the government is assumed to be committed to repay are likely to be too optimistic. Intuitively, one can infer that in the structural model, a given increase in the initial debt would be harder to offset with higher primary balances if the interest rate at which those primary balances are discounted rises with higher debt because of default risk. Moreover, the representative-agent assumption is also likely to lead to optimistic estimates of sustainable debt, because representative-agents models abstract from the strong incentives to use debt default as a tool for redistribution across heterogeneous agents. These incentives are likely to be weaker than in the model in practice, because tax and transfer policies that we did not include in the model can be used for redistribution as well. But when these other instruments have been exhausted, and if inequality in bond holdings is sufficiently concentrated, the incentives to default as vehicle for redistribution are likely to be very strong.

A second important insight from this analysis is that sustainable debt is higher if the government's payoff function is biased in favor of bondholders, and can even exceed debt that is sustainable without default risk when the government has a utilitarian social welfare function. Furthermore, D'Erasmo and Mendoza (2013) show that nonbondholders may prefer equilibria where the government favors bondholders, instead of being utilitarian, because higher sustainable debt help relax their liquidity constraints. Hence, at sufficiently high levels of concentration of bond ownership, a biased government can sustain high debt and the biased government can be elected as a majority government.

The main caveat of this analysis is that, because it was based on a two-period model, it misses important endogenous costs of default that would be added to the model by introducing a longer life horizon. In this case, default costs due to the reduced ability to smooth taxation and consumption when the debt market closes, and due to the loss of access to the self-insurance vehicle and the associated tightening of liquidity constraints, can take up the role of the exogenous default costs and/or government bias for bondholders, enabling the model to improve its ability to account for key features of the data and sustain higher debt levels at nontrivial default premia. D'Erasmo and Mendoza (2014) examine a model with these features and study its quantitative implications.

5. CRITICAL ASSESSMENT AND OUTLOOK

We started this chapter by noting that the question of what is a sustainable public debt has always been paramount in the macroeconomics of fiscal policy. The question will remain paramount for years to come, as the precarious public debt and deficit positions of many advanced and emerging economies today will make it a central focus of both policy analysis and academic research. This chapter aimed to demonstrate the flaws that affect the classic, but still widely used, approach to analyze public debt sustainability, and to show how three approaches based on recent research can provide powerful alternative ways to tackle the question. Two of these approaches, the empirical approach and the structural approach, assume that the government is committed to repay its debt, and the third approach, the domestic default approach, assumes that the government cannot commit to repay. In this section, we reflect further on the limitations of each of these approaches and suggest directions for further research.

The empirical approach has been widely studied and is by now very well established. Its strengths are in that it can easily determine whether debt has been consistent with fiscal solvency in available time-series data via straightforward estimation of a fiscal reaction function, and in that analyzing the characteristics of this FRF it can shed light on the dynamics of adjustment of debt and the primary balance. Unfortunately, as we explained earlier, it is not helpful for comparing alternative fiscal policy strategies to maintain debt sustainability and/or cope with public debt crises in the future.

The structural approach showed how an explicit dynamic general equilibrium model can be used to compare alternative fiscal policy strategies aimed at maintaining fiscals solvency at different levels of observed outstanding debt. We used a variation of the workhorse two-country Neoclassical framework with exogenous, balanced growth in which endogenous capacity utilization and a limited tax allowance for capital depreciation allow the model to match a key feature of the data for fiscal sustainability analysis: The observed elasticity of the capital tax revenue. Yet, the model is also very limited inasmuch as it abstracts from other important features of the data. In particular, the model is purely "real," and hence abstracts from the fact that public debt is largely nominal debt denominated in domestic currencies, and also abstracts from linkages between potentially important nominal rigidities, relative prices, and the evolution of government revenues and outlays.

The model used in the structural approach also has the drawbacks that it abstracts from heterogeneity in households and firms and assumes that agents are infinitely lived. Hence, while it takes into account important efficiency effects resulting from alternative fiscal policies, it cannot capture their distributional implications across agents and/or generations. The fiscal policy research on heterogeneous-agents and overlapping-generations models has shown that these distributional effects can be quite significant, and hence it is important to develop models of debt sustainability that incorporate them. For instance, Aiyagari (1995) showed that reductions in capital taxes have adverse distributional consequences that can offset the efficiency gains emphasized in representative agent models. Aiyagari and McGrattan (1998) showed that public debt has social value because it acts as vehicle to provide liquidity (ie, relax borrowing constraints) of the agents at the low end of the wealth distribution, and Birkeland and Prescott (2006) provide a setup in which using debt to save for retirement dominates a tax-and-transfer system. Imrohoroglu et al. (2016) and Braun and Joines (2015) also show how sophisticated overlapping-generations models can be applied to study debt-sustainability issues, with a particular focus on the implications of the adverse demographics dynamics facing Japan.

Of the approaches to debt sustainability analysis reviewed here, the domestic default approach is the one that has been studied the least. We provided a very simple canonical model in which default on domestic debt can emerge as an optimal outcome for a government with incentives to redistribute across debt holders and nonholders, but clearly significant further research in this area is needed (in addition to the recent work by D'Erasmo and Mendoza (2014) and Dovis et al. (2014) that we cited).

There are also two other directions in which research on debt sustainability should go. First, to model the role of public debt in financial intermediation in general and in financial stabilization policies in particular. In terms of the former, domestic banking systems are often large holders of domestic public debt, so a domestic default of the kind the third approach we examined seeks to explain tends to materialize in terms of a redistribution that hurts the balance sheets of banks. A deeper question in a similar vein is why public debt is such a high-demand asset, or liquidity vehicle, in modern financial systems. Macro/finance research is looking into this questions, but introducing these considerations into debt sustainability analysis is still a pending task. Regarding crisis-management policies, the aftermath of the global financial crisis has been characterized by strong demand for public debt instruments driven by quantitative easing policies and by the new regulatory environment. This may account for the apparent paradox between the pessimistic fiscal prospects that the analysis of this chapter presents and the observation that we currently observe near-zero and even negative yields on the public debt of some advanced economies (ie, demand for public debt remains very strong despite the highly questionable capacity of governments to repay it through standard improvements of the primary fiscal balance). But to be certain we need a richer model of debt sustainability that incorporates both the long-term forces that drive the government's capacity to repay and short-term debt dynamics around a financial crisis in which demand for risk-free asset surges.

The second direction in which debt sustainability analysis needs to branch out is to develop tools to incorporate considerations of potential multiplicity of equilibria in public debt markets. The seminal work of Calvo (1988) showed how debt can move between two equilibria supported by self-fulfilling expectations. In one the debt is repaid because agents expect that the government will be able to access the debt market, and thus maintain the efficiency losses of taxation small enough to indeed generate enough revenue to repay. In the other, the government defaults because agents expect that it will not be able to access the debt market and will be forced into highly distorting levels of taxation that indeed result in revenues that are insufficient to repay. The external default literature has explored models with this kind of equilibrium multiplicity extensively, as documented in the corresponding chapter of this handbook, and theoretical work applying these ideas to domestic debt crises is also available, but research to incorporate this mechanism into quantitative models of domestic debt sustainability is still needed.

6. CONCLUSIONS

What is a sustainable public debt? Assuming that the government is committed to repay, the answer is a debt that satisfies the intertemporal government budget constraint (ie, a debt that is equal to the present discounted value of the primary fiscal balance). In this chapter we showed that the traditional approach to debt sustainability analysis is flawed. This approach uses the steady-state government budget constraint to define sustainable debt as the annuity value of the primary balance, but it cannot establish if current or projected debt and primary balance dynamics are consistent with that debt level. We then discussed two approaches to study public debt sustainability under commitment to repay: First, an empirical approach, based on a linear fiscal reaction function, according to which a positive, conditional response of the primary balance to debt is sufficient to establish debt sustainability. Second, a structural approach based on a two-country variant of the workhorse Neoclassical dynamic general equilibrium model with an explicit fiscal sector. The model differs from the standard Neoclassical setup in that it introduces endogenous capacity utilization and a limited tax allowance for depreciation expenses in order to match the observed elasticity of the capital tax base to changes in capital taxes. In this setup, the initial debt that is sustainable is the one determined by the present value of primary balances evaluated using equilibrium allocations and prices.

Applications of these first two approaches to cross-country data produced key insights. With the empirical approach, we found in tests based on historical US data and cross-country panels that the sufficiency condition for public debt to be sustainable (the positive, conditional response of the primary balance to debt), cannot be rejected. We also found, however, clear evidence showing that the fiscal dynamics observed in the aftermath of the recent surge in debt in advanced economies represent a significant structural break in the estimated reaction functions. Primary deficits have been too large, and are projected to remain too large, relative to what the fiscal reaction functions predict, and they are also large compared with those observed in the aftermath previous episodes of large surges in debt.

The structural approach differs from the empirical approach in that it can be used to evaluate the positive and normative effects of alternative paths of fiscal adjustment to attain debt sustainability, whereas the empirical approach is silent about these effects. We calibrated the model to the United States and European data and used it to quantify the effects of unilateral changes in capital and labor taxes, particularly their effects on sustainable debt. The results suggest key differences across Europe and the United States. For the United States, the results suggest that changes in capital taxes cannot make the observed increase in debt sustainable, while small increases in labor taxes could. For Europe, the model predicts that the capacity to use taxes to make higher debt ratios sustainable is nearly fully exhausted. Capital taxation is highly inefficient (in the decreasing segment of dynamic Laffer curves), so cuts in capital taxes would be needed to restore fiscal solvency. Labor taxes are near the peak of the dynamic Laffer curve, and even if increased to the maximum point they do not generate enough revenue to make the present value of the primary balance match the observed surge in debt. In addition, international externalities of capital income taxes were quantitatively large, which suggest that incentives for strategic interaction are nontrivial and could lead to a classic race-to-the-bottom in capital income taxation.

The results of the applications of the empirical and structural approaches paint a bleak picture of the prospects for fiscal adjustment in advanced economies to restore fiscal solvency and make the post-2008 surge in public debt ratios sustainable. In light of these findings, and with the ongoing turbulence in European sovereign debt markets and recurrent debt ceiling debates in the United States, we examined a third approach to debt sustainability that relaxes the assumption of a government committed to repay and allows for the risk of default on domestic public debt. In this environment, debt is sustainable when it is part of the equilibrium that includes the optimal debt issuance and default choices of the government. The government has incentives to default as a vehicle for redistribution across agents who are heterogeneous in wealth. Public debt is not sustainable in the absence of default costs or a political bias to weigh the welfare of bond holders by more than their share of the wealth distribution. This is the case because without these assumptions default is always the optimal choice that maximizes the social welfare function of a government who values the utility of all agents, and this is the case regardless of the present value of primary balances used to characterize sustainable debt under the other two approaches.

Quantitatively, this domestic default approach adds valuable insights to those obtained from the empirical and structural approaches without default risk. In particular, sustainable debt falls sharply once risk of default is present, even when it is very small, and it also falls sharply with wealth inequality, because of the strengthened incentive to use default as a tool for redistribution. Hence, estimates of sustainable debt based on models in which the government is assumed to be committed to repay are too optimistic. Moreover, the representative-agent assumption is also likely to lead to optimistic estimates of sustainable debt, because models in this class abstract from the strong incentives to use debt default as a tool for redistribution across heterogeneous agents. A second important insight from the domestic default approach is that sustainable debt is higher if the government's payoff function weighs the welfare of bond holders more heavily than their share of the wealth distribution. In addition, it is possible that low-wealth agents may also prefer that the government weights high-wealth agents more heavily, instead of acting as a utilitarian government, because higher debt stocks help relax their liquidity constraints.

The three approaches reviewed in this chapter provide useful tools for conducting debt sustainability analysis. When applied to the current fiscal situation of advanced economies, all three suggest that substantial fiscal adjustment is still needed, is likely to entail substantial welfare costs, and is likely to continue to be challenged by potential default risk in domestic sovereign debt markets.

APPENDIX: DETAILS ON MEASUREMENT OF EFFECTIVE TAX RATES

Effective tax rates have been widely used in a number of studies including Carey and Tchilinguirian (2000), Sorensen (2001), and recently by Trabandt and Uhlig (2011, 2012). The MRT methodology uses the wedge between reported pretax and post-tax macro estimates of consumption, labor income and capital income to estimate the effective tax rate levied on each of the three tax bases. This methodology has two main advantages. First, it provides a fairly simple approach to estimating effective tax rates at the macro level using readily available data, despite the complexity of the various credits and deductions of national tax codes. Second, these tax rates correspond directly to the tax rates in a wide class of representative-agent models with taxes on consumption and factor incomes, including the model proposed here. The main drawback of the MRT tax rates is that they are average, not marginal, tax rates, but because they are intended for use in representative-agent models, this disadvantage is less severe than it would be in a model with heterogeneous agents. Moreover Mendoza et al. (1994) show that existing estimates of aggregate marginal tax rates have a high time-series correlation with the MRT effective tax rates, and that both have similar cross-country rankings.

Following Trabandt and Uhlig (2011), we modify the MRT estimates of labor and capital taxes by adding supplemental wages (ie, employers' contributions to social security and private pension plans) to the tax base for personal income taxes. These data were not available at the time of the MRT 1994 calculations and, because this adjustment affects the calculation of the personal income tax rate, which is an initial step for the calculation of labor and capital income tax rates, it alters the estimates of both. In general, this adjustment makes the labor tax base bigger and therefore the labor tax rate smaller than the MRT original estimates. [bh]

[bh] Trabandt and Uhlig make a further adjustment to the MRT formulae by attributing some of the operating surplus of corporations and nonincorporated private enterprises to labor, with the argument that this represents a return to entrepreneurs rather than to capital. We do not make this modification because the data do not provide enough information to determine what fraction of the operating surplus should be allocated to labor.

ACKNOWLEDGMENTS

We are grateful to Juan Hernandez, Christian Probsting and Valentina Piamiotti for their valuable research assistance. We are also grateful to our discussant, Kinda Hachem, the Handbook editors, John Taylor and Harald Uhlig, and Henning Bohn for their valuable suggestions and comments. We also acknowledge comments by Jonathan Heathcote, Andy Neumeyer, Juan Pablo Nicolini, Martin Uribe, and Vivian Yue, and by participants attending presentations at the Bank for International Settlements, Emory University, the third RIDGE seminar on International Macroeconomics, and the April 2015 Handbook conference at the University of Chicago. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Chicago, the Federal Reserve Bank of Philadelphia, or the Federal Reserve System.

REFERENCES

Afonso, A., 2005. Fiscal Sustainability: the Unpleasant European Case. FinanzArchiv 61 (1), 19–44. http://ideas.repec.org/a/mhr/finarc/urnsici0015-2218(200503)611_19fstuec_2.0.tx_2-r.html.

Aghion, P., Bolton, P., 1990. Government domestic debt and the risk of default: a political-economic model of the strategic role of debt. In: Dornbusch, R., Draghi, M. (Eds.), Public Debt Management: Theory and History. Cambridge University Press, Cambridge, pp. 315–344.

Aguiar, M., Amador, M., 2013. Fiscal policy in debt constrained economies. NBER Working Papers 17457.

Aguiar, M., Amador, M., Farhi, E., Gopinath, G., 2013. Crisis and commitment: inflation credibility and the vulnerability to sovereign debt crises. NBER Working Papers 19516.

Aiyagari, S.R., 1995. Optimal capital income taxation with incomplete markets, borrowing constraints, and constant discounting. J. Polit. Econ. 103 (6), 1158–1175.

Aiyagari, R., McGrattan, E., 1998. The optimum quantity of debt. J. Monet. Econ. 42, 447–469.

Aiyagari, R., Marcet, A., Sargent, T., Seppala, J., 2002. Optimal taxation without state-contingent debt. J. Polit. Econ. 110 (6), 1220–1254.

Alesina, A., Tabellini, G., 1990. A positive theory of fiscal deficits and government debt. Rev. Econ. Stud. 57, 403–414.

Alesina, A., Tabellini, G., 2005. Why is fiscal policy often procyclical? National Bureau of Economic Research, Working Paper 11600. doi: 10.3386/w11600, http://www.nber.org/papers/w11600.

Amador, M., 2003. A political economy model of sovereign debt repayment. Mimeo, Stanford University.

Andreasen, E., Sandleris, G., der Ghote, A.V., 2011. The political economy of sovereign defaults. Universidad Torcuato Di Tella, Business School Working Paper.

Arellano, C., 2008. Default risk and income fluctuations in emerging economies. Am. Econ. Rev. 98 (3), 690–712.

Auray, S., Eyquem, A., Gomme, P., 2013. A tale of tax policies in open economies. Mimeo, Department of Economics, Concordia University.

Azzimonti, M., de Francisco, E., Quadrini, V., 2014. Financial globalization, inequality, and the rising public debt. Am. Econ. Rev. 104 (8), 2267–2302.

Barnhill Jr., M.T., Kopits, G., 2003. Assessing fiscal sustainability under uncertainty. IMF Working Paper, WP 03-79.

Barro, R., 1979. On the determination of the public debt. J. Polit. Econ. 87 (5), 940–971.

Basu, S., 2009. Sovereign debt and domestic economic fragility. Manuscript, Massachusetts Institute of Technology.

Birkeland, K., Prescott, E.C., 2006. On the needed quantity of government debt. Research Department, Federal Reserve Bank of Minneapolis, Working Paper 648.

Blanchard, O.J., 1990. Suggestions for a new set of fiscal indicators. OECD Economics Department Working Papers 79, OECD Publishing, http://ideas.repec.org/p/oec/ecoaaa/79-en.html.

Blanchard, O.J., Chouraqui, J.C., Hagemann, R.P., Sartor, N., 1990. The sustainability of fiscal policy: new answers to an old question. OECD Econ. Stud. 15 (2), 7–36.

Bocola, L., 2014. The pass-through of sovereign risk. Manuscript, University of Pennsylvania.

Bohn, H., 1995. The sustainability of budget deficits in a stochastic economy. J. Money Credit Bank. 27 (1), 257–271. http://ideas.repec.org/a/mcb/jmoncb/v27y1995i1p257-71.html.

Bohn, H., 1998. The behavior of U.S. public debt and deficits. Q. J. Econ. 113 (3), 949–963. http://ideas.repec.org/a/tpr/qjecon/v113y1998i3p949-963.html.

Bohn, H., 2007. Are stationarity and cointegration restrictions really necessary for the intertemporal budget constraint? J. Monet. Econ. 54 (7), 1837–1847. http://ideas.repec.org/a/eee/moneco/v54y2007i7p1837-1847.html.

Bohn, H., 2008. The sustainability of fiscal policy in the United States. In: Neck, R., Sturm, J.E. (Eds.), Sustainability of public debt. MIT Press, Cambridge, MA.

Bohn, H., 2011. The economic consequences of rising U.S. government debt: privileges at risk. Finanzarchiv 67 (3), 282–302.

Boz, E., D'Erasmo, P., Durdu, B., 2014. Sovereign risk and bank balance sheets: the role of macroprudential policies. Manuscript.

Braun, R.A., Joines, D.H., 2015. The implications of a graying Japan for government policy. J. Econ. Dyn. Control 57, 1–23.

Broner, F., Martin, A., Ventura, J., 2010. Sovereign risk and secondary markets. Am. Econ. Rev. 100 (4), 1523–1555.

Broner, F., Ventura, J., 2011. Globalization and risk sharing. Rev. Econ. Stud. 78 (1), 49–82.

Brutti, F., 2011. Sovereign defaults and liquidity crises. J. Int. Econ. 84 (1), 65–72.

Buiter, W.H., 1985. A guide to public sector debt and deficits. Econ. Policy 1 (1), 13–61.

Calvo, G., 1988. Servicing the public debt: the role of expectations. Am. Econ. Rev. 78 (4), 647–661.

Carey, D., Tchilinguirian, H., 2000. Average effective tax rates on capital, labour and consumption. OECD Economics Department Working Papers: 258.

Chalk, N.A., Hemming, R., 2000. Assessing fiscal sustainability in theory and practice. International Monetary Fund.

Chari, V.V., Christiano, L.J., Kehoe, P.J., 1994. Optimal fiscal policy in a business cycle model. J. Polit. Econ. 102 (4), 617–652.

Cooley, T.F., Hansen, G.D., 1992. Tax distortions in a neoclassical monetary economy. J. Econ. Theory 58 (2), 290–316. ISSN 0022-0531. doi:10.1016/0022-0531(92)90056-N. http://www.sciencedirect.com/science/article/pii/002205319290056N.

Cuadra, G., Sanchez, J., Sapriza, H., 2010. Fiscal policy and default risk in emerging markets. Rev. Econ. Dyn. 13 (2), 452–469.

Davies, J., Sandström, S., Shorrocks, A., Wolff, E., 2009. The level and distribution of global household wealth. NBER Working Paper 15508.

D'Erasmo, P., 2011. Government reputation and debt repayment in emerging economies. Mimeo.

D'Erasmo, P., Mendoza, E., 2013. Distributional incentives in an equilibrium model of domestic sovereign default. National Bureau of Economic Research, No. w19477.

D'Erasmo, P., Mendoza, E., 2014. Optimal domestic sovereign default. Manuscript, University of Pennsylvania.

Di Casola, P., Sichlimiris, S., 2014. Domestic and external sovereign debt. Stockholm School of Economics, Working Paper.

Dias, D., Richmond, C., Wright, M., 2012. In for a penny, in for a 100 billion pounds: quantifying the welfare benefits from debt relief. Mimeo.

Dixit, A., Londregan, J., 2000. Political power and the credibility of government debt. J. Econ. Theory 94, 80–105.

Dovis, A., Golosov, M., Shourideh, A., 2014. Sovereign debt vs redistributive taxes: financing recoveries in unequal and uncommitted economies. Mimeo.

Durdu, B.C., Mendoza, E.G., Terrones, M.E., 2013. On the solvency of nations: cross-country evidence on the dynamics of external adjustment. J. Monet. Econ. 32, 762–780.

Dwenger, N., Steiner, V., 2012. Profit taxation and the elasticity of the corporate income tax base: evidence from German corporate tax return data. Natl. Tax J. 65 (1), 117–150.

Eaton, J., Gersovitz, M., 1981. Debt with potential repudiation: theoretical and empirical analysis. Rev. Econ. Stud. 48 (2), 289–309.

Eichengreen, B., 1989. The capital Levy in theory and practice. National Bureau of Economic Research, Working Paper Series 3096.

Escolano, J., 2010. A practical guide to public debt dynamics, fiscal sustainability, and cyclical adjustment of budgetary aggregates. International Monetary Fund.

Ferraro, D., 2010. Optimal capital income taxation with endogenous capital utilization. Mimeo, Department of Economics, Duke University.

Ferriere, A., 2014. Sovereign default, inequality, and progressive taxation. Mimeo.

Floden, M., 2001. The effectiveness of government debt and transfers as insurance. J. Monet. Econ. 48, 81–108.

Frenkel, J., Razin, A., Sadka, E., 1991. The sustainability of fiscal policy in the United States. In: International Taxation in an Integrated World. MIT Press, Cambridge, MA.

Gali, J., 1991. Budget constraints and time-series evidence on consumption. Am. Econ. Rev. 81 (5), 1238–1253. http://ideas.repec.org/a/aea/aecrev/v81y1991i5p1238-53.html.

Gennaioli, N., Martin, A., Rossi, S., 2014. Sovereign default, domestic banks, and financial institutions. The Journal of Finance 69 (2), 819–866.

Ghosh, A.R., Kim, J.I., Mendoza, E.G., Ostry, J.D., Qureshi, M.S., 2013. Fiscal fatigue, fiscal space and debt sustainability in advanced economies. Econ. J. 123, F4–F30.

Golosov, M., Sargent, T., 2012. Taxation, redistribution, and debt with aggregate shocks. Princeton University, Working Paper.

Greenwood, J., Huffman, G.W., 1991. Tax analysis in a real-business-cycle model. J. Monet. Econ. 22 (2), 167–190.

Gruber, J., Rauh, J., 2007. How elastic is the corporate income tax base? In: Taxing Corporate Income in the 21st Century. Cambridge University Press, New York.

Guembel, A., Sussman, O., 2009. Sovereign debt without default penalties. Rev. Econ. Stud. 76, 1297–1320.

Hall, G., Sargent, T., 2014. Fiscal discrimination in three wars. J. Monet. Econ. 61, 148–166.

Hamilton, J.D., Flavin, M.A., 1986. On the Limitations of Government Borrowing: a Framework for Empirical Testing. Am. Econ. Rev. 76 (4), 808–819. http://ideas.repec.org/a/aea/aecrev/v76y1986i4p808-19.html.

Hansen, G., Imrohoroglu, S., 2013. Fiscal reform and government debt in Japan: a neoclassical perspective. National Bureau of Economic Research, No. w19477.

Hansen, L.P., Roberds, W., Sargent, T.J., 1991. Time series implications of present value budget balance and of martingale models of consumption and taxes. In: Hansen, L.P., Sargent, T.J., Heaton, J., Marcet, A., Roberds, W. (Eds.), Rational xpectations econometrics. Westview Press, Boulder, CO, pp. 121–161.

Hatchondo, J.C., Martinez, L., Sapriza, H., 2009. Heterogeneous borrowers in quantitative models of sovereign default. Int. Econ. Rev. 50, 129–151.

Heathcote, J., 2005. Fiscal policy with heterogeneous agents. Rev. Econ. Stud. 72, 161–188.

House, C.L., Shapiro, M.D., 2008. Temporary investment tax incentives: theory with evidence from bonus depreciation. Am. Econ. Rev. 98 (3), 737–768. doi: 10.1257/acr.98.3.737.

Huizinga, H., 1995. The optimal taxation of savings and investment in an open economy. Econ. Lett. 47 (1), 59–62.

Huizinga, H., Voget, J., Wagner, W., 2012. Who bears the burden of international taxation? Evidence from cross-border m&as. J. Int. Econ. 88, 186–197.

IMF, International Monetary Fund, 2003. World economic outlook. IMF Occasional Papers 21, International Monetary Fund.

IMF, 2013. Staff guidance note for public debt sustainability in market access countries. http://www.imf.org/external/np/pp/eng/2013/050913.pdf.

Imrohoroglu, S., Sudo, N., 2011. Productivity and fiscal policy in Japan: short-term forecasts from the standard growth model. Monetary Econ. Stud. 29, 73–106.

Imrohoroglu, S., Kirao, S., Yamada, T., 2016. Achieving fiscal balance in Japan. Int. Econ. Rev. 57 (1), 117–154.

Jeon, K., Kabukcuoglu, Z., 2014. Income inequality and sovereign default. University of Pittsburgh, Working Paper.

Kaminsky, G.L., Reinhart, C.M., Vegh, C.A., 2005. When it rains, it pours: procyclical capital flows and macroeconomic policies. In: NBER Macroeconomics Annual 2004, NBER Chapters, National Bureau of Economic Research, Inc., pp. 11–82. vol.19. http://ideas.repec.org/h/nbr/nberch/6668.html

King, R.G., Plosser, C.I., Rebelo, S.T., 1988. Production, growth and business cycles: I. the basic neoclassical model. J. Monet. Econ. 21 (2), 195–232.

Klein, P., Quadrini, V., Rios-Rull, J.V., 2007. Optimal time-consistent taxation with international mobility of capital. B.E. J. Macroecon. 5.1, 186–197.

Ljungqvist, L., Sargent, T.J., 2012. Recursive Macroeconomic Theory, third ed. The MIT Press, Cambridge, Massachusetts.

Lorenzoni, G., Werning, I., 2013. Slow moving debt crises. NBER Working Paper No. w19228.

Lucas, R.E., 1987. Models of Business Cycles. Basil Blackwell, Oxford.

Lucas, R.E., 1990. Why doesn't capital flow from rich to poor countries? In: Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association, Am. Econ. Rev. vol. 80, pp. 92–96.

Lucas, D., 2012. Valuation of government policies and projects. Ann. Rev. Financ. Econ. 4, 39–58.

Mendoza, E.G., Ostry, J.D., 2008. International evidence on fiscal solvency: is fiscal policy responsible. J. Monet. Econ. 55, 1081–1093.

Mendoza, E.G., Oviedo, P.M., 2006. Fiscal policy and macroeconomic uncertainty in emerging markets: the tale of the tormented insurer. 2006 Meeting Papers 377, Society for Economic Dynamics, http://ideas.repec.org/p/red/sed006/377.html, 2006 Meeting Papers.

Mendoza, E.G., Oviedo, P.M., 2009. Public debt, fiscal solvency and macroeconomic uncertainty in Latin America the cases of Brazil, Colombia, Costa Rica and Mexico. Econ. Mex. NUEVA POCA XVIII (2), 133–173. http://ideas.repec.org/a/emc/ecomex/v18y2009i2p133-173.html.

Mendoza, E.G., Tesar, L.L., 1998. The international ramifications of tax reforms: supply-side economics in a global economy. Am. Econ. Rev. 88 (1), 226–245.

Mendoza, E.G., Tesar, L.L., 2005. Why hasn't tax competition triggered a race to the bottom? Some quantitative lessons from the EU. J. Monet. Econ. 52 (1), 163–204.

Mendoza, E.G., Yue, V.Z., 2012. A general equilibrium model of sovereign default and business cycles. Q. J. Econ. 127 (2), 889–946.

Mendoza, E.G., Razin, A., Tesar, L.L., 1994. Effective tax rates in macroeconomics: cross-country estimates of tax rates on factor incomes and consumption. J. Monet. Econ. 34 (3), 297–323.

Mendoza, E.G., Milesi-Ferretti, G.M., Asea, P., 1997. On the ineffectiveness of tax policy in altering long-run growth: Harberger's superneutrality conjecture. J. Public Econ. 66 (2), 99–126.

Mendoza, E.G., Tesar, L.L., Zhang, J., 2014. Saving Europe? the unpleasant arithmetic of fiscal austerity in integrated economies. University of Michigan Working Paper.

Mengus, E., 2014. Honoring sovereign debt or bailing out domestic residents? A theory of internal cost of default. WP Banque de France 480.

Neck, R., Sturm, J.E., 2008. Sustainability of Public Debt. MIT Press, Cambridge.

Ostry, J.D., David, J., Ghosh, A., Espinoza, R., 2015. When should public debt be reduced? International Monetary Fund, Staff Discussion Notes No. 15/10.

Perez, D., 2015. Sovereign debt, domestic banks and the provision of public liquidity. Manuscript.

Persson, T., Tabellini, G., 1995. Double-edged incentives: institutions and policy coordination. In: Grossman, G., Rogoff, K. (Eds.), Handbook of International Economics, vol. III. North-Holland, Amsterdam.

Pouzo, D., Presno, I., 2014. Optimal taxation with endogenous default under incomplete markets. U.C. Berkeley, Mimeo.

Prescott, E.C., 2004. Why Do Americans Work So Much More Than Europeans? Federal Reserve Bank of Minneapolis Quarterly Review 28 (1), 2–13.

Quintos, C.E., 1995. Sustainability of the Deficit Process with Structural Shifts. J. Bus. Econ. Stat. 13 (4), 409–417. http://ideas.repec.org/a/bes/jnlbes/v13y1995i4p409-17.html.

Reinhart, C.M., Rogoff, K.S., 2011. The forgotten history of domestic debt. Econ. J. 121 (552), 319–350. ISSN 1468-0297. doi:10.1111/j.1468-0297.2011.02426.x.

Sorensen, P., 2003. International tax coordination: regionalism versus globalism. J. Public Econ. 88, 1187–1214.

Sorensen, P.B., 2001. Tax coordination and the European Union: what are the issues? University of Copenhagen, Working Paper.

Sosa-Padilla, C., 2012. Sovereign defaults and banking crises. Manuscript.

Tabellini, G., 1991. The politics of intergenerational redistribution. J. Polit. Econ. 99, 335–357.

Talvi, E., Vegh, C.A., 2005. Tax base variability and procyclical fiscal policy in developing countries. J. Dev. Econ. 78 (1), 156–190. http://ideas.repec.org/a/eee/deveco/v78y2005i1p156-190.html.

Tauchen, G., 1986. Finite state Markov-chain approximation to univariate and vector autoregressions. Econ. Lett. 20, 177–181.

Trabandt, M., Uhlig, H., 2011. The Laffer curve revisited. J. Monet. Econ. 58 (4), 305–327.

Trabandt, M., Uhlig, H., 2012. How do laffer curves differ across countries? BFI Paper no. 2012-001.

Trehan, B., Walsh, C., 1988. Common trends, the government's budget constraint, and revenue smoothing. J. Econ. Dyn. Control 12 (2-3), 425–444. http://EconPapers.repec.org/RePEc:eee:dyncon:v:12:y:1988:i:2-3:p:425-444.

Vasishtha, G., 2010. Domestic versus external borrowing and fiscal policy in emerging markets. Rev. Int. Econ. 18 (5), 1058–1074.

CHAPTER 33

The Political Economy of Government Debt

A. Alesina[*,†], A. Passalacqua[*]
[*]Harvard University, Cambridge, MA, United States
[†]IGIER, Bocconi University, Milan, Italy

Contents

Handbook of Macroeconomics, Volume 2B
ISSN 1574-0048, http://dx.doi.org/10.1016/bs.hesmac.2016.03.014

Abstract

This chapter critically reviews the literature which explains why and under which circumstances governments accumulate more debt than it would be consistent with optimal fiscal policy. We also discuss numerical rules or institutional designs which might lead to a moderation of these distortions.

Keywords

Political economy, Optimal taxation, Budget rules, Government debt

JEL Classification Codes

E62, H63, H21

1. INTRODUCTION

Fiscal policy is deeply intertwined with politics since it is mostly about redistribution across individuals, regions, and generations: the core of political conflict. The redistributive role of governments has been increasing over time starting with the welfare programs introduced during the Great Depression and then with the additional jumps in the sixties and seventies of last century. But even recently the size of social spending (as defined by the OECD[a]) in 18 OECD countries jumped from 18% of GDP in 1980 to 26% in 2014.[b] In addition, the provision of public goods, which is therefore not classified as directly redistributive, has a redistributive component to the extent that public goods are used more or less intensively by individuals in different income brackets. The structure of taxation, such as the progressivity of the income tax brackets, also implies redistributions.[c] Politics matter for other macro policy areas, such as monetary policy and financial regulation. The recent financial crisis, for example, has reopened issues regarding the desirable conduct of monetary policy and the connection between

[a] OECD defines Social Expenditure as the provision by public (and private) institutions of benefits to, and financial contributions targeted at, households and individuals in order to provide support during circumstances which adversely affect their welfare, provided that the provision of the benefits and financial contributions constitutes neither a direct payment for a particular good or service nor an individual contract or transfer. Such benefits can be cash transfers, or can be the direct (in-kind) provision of goods and services.

[b] OECD (2014). The list of countries is: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Ireland, Italy, Japan, Netherlands, Norway, Portugal, Spain, Sweden, United Kingdom, United States.

[c] Alesina and Giuliano (2012) review the vast literature which has investigated the political and social determinants for the demand of redistribution.

monetary and fiscal policy. The ECB is at the center stage of the political discussion about institutional building in the Euro area. In the present chapter we focus exclusively on fiscal policy.[d]

The politics of fiscal policy could cover issues as diverse as the level of centralization vs decentralization, the structure of taxation, pension systems, the design of insurance programs like health care and unemployment subsidies, the optimal taxation of capital, international coordination of tax systems, just to name a few topics. In this chapter we focus on debt. Many countries have been struggling with large debt over GDP ratios even before the financial crisis: countries which faced the Great Recession starting with large debt risked (or experienced) debt crises, like Greece, Italy, and Portugal putting at risk even the survival of the Monetary Union. Japan has a public debt held by the private sector of at least 140% of GDP.[e] The political debate on how and at what speed to reduce the public debt after the Great Recession is at the center stage of the political debate.[f] When adding expected future liabilities of entitlements and pensions the public budget of most OECD countries, including the Unites States, look bleak. Debt problems in developing countries, especially in Latin America have been common. Any attempt to explain all of these phenomena leaving politics out is completely pointless.

In particular we ask two broad questions. First, is there a tendency in democracies to pursue suboptimal fiscal policies which lead to the accumulation of excessive debt, where "excessive" is in reference of what a benevolent social planner would do? In other words, how far are the observed pattern of debt accumulation and fluctuations in line with normative prescription of the literature on debt management like, in particular, Barro (1979), Lucas and Stokey (1983), and Aiyagari et al. (2002)? What explains substantial departure from optimality?[g] Second, are fiscal rules (and which ones) a possible solution to limit the extent of the problem of excessive deficits? The balanced budget rule is the most famous one, but may other have been proposed, especially in the Euro area. Two are the key issues in this debate. The trade off between the rigidity of a rule and the lack of flexibility which these rules create. More flexible rules may be superior but harder to enforce because they have too many escape clauses. Finally, assuming that a rule would work, would a country adopt it? Or would political distortions prevent it?[h]

We shall begin with a brief sketch of the prescriptions of the optimal debt management in order to identify the normative implication against which to confront actual

[d] Alesina and Stella (2010) address old and new issues regarding the politics of monetary policy.

[e] The gross figure is well above 200% but it includes debt held by various public institutions.

[f] Reinhart and Rogoff (2010) and Rogoff (1990) have emphasized the cost of debt burden for long run growth.

[g] For a review of an early literature on this point see Alesina and Perotti (1995). For more recent surveys see Persson and Tabellini (2000) and Drazen (2000).

[h] An issue which we do not consider in this chapter is the question of procyclicality of budget deficits and the political distortions which may lead to this problem. See Gavin and Perotti (1997) and Alesina et al. (2008).

policies. The goal of this chapter is not to review in detail the optimal debt literature. We will exclusively focus on models with distortionary taxation and we will not enter the discussion of the Ricardian equivalence. We will not discuss issues regarding governments' defaults on their liabilities, a topic which would deserve an entire chapter on its own. After having described which are the implications of the optimal taxation theory regarding debt management, we show that even a cursory look at the empirical evidence suggest substantial deviations from these prescriptions even amongst OECD countries. In fact, in terms of empirical evidence we will focus almost exclusively on OECD economies. Then, we discuss several different approaches which have tried to explain these deviations from optimality, by introducing political variables in debt management models. Finally, we return to a normative question. Given the presence of all of the potential political distortions examined above, which rules, institutions, procedures or a combination of them is more likely to bring actual fiscal policy closer to the social planner ideal policy? In addition, are these rule and procedures likely to be chosen? Have they worked in the past?.

This chapter is organized as follows. In Section 2, we briefly review the theories of optimal deficit management and the related empirical evidence. In Section 3 to 7, we address the first question, namely whether or not there is a deficit bias in modern economies, and what explains it. In Sections 7 to 10, we cover the question of fiscal rules and of which institutional arrangement would be more suitable to limit suboptimal conduct of fiscal policy. The last section discusses open issues for future research.

2. OPTIMAL DEBT POLICIES: A BRIEF REVIEW

2.1 Tax Smoothing

The theory of tax smoothing is due to Barro (1979) in a model where debt is not contingent and risk free, spending needs are exogenously given and known, taxes have convex costs. The public debt takes the form of one-period, single-coupon bond and the rate of return on public and private debt is constant over time. The government raises in each period tax revenues τ_t. Government spending is indicated with G_t and debt with b_t and the interest rate on debt with r. Thus the government budget constraint in each period is given by:

$$G_t + rb_{t-1} = \tau_t + (b_t - b_{t-1}) \tag{1}$$

The lifetime government budget constraint is given by:

$$\sum_{t=1}^{\infty}\left[\frac{G_t}{(1+r)^t}\right] + b_0 = \sum_{t=1}^{\infty}\left[\frac{\tau_t}{(1+r)^t}\right] \tag{2}$$

Raising taxes generates some extra costs which can be interpreted as collection costs, or more in general deadweight losses or excess burden of taxes and the timing in which taxes are collected. Let Z_t be this cost which depends on the taxes of that period τ_t and negatively on the pool of taxable income/resources Y_t. In particular, let Z_t be defined as:

$$Z_t = F(\tau_t, Y_t) = \tau_t f(\frac{\tau_t}{Y_t}) \tag{3}$$

with $f'(\cdot) > 0$ and $f''(\cdot) > 0$. The present discounted value of these costs is:

$$Z = \sum_{t=1}^{\infty} \tau_t \frac{f(\frac{\tau_t}{Y_t})}{(1+r)^t} \tag{4}$$

The social planner chooses τ_t in order to minimize (4) subject to the budget constraint (2). From the first order conditions, one can find that the tax–income ratio $\frac{\tau}{Y}$ is equal in all periods. Given that, the level of taxes in each period is determined from the values of income $(Y_1, Y_2, \ldots)$, government expenditure $(G_1, G_2, \ldots)$, interest rate r and the initial debt stock b_0. The properties of the solution are considered under different assumptions about the time paths of income Y and government expenditure G. With constant income and government expenditure (ie, $Y_t = Y_{t+1} = \ldots = Y$ and $G_t = G_{t+1} = \ldots = G$) since the tax–income ratio is constant, this implies that τ is also constant and the government budget is always balanced. With transitory income and government expenditure (eg, transitory expenditure during wartime or during recessions) deficits are larger the longer and the larger is the transitory shock. The debt–income ratio would be expected to be constant on average, but would rise in periods of abnormally high government spending or abnormally low aggregate income.

2.2 Keynesian Stabilization

This is not the place to discuss the potential benefits of discretionary countercyclical fiscal policy actions, namely increases in discretionary spending during recessions and reductions during booms. According to Keynesian theories, higher government spending or lower taxes during a recession may help economic recovery. The reason is that under high unemployment and low capacity utilization, higher government spending, and lower tax rates may increase aggregate demand. Note that Keynesian models would prescribe that deficits should be countercyclical (ie, increase in recessions), but should not lead to a secular increase in debt over GDP. The reason being that spending increases during recessions should be compensated by discretionary spending cuts during booms.

We only note that the "long and variable lags" argument raised by Milton Friedman regarding monetary stabilization policy applies even more to fiscal policy where the lags are even longer and less predictable than for monetary policy. Friedman's original argument was applied to monetary policy. He argued that the lags in between the uncovering of the need of, say, a stimulus, the discussion of it, the implementation and the realization of its effects were "long and variable." Therefore, by the time the expansionary policy came into action it was too late and it was counterproductive. This argument applies even more strongly to fiscal policy since the latter requires also an explicit political process, debate, and approval in parliaments. The recent Great Recession and the lower bound issue for monetary policy has made popular the view that in this scenario, aggressive discretionary fiscal policies are necessary since automatic stabilizers are not enough. We do not enter in the zero lower bound debate in the present chapter.

2.3 Contingent Debt

Lucas and Stokey (1983) build on Ramsey (1927) and show that Barro's intuition does not generally apply. The main difference with Barro (1979) is in the set of instruments available to the government to smooth the distortionary cost of taxation. While Barro (1979) focuses in only one instrument, namely noncontingent one-period bonds, Lucas and Stokey (1983) consider a model with complete markets, no capital, exogenous Markov government expenditures, state-contingent taxes, and government debt. In this, environment optimal tax rates and government debt are not random walks, and the serial correlations of optimal taxes are tied closely to those for government expenditures. Moreover, they find that taxes should be smooth, not by being random walks, but in having a smaller variance than a balanced budget would imply. Thus, to some extent, the idea of tax smoothing holds but not in the extreme version as in Barro (1979).[i]

2.4 Accumulation of Government Assets

Aiyagari et al. (2002) reconsider the optimal taxation problem in an incomplete markets setting. They begin with the same economy as in Lucas and Stokey (1983), but allow only risk-free government debt. Under some restrictions on preferences and the quantities of risk-free claims that the government can issue and own, it is possible to obtain back Barro's random walk characterization of optimal taxation. However, by dropping the restriction on government asset holdings (or modifying preferences) generates different results.

More specifically, under the special case of utility linear in consumption and concave in leisure, the authors show that as long as the government can use lump-sum transfers and spending shocks are bounded, then distortionary labor taxes converge to zero in the long run. The optimal solution prescribes reducing debt in good times, so that eventually the government has accumulated enough assets to finance the highest possible

[i] Interestingly, Klein et al. (2008) address the same issue raised in Lucas and Stokey (1983) but find different and strikingly results. In particular, they find that the time series of debt in the economy without commitment is extremely similar to that with commitment. Welfare is very similar as well. This result is surprising: under commitment, there is always an incentive for a once-and-for-all tax cut/debt hike, thus suggesting ever-increasing debt under lack of commitment. However, they show that the incentives that naturally arise in the dynamic game between successive governments actually help limit the time-consistency problem: they lead to very limited debt accumulation, and long-run debt levels can even be lower than under commitment. This incentive mechanism is a result of forward looking and strategic use of debt.

expenditure shock with the interest earned on its stock of assets. This is the so-called "war chest of the government." Instead, if one set a binding upper bound on the government asset level (Ad Hoc Asset Limit) the Ramsey solution for taxes and government debt will resemble the results stated in Barro (1979).[j]

2.5 Evidence on Optimal Policy

The very basic principles of optimal debt policies, namely the debt–income ratio would be expected to be constant on average, but would rise in periods of abnormally high government spending or abnormally low aggregate income, are generally not satisfied by the data.

Government debts do go up during wars and major recessions, but beyond that, deviations from optimal policy are widespread. Figs. 1 and 2 clearly show that government debts do go up in wars and recessions in the United Kingdom and United States.

The major role played by wars is evident in these graphs. However, even the United States shows anomalous features, like the accumulation of debt in the eighties, which is a

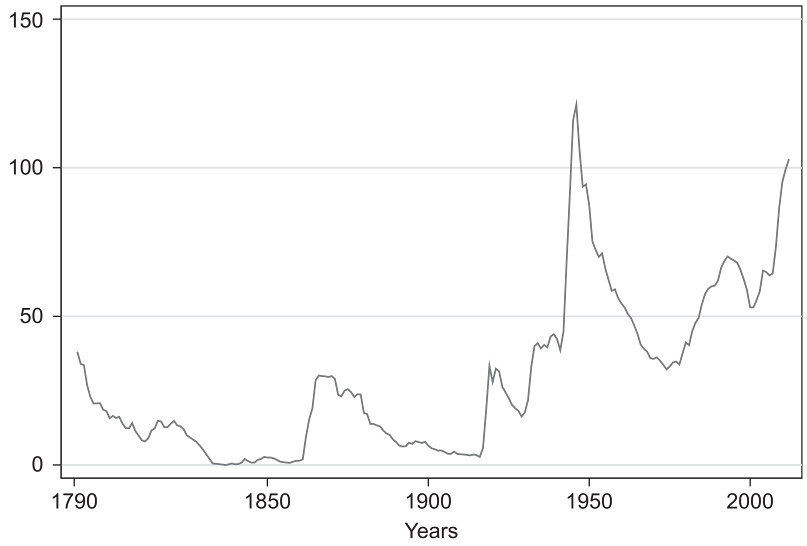


Fig. 1 Ratio of public debt to trend real GDP, the United States, 1790–2012. *Source: Abbas, S.A., Belhocine, N., Elganainy, A., Horton, M. 2010. A historical public debt database. Working Papers 245, International Monetary Fund.*

[j] By imposing a time invariant ad hoc limit on debt, the distribution of government debt will have a nontrivial distribution with randomness that does not disappear even in the limit. In particular, rather than converging surely to a unique distribution, it may continue to fluctuate randomly if randomness on government expenditures persists sufficiently.

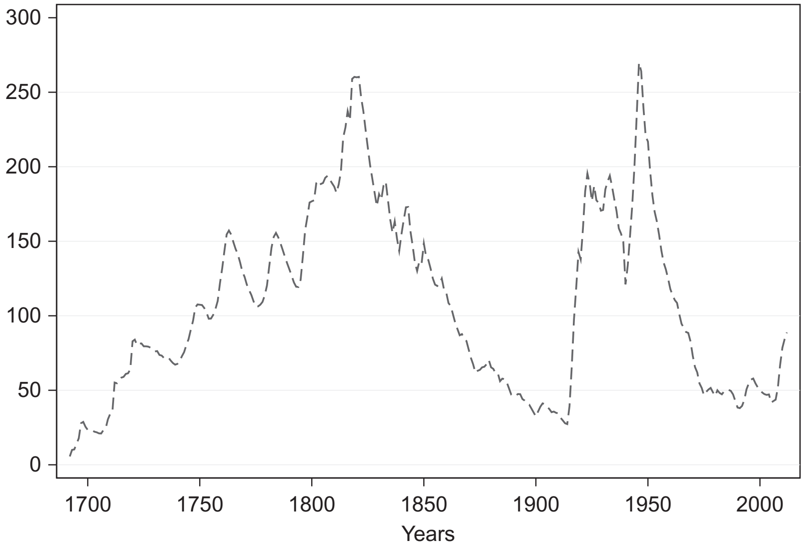


Fig. 2 Ratio of public debt to trend real GDP, United Kingdom, 1692–2012. *Source: Abbas, S.A., Belhocine, N., Elganainy, A., Horton, M. 2010. A historical public debt database. Working Papers 245, International Monetary Fund.*

period of peace. This episode (the so-called "Reagan deficits") in fact inspired a few papers reviewed later and, at the time, generated a major policy debate about the political forces which led to these deficits. Other OECD countries show remarkable deviation from optimality.

We show in Figs. 3 and 4 two graphs for a group of relatively high and low debt countries.

Several observations are in order. First, the decline in the debt ratios after the Second World War in both groups of countries stopped in the seventies. In both groups of countries it increased for several decades in peace time, obviously much more in the high debt group. For instance, in Italy and Greece the debt to GDP ratio skyrocketed in the eighties and nineties in a period of relatively rapid growth for these countries. Belgium and Ireland as well entered the nineties with debt level normally typical of postwar periods well above 100% of GDP. Second, several countries (ie, Ireland, Belgium, Denmark) had massive variations up and down of their debt ratios in peace time. Third, very few countries when they adopted the Euro satisfied the requirement of a less than 60% debt over GDP ratio. In addition, in the first decade of the Euro, up to the financial crisis, there was not much of an effort to converge to the prescribed target of 60%. Fourth, no country comes even close to a policy as prescribed by Aiyagari et al. (2002) which would imply the accumulation of assets to build a "war chest." Fifth, the Great Recession has led to very large accumulation of government debts and this is, at least in large part, consistent

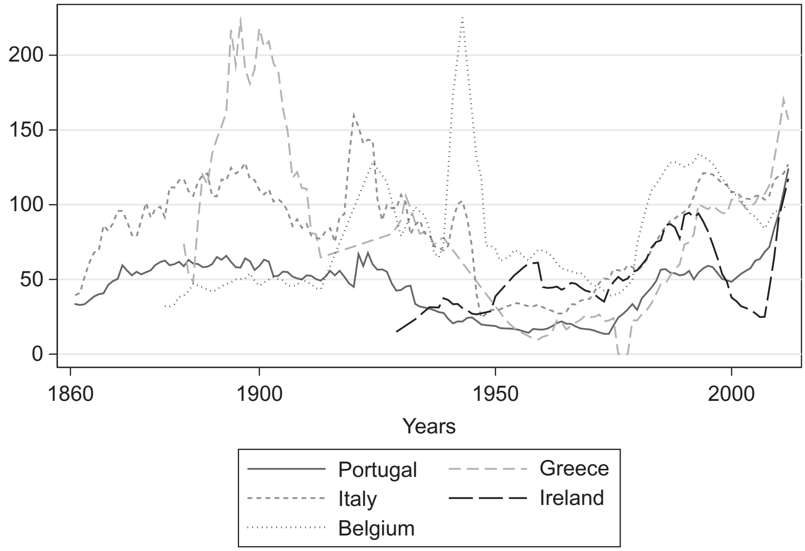


Fig. 3 High debt countries, ratio of public debt to trend real GDP. *Source: Abbas, S.A., Belhocine, N., Elganainy, A., Horton, M. 2010. A historical public debt database. Working Papers 245, International Monetary Fund.*

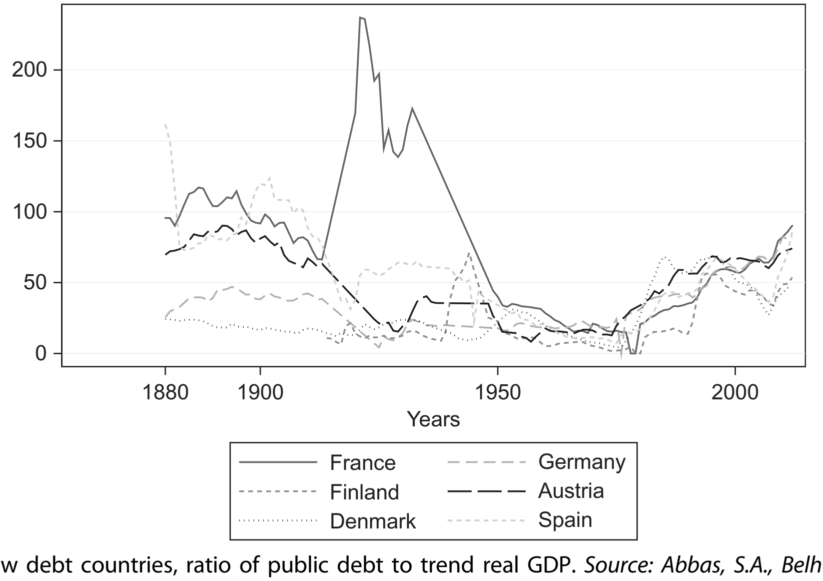


Fig. 4 Low debt countries, ratio of public debt to trend real GDP. *Source: Abbas, S.A., Belhocine, N., Elganainy, A., Horton, M. 2010. A historical public debt database. Working Papers 245, International Monetary Fund.*

with the tax smoothing hypothesis. However, countries which had already accumulated large debts for no obvious reasons before the crisis were constrained in how much they could accumulate more. Some additional accumulation created market panics; Greece had a partial default; Italy in 2011 was on the brink of a major crisis. Fifth, a few countries like Ireland and Spain entered the Great Recession with relatively low debt/GDP ratio but their fiscal position looked better than they really were due to extraordinarily and temporary tax revenues, namely the housing boom. When this became apparent these countries also faced debt panics. In fact, public debt problem in Europe almost degenerated to the point of a collapse of the Euro.

Table 1 shows that out of 20 OECD countries only 4 had a deficit for less than 50% of the time since 1960, and 11 countries had a deficit for more than 80% of the years. Italy and Portugal achieved a "perfect" 100%! These data do not distinguish between primary and total deficit, do not account for the cycle but nevertheless raise a significant flag about government profligacy. After the first oil shock of 1973–74, surpluses close to disappear. Easterly (1993) suggests that at that time (early seventies) many countries did not internalize a secular downturn of their growth process which would have required a reduction in the growth of government spending to keep the size of government constant. This lead to an accumulation of debt. Whether this misperception was an "honest mistake" or it was due to political distortions is a topic of discussion. In fact, it is pretty common for governments to justify large spending programs with very optimistic growth forecasts.

When considering the future liabilities of government, the picture regarding debt levels, appears substantially worse. The aging of the population (and the retirement of the baby boomers) will induce substantial strains over the social security budgets.

Table 1 Percent years of deficit over 1960–2011

	Australia	**Austria**	**Belgium**	**Canada**	**Germany**
Percent Last surplus	80 2008	82 1974	96 2006	76 2007	78 2008
	Denmark	**Spain**	**Finland**	**France**	**United Kingdom**
Percent Last surplus	48 2008	78 2007	20 2008	90 1974	84 2001
	Greece	**Ireland**	**Italy**	**Japan**	**Netherlands**
Percent Last surplus	80 1972	80 2007	100	68 1992	88 2008
	Norway	**New Zealand**	**Portugal**	**Sweden**	**United States**
Percent Last surplus	4 2011	46 2008	100	42 2008	92 2000

Source: Wyplosz (2014). Fiscal rules: theoretical issues and historical experiences. In: Alesina, A., Giavazzi, F. (Eds.), Fiscal Policy After the Financial Crisis, Volume Fiscal Rules: Theoretical Issues and Historical Experiences. University of Chicago Press and National Bureau of Economic Research, pages 495–529.

In different degrees, in various countries health expenses (also related to the aging of the population) are rising at phenomenal rates. The US Congressional budget office (CBO (2014)) predicts that with unchanged legislature, the debt over GDP ratio in the United States will never fall in the most optimistic scenarios in the next couple of decades. With "middle range" assumptions, the (net) debt over GDP ratio may be well above 100%. The forecasts of the social security administration have been called into question for being too optimistic and not transparent (Kashin et al., 2015). Similar considerations apply to Japan and European countries. There is a large difference between United States vs European countries, and also within European countries. Specifically, in the United States these entitlement programs are about 18.5% of the GDP, while in the European countries between 20% and 30% of GDP. Within the European countries, Norway is the leading country which spent about 30% of the GDP in Entitlement programs. Regarding the type of entitlement programs, pension expenditures account for more than half of the entitlements in Italy and Greece, while they are less than 20% in Ireland and Denmark.[k] In countries like Italy, we are reaching paradoxes in which youngsters do not find jobs because of high labor taxes and high labor cost for firms to collect tax revenues needed to pay pensions for the parents who then support the unemployed children.

The intergenerational accounting procedure for evaluating liabilities of the government offers an alternative measure to federal budget deficit to gauge intergenerational policy. It was developed by Auerbach et al. (1991) and it computes the net amount in present value that current and future generations are projected to pay to the government now and in the future. If one thinks that the government has an intertemporal budget constraint, then this constraint would require that the sum of generational accounts of all current and future generations plus existing government net wealth be sufficient to finance the present value of current and future government consumption. The generational accounts can be viewed simply as a tabulation of the net effect of future taxes paid and transfers received by various generations, assuming that current policy remains unchanged into the indefinite future. Auerbach et al. (1991) compute the "lifetime net tax rate," which measures the burden of taxes minus transfer payment on a generation over its lifetime. The Generational accounting criteria presumes that fiscal policies should be generational balanced. This would imply that the net tax rate for current and future generations should be the same. If the net tax rate for future generations exceeds the net tax rate for newborns, then according to this criteria, fiscal policy is not in generational balance. Haveman (1994) provides an excellent discussions of the pros and cons of generational accounting methods.

[k] Specifically, in 2011 pensions account for 51.9% of total Entitlement programs in Italy, 51.1% in Greece, 19.6% in Denmark, and 16.8% in Ireland. *Source*: OECD (2015).

3. DEFICITS AND ELECTIONS

3.1 Fiscal Illusion

The idea of "fiscal illusion" is due to the public choice school (see in particular Buchanan and Wagner, 1977). According to this argument voters do not understand the notion of intertemporal budget constraint for the government, therefore when (especially close to elections) voters see pending hikes or tax cuts (the public choice schools was especially concerned with the former) they reward the incumbent, and remain unaware of the consequences of such policies on public debt and the future costs of taxation needed to service it. The problem, according to the Public Choice school, is aggravated by the "Keynesian" policy stand. Politicians are eager to follow the Keynesian rule of increasing discretionary spending during recessions, but then they do not counterbalance it with cuts during booms. Thus, the result of keynesianism and fiscal illusion leads to persistent deficits and explosive debt levels.

In general, the view that the best way to please the voters is to spend more and tax less is so pervasive that it is assumed to be an obvious fact. As we show later, the evidence is much more nuanced than it would appear. In addition, given the extensive discussion of the deficits, the pros and cons of austerity policies in the United States and Europe, it is hard to believe that today's voters are unaware of the potential cost of deficits because of fiscal illusion, even though there may be disagreement on what policies to follow to respond to deficits. The fiscal illusion argument is overly simplistic although it does raise important warning bells on the conduct of fiscal policies in democracies.

3.2 Political Budget Cycles: Theory

The traditional fiscal illusion argument rely on some form of irrationality or ignorance on the part of the voters. However, political budget cycles can be derived also in models where voters are fully rational but imperfectly informed as in Rogoff (1990) and Rogoff and Sibert (1988). What leads to these cycles is a combination of delays in the acquisition of information on the part of the voters regarding the realization of certain policy variables and different degrees of "competence" of policymakers.[1]

In Rogoff and Sibert (1988) more competent governments can tax less to provide public goods, because they introduce less wastage in the fiscal process. However, the full combination of income taxes, spending, seigniorage, and government wastage (ie, negative competence) is learned with one period delay by the voters. A higher level of competence implies that the government can provide public goods with lower taxes (or seigniorage). Suppose that before an election voters see a tax cut. They cannot distinguish whether the cut is due to a high realization of competence (which is unobservable by them immediately) or transitory deficit which they do not fully observe.

[1] For a review of political business cycles in general see Alesina et al. (1993) and Drazen (2000).

After the election, a less competent government would have to increase seigniorage generating also an inflation cycle. With a finite time horizon the only equilibrium that exists is a separating equilibrium, ie, the one in which voters are able to infer exactly the incumbent's level of competency from the tax she selects in order to signal her competence. The competent policymaker cut taxes before election to a level that cannot be matched by the less competent one. A somewhat unpleasant feature of these models is that the more competent policymakers engages in budget cycles by cutting taxes before elections to signal their competence and distinguish themselves from the less competent ones who cannot afford such a large tax cut. Rogoff (1990) adds a distinction between two types of public goods, those that are clearly visible before an election, say fixing the holes in the street, and those less immediately visible, like increasing the quality of the training of teachers. In this model politicians have an interest in overspending in more visible but not necessarily the most productive public goods close to election time.

While, in principle, the implication of rationally based modern theories of political business cycles may be similar to the traditional one, they differ in two ways. First, the rationality of voters output a limit on the extent of these policies. Second, and this will be revealed by the empirical evidence, the more the voters are informed and understand the incentive of policymakers, the less they reward them for their behavior; thus for instance more freedom of the press in established democracies would be a constraint on this behavior.[m]

Drazen and Eslava (2010b) present models of political budget cycles in which the incumbent favors with certain spending projects specific and critical to constituencies and/or localities. By varying the composition of government spending the incumbent can target swing voters before elections. Incidentally, this imply that a political budget cycles may imply distribution of spending from one district to another, holding constant the total amount of government spending.[n]

3.3 Political Budget Cycles: Evidence

Are political budget cycles common? Persson and Tabellini (2000) argue that the answer depends upon the nature of the political institutions of the country. In particular, they argue that political budget cycles are less likely to occur in majoritarian systems rather than proportional representation systems. Brender and Drazen (2005), however, challenge these results. They find that the existence of political budget cycles do not depend on voting rules. Political budget cycles exist only in "new democracies," where fiscal

[m] For instance, Besley and Prat (2006) develop a model in which more press freedom reduces the space for policymakers to extract rents. For a review of the political economy of mass media refer to Prat and Stromberg (2013).

[n] Hassler et al. (2005) show an interesting result, namely that the introduction of political distortions would reduce, instead of exacerbate, oscillations in tax rates. This is contrast with the predictions of the literature on political business cycles.

manipulation may work because voters are inexperienced with electoral politics or may simply lack information, which may be one of the main factors generating the political budget cycle, as implied by the models reviewed earlier.

The role of information is tested by Brender (2003) for local elections in Israel. Peltzman (1992) and Drazen and Eslava (2010a) perform an analogous analysis in the United States and Colombia, respectively.[o] Gonzalez (2002) and Shi and Svensson (2006) test the importance of transparency, which ultimately means the probability that voters at no costs learn the incumbent's characteristics. They find that the higher the degree of transparency, the smaller the political budget cycle. Moreover, while the proportion of uninformed voters may be initially large, it is likely to decrease over time, thus decreasing the magnitude of the budget cycle. Akhmedov and Zhuravskaya (2003) find that measures of the freedom of the regional media and the transparency of the regional governments are important predictors of the magnitude of the cycle. Alt and Lassen (2006) find that, in the sample of OECD countries, higher fiscal transparency eliminates the electoral cycle.[p]

The other important aspect is whether or not governments which generated political budget cycles are more easily reelected. Brender and Drazen (2008) consider the effect of deficits on the probability of reelection and show that voters are (weakly) likely to punish rather than reward budget deficits over the leader's term in office. Their results are robust by considering different subsamples: (i) developed countries and less developed countries; (ii) new and old democracies; (iii) countries with presidential or parliamentary government systems; (iv) countries with proportional or majoritarian electoral systems; (v) countries with different levels of democracy.

A related literature directly tests the political consequences of large fiscal adjustments, ie, whether large reductions of budget deficit have important negative political consequences. Alesina et al. (1998) consider a sample of OECD countries and they find that fiscal austerity has a weakly positive, rather than negative, electoral effect. However, they focus on cabinet changes and opinion pools, rather than on election results. Alesina et al. (2012) fill this gap, by looking directly at the election results. They find no evidence of a negative effect on the election results due to a fiscal adjustment. Buti et al. (2010) find that the probability of reelection for the incumbent politicians are not affected by their efforts in implementing pro-market reforms. This literature, however, suffers from a potential sort of reverse causality problem, namely governments which are especially popular for whatever reasons, manage to get reelected despite their deficit reduction policies, not

[o] Schuknecht (2000) presents evidence of cycles in 35 developing countries and Buti and Van Den Noord (2004) some evidence on European Union countries.

[p] Alesina and Paradisi (2014) show evidence of political budget cycles in Italian cities. Foremny et al. (2015) provide evidence on political budget cycles using data on two German regions. Arvate et al. (2009) find evidence on localities in Brazil.

because of them. While the authors are aware of this issues and try to asses it, measuring the "popularity" of a government is not always straightforward.

The bottom line is that political budget cycles may explain relative small departures from optimal policy around election times, especially in new democracies. However, they cannot be the main explanation for large and long lasting accumulation of public debt, as we documented earlier. Also, the cross country empirical evidence seems to have been exhausted. Perhaps natural experiments at the local level might be interesting.

4. SOCIAL CONFLICT: WAR OF ATTRITION AND RIOTS

4.1 War of Attrition: Theory

War of attrition models do not explain "why" a deficit occurs, but they explain why deficit reduction policies are postponed. Alesina and Drazen (1991) focus on the case of a country that for whatever reason, due a permanent shock on revenues (or on expenditures), is on a "nonsustainable" path of government debt growth. The debt is held by foreigners and the interest rate is constant and exogenously given and there is no default. The longer the country waits to raise tax rates to stop the growth of debt, the more the interest burden accumulates and the more expensive the stabilization will be. The latter implies a reduction to zero of total deficits.

There are two equally sized groups of equal (exogenous) income which cannot agree on how to share the costs of the stabilization. The social planner would choose an equal division of costs for each group since the groups have the same income and size. In this case, stabilization would occur immediately since delays only create inefficient costs, namely higher interests on the accumulated foreign debt. The critical feature of the model is that without a social planner political polarization leads to an uneven distribution of the costs of the stabilization. In particular, one group has to pay more than 1/2 of the taxes needed for the stabilization and in every period after that. When both groups perceive the possibility of shifting this burden elsewhere, each group attempts to wait the other out. In order for this to happen there has to be some uncertainty about the costs of each group to wait the other out, namely how long a group can bear the costs of delaying the stabilization. These costs are modeled as the economic costs of living in the distorted prestabilization economy (for instance with inflation) or the political cost of "blocking" attempts of the opponent to impose an undesired stabilization plan. This war of attrition ends, and a stabilization is enacted, when a group concedes and allows its political opponents to be the winner. The loser then pays more than half of the costs of the stabilization, allowing the winner to pay less. The condition which determines the concession time is the one which equals the marginal cost living an extra moment in the unstable economy to the probability that in the next moment the opponent group will concede, multiplied by the differences the costs of being the winner rather than the loser. This is why uncertainty about the strength of the groups is critical. If one group knew from the beginning that its cost of living in an unstable economy were larger than those of the other group, it would know that it may end up losing the war of attrition and therefore it would concede immediately; this would be cheaper than postponing the inevitable loss. The passage of time reveals the type of the groups, namely which one is stronger. The more unequal are the divisions of the cost of the stabilization, which can be interpreted as a degree of polarization of a society, the longer the war of attrition and the higher the level of debt accumulated since the relative benefit of winning increase.

The war of attrition implies that individually (group level) rational strategies lead to a suboptimal accumulation of debt. The group which will end up being the loser is the one with the highest cost of prolonging the war of attrition. This is why uncertainty about these cost are critical. If it was common knowledge which was the weaker groups, the latter would capitulate immediately, since waiting adds to the costs and this group would lose anyway. Therefore anything that eliminates this uncertainty ends the war of attrition.

4.2 War of Attrition: Empirical Evidence

The model has several empirical implications. The first one is that the passage of time may lead a country to stabilize even if nothing observable happens, simply because one group has reached the condition of "conceding," namely has learned its relative strength to that of the opponent. Second, an electoral or legislative victory of one of the groups may signal its superior political strength and may lead the opponent to concede. Third, longer delays and higher debt should occur in polarized societies which cannot reach a "fair" and acceptable distribution of costs. In addition, delays are longer when many groups have a "veto power" to block policy decisions which they do not like. Fourth, a worsening of the economic crisis may lead to a resolution of the war of attrition. When the costs of delay increase for one of the groups the latter may concede sooner. Drazen and Grilli (1993) show that in their case a "crisis" can be beneficial, since it worsens the utility level of one of the groups in the short run, but it may be welfare improving for all in the long run since the war of attrition ends sooner. Fourth, for the opposite reason foreign aid can be counterproductive (Casella and Eichengreen, 1996). If foreign aid makes life easier before the stabilization, delays are longer and in the long run welfare is lower. The result, however, depends on how aid is disbursed; for instance foreign aid that implicitly "picks" a winner would end the war of attrition sooner. Finally, an external commitment, say an IMF conditionality agreement, may accelerate the resolution of the war of attrition making it more costly to "fight it." Several authors have suggested empirical observations consistent with the implications of the war of attrition model. Alesina and Drazen (1991) discuss a few historical examples of cases in which the same government first fails to stabilize because it encounters political opposition then it succeeds because the opposition is defeated. The idea that multiple veto players delay the elimination of deficits is consistent with the evidence by Grilli et al. (1991) and Kontopoulos and Perotti (1999).

The former argue that in the eighties, debt accumulated more in parliamentary democracies with multiparty systems. The latter argue that the number of spending ministers is associated with looser fiscal controls, an issue upon which we return later. Volkerink and De Haan (2001) and Elgie and McMenamin (2008) provide evidence on a sample of 22 advanced economies showing that more fragmented governments with smaller majorities in parliaments have larger deficits. Persson and Tabellini (2000) review and add to this line of research with additional evidence. These authors and Milesi-Ferretti et al. (2002) show also that coalition governments spend more on welfare, a point analyzed also by Alesina and Glaeser (2005) in a comparison of United States vs Europe. As we discussed earlier, Easterly (1993) noted that countries accumulated debt because they did not adjust their spending programs to the secular reduction of growth which started in the late seventies. These delays in adjusting to a permanent shock is consistent with the general message of the war of attrition. Various constituencies objected to reducing the growth of their favorite spending programs.

A second line of inquiry has focused on the idea that "crisis generates reforms," as in Drazen and Easterly (2001). Needless to say, the evidence suffers from problems of reverse causality: why would you need a reform if you did not have a problem to begin with?[q] Alesina et al. (2010) combine these institutional hypothesis with the crisis hypothesis, making a step closer toward testing the war of attrition model. In particular, they test whether certain institutions are more likely and rapid to resolve crisis, a result consistent with the model by Spolaore (2004). Alesina et al. (2010) define a country as being in a "crisis" if at time *t* the country is in the "worst" 25% of the countries in the (large) sample in terms of budget deficits.[r] They find support for the view that "stronger governments" stabilize more in time of crisis, ie, when a crisis comes, strong governments adjust more and exit more quickly from the state of "crisis." Strong governments are presidential systems and amongst parliamentary systems those in which the majority has a greater share advantage over the minority. They also find that stabilization (ie, exit from crisis) are more likely to occur at the beginning of a term of office of a new government. These results are consistent with the war of attrition model in the sense that in an unstable situation (ie, a crisis), a stabilization occurs sooner with fewer veto players or with a clear political winner. Results on the effect of IMF programs are inconclusive but, as discussed earlier, causality problems are especially serious in this case.

4.3 War of Attrition: Summing up

The war of attrition model has proven to be successful as an explanation of observed characteristics of run away debts and the timing of stabilization. One issue with this model is that it has been proven difficult to extend it. In particular, the division of costs of the

[q] Similar issues arise on the huge literature on foreign aid, which we can not even begin to survey here.

[r] In their paper these authors also consider inflation crisis, not only deficit crisis.

stabilization is taken as exogenous and not bargained amongst groups. Moving in that direction would lead to bargaining models where institutional details on how the game occurs are critical. Perhaps one may think about connecting this approach with the one discussed later on voting in legislatures. Also the extensions to n rather than 2 groups implies results which are not clear cut and the formation of coalitions amongst n groups are intractable (thus far) problems. Finally, in the model, a stabilization is a zero-one event. Partial or failed attempts are not explicitly modeled even though in reality are quite common.

4.4 Riots

Passarelli and Tabellini (2013) provide a model of political competition which has some connection to the war of attrition although with substantial differences and a "behavioral" bend. In their model several social groups have views about what is a "fair" allocation of resources. The sum of those views about what is fair for each group may be larger than the available resources. In addition groups are willing to engage in costly political actions (riots) when they feel that they have not obtain their fair allocation. When a group perceives that fairness (according to this group's view) has been violated, individuals are willing to engage in costly political actions, like riots, because of this emotional reaction to a perceived unfair behavior. The groups which are more homogeneous are also more likely to be more successful in organizing riots. This feeling of "anger" when perceived fairness has been violated solves the free rider problem of political actions. In a dynamic setting the threats of riots pose constraints to the government. In particular, even a benevolent government may be forced to accumulate excessive debt (above the optimal level) to reduce the threats of riots. Empirically, Woo (2003) shows that public debt accumulation is associated with the occurrence of riots. Ponticelli and Voth (2011) and Passarelli and Tabellini (2013) show how budget cuts are sometimes followed by riots.

It is interesting to compare this evidence on riots and the one reviewed earlier by Brender and Drazen (2008) and Alesina et al. (2012) which suggest that, at least in democracies fiscal adjustments are not associated with consistent electoral losses for the incumbent. Perhaps homogeneous and organized groups organize riots while the less organized median voter is much more prone to accept fiscal retrenchments when necessary. In other words, a government may face strikes and riots organized by specific homogeneous constituencies and those actions may block fiscal adjustment policies and increase public debt. However, the unorganized voters (which may be the majority) may not approve those policies.

It would be interesting to expand Passarelli and Tabellini (2013) framework to incorporate these features in which part of the electorate is organized and has this behavioral bend about fairness, and another part of the electoral is unorganized and does not have self serving feelings of fairness.

5. DEBT AS A STRATEGIC VARIABLE

Government debt is a state variable which "links" several successive governments. Different governments may have different preferences over fiscal policy, say the level and/or composition of public spending. If the current government is not sure of its reappointment, it may want to choose a level of deficit while in office (thus a level of debt) in order to influence the fiscal choices of future governments. In these models, deficits do not affect the probability of reelection since the voters are fully rational, fully informed, and forward looking, but deficits serve the purpose of insuring that future governments follow policies closer to the preference of the current government by constraining future governments' actions. The asymmetry of information that would lead to political business cycles, as we discussed earlier, are assumed away here, and the strategic manipulation of the debt by the current government or majority in office is fully in the interest of those who supported the current government. Another way to put this is the following. Given the inability of current government to control future public spending, it may prefer to take a $1 of tax revenue away from the future government by borrowing because it may not be in power and be able to decide how that $1 is spent in future, but it can decide how it can be spent today. Clearly, this logic applies only if there is political turnover and heterogeneity of preferences over fiscal policy amongst the different potential governments.

In Alesina and Tabellini (1990) two parties, with exogenously given preferences, stochastically alternate in office. They care about the level of income of the representative individual and care about two different public goods, say military spending vs domestic spending (more generally they place different weights on these two public goods). In the model there is a representative voter/citizen in terms of his/her choices of labor and leisure but with a distribution of preferences about the type of public goods that they prefer, so they would vote for different parties depending on the parties' choice of public goods. Private and public goods enter separately in the utility function. If a party is unsure of being reappointed, it will issue debt. By doing so it "forces" the following government (possibly of a different party) to spend less on the public good the current government does not care as much. In other words, the current government chooses to distort the path of income taxation in order to spend more on the public goods that it prefers leaving future governments with the task of reducing the debt since default is ruled out by assumption. The future government will do so, at least in part, by cutting spending on the public good the current government does not care much about.[s] The lower is the probability of reappointment of the current government the higher the level of

[s] When both parties care (with different weights) about the two public goods the result about excessive deficit require a weak condition on the third derivative of the utility function on the public goods.

debt chosen. Only a government sure of reappointment would issue no debt. The social planner would issue no debt since there is no reason to do so and would choose a stable combination of the two public goods in order to satisfy, say, utilitarian social preferences. Tabellini and Alesina (1990) provide analogous results in a model in which fiscal decisions are taken by the median voter. The current median voter is uncertain about the preferences of future median voters, because of shocks to the distribution of preferences. Today's median voter choose to issue debt for the political incentives of creating "facts" for future majorities. Alesina and Tabellini (1989) extend this type of model to a small open economy and show a connection between excessive public debts and private capital flights.

Persson and Svensson (1989) provide a related model which, however, does not imply a deficit bias but nonobvious implications about which government would lead a deficit and which would run a surplus. In their model, there are two parties, one of the left who likes a large amount of public goods even at the cost of high taxes, and a party of the right which, on the contrary dislikes public spending and taxation. The public debts links the two alternating parties in office. When the left is in office it chooses to leave a surplus by taxing more in order to generate an incentive for the right when in office to spend more on public goods. The right, when in office, will cut taxes creating a deficit in order to prevent easy spending when the left comes in to office.[t]

In a similar vein Aghion and Bolton (1990) consider the commitment effect of debt in two ways. First, by limiting future expenditure on public goods. Second, in forcing to raise higher tax revenues to repay the debt. Lizzeri (1999) uses similar insights, linking excessive debt accumulation and redistributive policies. In his model, two candidates, motivated purely by the desire of winning elections, can redistribute to some citizens and cannot make promises on future redistribution. In the first period, by running deficits they can target with "excessive" redistribution of transfer skewed in favor of a majority and against a minority.

6. THE COMMON POOL PROBLEM

In these types of models agents do not fully internalize the tax burden of spending decisions leading to "excessive" spending. The most widely studied "common pool problem" is the one of legislators (like the United States Congress) which would like to approve spending programs for their districts without fully internalizing the cost of taxation; in fact, the latter are spread on all (or many other) districts. As we discuss later, similar political distortions arise in different institutional settings.

[t] Pettersson-Lidbom (2001) presents supporting evidence for this model using Swedish data on localities.

6.1 Bargaining in Legislatures

Weingast et al. (1981) provide a model of excessive spending on pork barrel projects which was later extended to various voting rules and applied to study debt accumulation. These authors show how representatives with a geographically based constituency overestimate the benefits of public projects in their districts relative to their financing costs, which are distributed nationwide. The voters of district i receive benefits equal to B_i for a project, but have to pay $1/N$ of the total costs if taxes are equally distributed among districts. Thus, a geographically based representative does not internalize the effect of his proposals on the tax burden of the nation. The aggregate effect of rational representatives facing these incentives is an oversupply of geographically based public projects. Specifically, the size of the budget is larger with N legislators elected in N districts than with a single legislator elected nationwide, and the budget size is increasing in N, the number of districts.

Baron and Ferejohn (1989) substantially improve upon this model by considering voting on the distribution of taxes rather than assuming that every district pays $1/N$ of the cost of every project. They study decisions with majority rule with various alternative procedural rules. In their model there are n members (they can be interpreted as people, districts, or States) in the legislature. The task of the legislature is to choose the distribution of one unit of benefits among the n districts, with no side payments outside the legislature. A "recognition rule" defines who, at each session is going to be the agenda setter with the task of making a proposal. In each session, member i is chosen with probability p_i. Member i then puts forward a bargaining proposal of the form $x^i = (x_1^i, x_2^i, \ldots, x_n^i)$ such that $\sum_j^n x_j^i \leq 1$. If no proposal is approved, each member of the legislature gets zero benefits, the status quo. Members of the legislature have a common discount factor δ.

These authors distinguish between a "close amendment rule" and an "open amendment rule." In the first case, the proposal on the floor is voted upon against the status quo, with no amendments. If the proposal is approved, then the benefits are distributed and the legislature adjourns. If the proposition is rejected the benefits are not distributed and the legislature moves to the next turn. In this case the process starts over, but the benefits are discounted by the factor δ. With an "open amendment rule," after the member is randomly chosen to make the proposal, another member can be recognized at random and may either offer an amendment (ie, an alternative allocation) or move to vote. If the proposal is seconded, the legislature votes as previously. If the proposal is amended, a runoff election is held to determine which proposal will be on the floor. The process is repeated until a recognized member moves the previous question and a yes vote is reached.

In the case of closed amendment Rule, the subgame perfect equilibrium has the following characteristics: (i) the equilibrium distributions of benefits is majoritarian, ie, only a minimum majority gets something; (ii) the agenda setter can get a strictly greater allocation; and (iii) the legislature completes its task in the first session. In the case of open amendment rule, the agenda-setting power of the first proposer is diminished. Indeed, each member must consider the fact that her proposal may be pitted against an amendment. Thus, she has to take this into account when making the proposal. In particular, the proposing member must make a proposal acceptable for at least m out of $n - 1$ other members in the legislature. By choosing m, the original proposer determines the likelihood of acceptance. The higher is m, the higher the probability that the section rule will choose one of the m legislators and the proposal is accepted, but also the lower the benefits that the agenda setter can keep for himself.

6.2 Bargaining in Legislatures and Government Debt

In Velasco (1999, 2000) several interest groups benefit from a particular kind of government spending. Each group can influence the central fiscal authorities to set net transfers on the group's target item at some desired level. The equilibrium implies a debt level at the maximum feasible level. In fact each group demands transfers large enough to cause fiscal deficits and a sustained increase in government debt. Eventually, the government hits its credit ceiling and is locked forever in a position of paying sufficient taxes to service the associated maximal debt level. The intuition for this result is simple. Property rights are not defined over each group's share of overall revenue or assets. A portion of any government asset, which is not spent by one group, will be spent by the other group. Hence, there are incentives to raise net transfers above the collectively efficient rate. Groups do not fully internalize the costs of public spending, namely each of them uses the whole stock of resources instead of a fraction, as the basis for consumption of spending decisions. Krogstrup and Wyplosz (2010) provide a related common pool model of deficit bias in an open economy.

Battaglini and Coate (2008) adopt the Baron and Ferejohn (1989) framework described earlier and study how such bargaining leads to deviations from the optimal path of debt. They focus on the case in which a social planner would implement the solution by Aiyagari et al. (2002). Battaglini and Coate (2008) link the Baron and Ferejohn (1989) model of bargaining in a legislature with the insight of the literature on strategic debt which we have reviewed earlier, in particular the model by Tabellini and Alesina (1990). Current majorities in the legislature will bargain over spending with uncertainty about the nature of future majorities and the debt becomes, as earlier, a strategic tool to control future fiscal decisions.[u] While in Tabellini and Alesina (1990) the will of the majority is simply represented by the optimal policy of the median voter, Battaglini

[u] In a related work Barseghyan et al. (2013) consider, as a driver of fiscal policy persistent tax revenue shocks, which come from business cycle impacts on the private sector. Battaglini and Coate (2015) consider an economic model with unemployment and the distinction between private and public sector jobs. They explore the relationship between debt, unemployment, and the relative size of the public and private sector.

and Coate (2008) provide a much richer institutional setting to characterize decision making.

Battaglini and Coate (2008) model a continuum of infinitely lived citizens located in n identical districts. A single (nonstorable) consumption good z and a public good g are produced using labor. Citizens maximize their lifetime utility which depend on consumption, labor supply, and a parameter A_t, which is the realization at time t of a random variable, which represents the value of the public good for citizens at time t. If, for instance, the public good is defense spending, we value it a lot higher during a war. The legislature provides the public good g and it can finance targeted-district specific transfers s_i, ie, "pork barrel" spending. To finance its activities, the legislature can either set a proportional tax on labor τ or issue one-period risk free bonds x. The legislature faces three different constraints. A feasibility constraint, which imposes that the government revenues have to be high enough to cover expenditures. The "District Transfer Constraint," which imposes that the district-specific transfers must be nonnegative. This constraint excludes lump negative transfers (lump sum taxes) to finance government spending. Finally, the government has to satisfy the Borrowing Constraint, which implies setting an upper and lower bound on the amount of bonds that can be issued or bought back each period. The lower bound is set without loss of generality. Indeed, the government would never need more than the assets the lower bound implies so the constraint never binds. An upper bound is necessary to avoid the government to issue an amount of debt which is unable to pay back the next period. A lower bound is defined by the level according to which it is possible to finance the optimal level of public good just with the interests on the assets the government has accumulated.[v] The legislature, consisting of a representative from each of the n districts, make decisions with closed rules. The legislature meets at the beginning of each period knowing both b_t and A_t. One representative is randomly selected to make the government policy proposal, which consists of the tax rate on labor r_t, the level of public good g_t, the level of bonds x_t, and the district-specific transfers $(s_1,\ldots,s_n)$. The proposal requires consensus of a minimum winning coalition of $q < n$ legislators to be accepted and implemented. If the proposal is rejected another legislator is randomly chosen to make a new proposal. If, after τ rounds, all the proposals are rejected, then the government implements the "Default Policy," which has to satisfy the feasibility constraint and has to treat all the districts equally, ie, $s_1 = \ldots = s_n$.

In this model a social planner would choose the optimal debt path as in Aiyagari et al. (2002). More specifically, the social planner takes as given (b,A) and chooses a policy $\{r, g, x, s_1, \ldots, s_n\}$ which maximizes the utility of citizens in all district. Given (b,A) there

[v] The optimal level of public good is the one which satisfies the Samuelson Rule, ie, the level at which the sum of marginal benefits is equal to the sum of marginal costs.

are two possible cases, namely with or without transfers to the districts. In the first case, with positive pork barrel transfers, the optimal tax rate on labor is set to zero and the optimal level of public good is set to $g_S(A)$, ie, the level that satisfies the Samuelson's Rule. The reason is straightforward. Suppose that the tax rate is positive. Then, the Social planner finds strictly dominant to reduce the pork barrel transfers and to reduce the (distortionary) tax. If the Social Planner does not make any pork barrel transfer, it must be the case that the tax rate is positive, the level of public good provided is less than $g_S(A)$ and the level of public debt exceeds the one with transfers. Thus, pork barrel transfers depend upon the realization of the value for the public good, A. In particular, for high enough values of A, the optimal policy has no transfers: g is high and no room is left for pork barrel. Instead, if the government has resources left to provide pork barrel transfers, then the level of debt must be the lowest possible, ie, the lower bound $\underline{x}$. (Remember that the lower bound implies accumulation of assets). Intuitively, if the planner is willing to give revenues back to citizens through district transfers $(s_1,\ldots,s_n)$, then it must expect not to be imposing taxes in the next period; otherwise, he would be better off reducing transfers and acquiring more bonds. This suggests that the steady state debt level must be such that future taxes are equal to zero, implying it to be equal to $\underline{x}$.

Consider now bargaining in the legislature. The agenda setter has to find $q-1$ supporters for his proposal to pass. The equilibrium policies are driven by the realization of the value of the public good, A, and the value of the public debt left from the previous period. For high enough values of A and/or b, the marginal value of the public good is so high that the proposer does not find it optimal to make positive pork barrel transfers. Thus, the equilibrium policy consists of the outcome as the proposer maximize the utility of all representatives. In other words, we are back to the Social Planner solution with no transfer. For low levels of b and/or A, there may be resources left that can be transferred to the q districts. This implies there exists a cutoff value A^* which divides the space into two different regimes. For $A > A^*$ the economy is in the "responsible policy making" regime (RPM). In this case, the optimal level of the tax rate, the public good and the debt to issue are defined by the Social Planner's optimal conditions with no pork barrel. For $A < A^*$ the economy is in the "business-as-usual" regime (BAU). In this case the proposer defines $(r^*, g^*(A), x^*)$ by maximizing the utility for the q districts included in the "Minimum Winning Coalition." This equilibrium includes also transfers $(s_1,\ldots,s_q)$ high enough to induce the member of the coalition to accept the proposal.

The same optimal conditions can be defined in terms of the public debt. In particular, the equilibrium debt distribution converges to a unique invariant distribution whose support is a subset $[x^*, -x]$. When the debt level is x^*, then the optimal conditions for the tax rate and the public good are those defined by the BAU, with the proposer who makes pork barrel transfers to the q districts. If instead the debt level exceeds x^*, then the economy is in the RPM regime where the tax rate is higher than the one defined in BAU, the provision of public good is lower, and no districts receive transfers.

In the long run, the economy oscillates between BAU and RPM regimes, depending on the realization of the value of the public good A. For instance, pork barrel would disappear during a war when A is large.[w]

In summary, the political distortions which make the social planner solution differs from the political equilibrium arises for two specific reasons. The first one, which can be related to the "Common Pool problem" discussed in the previous section. The minimum winning coalition does not fully internalize the costs of raising taxes or reducing the public good but it fully enjoys the benefit of receiving the pork barrel transfers. The other distortion comes from the uncertainty suffered by the legislators. They do not know ex-ante whether they are going to be included in the minimum winning coalition next period. Thus, they do not fully internalize costs and benefits across periods. In particular, they compare $\$\frac{1}{q}$ benefit today by belonging to the coalition, vs $\$\frac{1}{n}$ expected costs tomorrow. This intuition is similar to the strategic model of debt of Tabellini and Alesina (1990) reviewed earlier. In conclusion, this section makes two important contributions. First, it merges the results found in Tabellini and Alesina (1990) by using Baron and Ferejohn (1989) type of model. Second, it shows that taxation smoothing "a la Barro" is still an important factor in a political economy model, but distortion smoothing through debt is inefficient, and therefore not only this results in excessive accumulation of debt, but also in excessive volatility of the policies in the steady state. From an empirical standpoint, Baqir (2002) shows results consistent with the common pool problem using data from US cities. He shows that larger city council, where the common pool problems may be larger, are associated with more public spending, holding other determinants of the latter constant.

There is also a potential connection with the war of attrition model discussed earlier. In these bargaining models the passage of time is not considered. With a closed rule agreement is immediate but even with an open rule to the extent that proposals and amendments can be made instantaneously time does not matter. In reality, bargaining in legislatures takes time, and the passage of time is critical in the war of attrition models to allow the game to be resolved. At the same time the passage of time leads to the accumulation of debt. Allowing for a realistic consideration of time in these bargaining model could be an interesting avenue for theoretical and empirical research.

6.3 The Common Pool Problems in Other Institutional Settings

The general idea of the common pool problem with strategic debt is relevant for other institutional settings beyond the US Congress.

[w] Battaglini (2014) illustrates an extension of that model, which includes two-party competition in a legislature modeled as earlier.

In particular, in many democracies the budget is crafted by a government (possibly formed by more than one party), it is presented in the legislature and approved, if the parties of the government have a majority, with or without amendments. In this case, we may have a common pool problem with the spending ministers in the government even before the budget reaches the legislature. Each spending minister would generally like to obtain more spending for its own ministry, often pushed by the bureaucracy of the latter. A winning coalition of spending ministers may lead to the approval of a budget which, like in the BAU regime of Battaglini and Coate lead to a sort of "pork barrel" transfers to a minimum winning coalition of spending ministers. These pork barrel spending may be geographically or functionally defined and the bargaining may get especially complicated when different spending ministers belong to different competing parties. In this institutional setting normally the Treasury Minister has the task of preventing spending ministers to overspend but he or she may be overruled by a minimum winning coalition of spending ministers. In fact, as we shall discuss later, different institutional settings attribute different levels of prerogatives to spending ministers vs the Treasury, making the problem arising in the BAU regime more or less serious. In addition, even in parliamentary democracies, legislatures have the ability of proposing and voting upon amendments on the budget presented by the government.[x]

Often budget deficits at the national levels originate at subnational levels of governments. Some famous examples are both from Latin America (ie, Argentina) and European countries (Italy and Spain, for instance). This is related to suboptimal allocation of spending and taxing prerogatives amongst various level of governments. Suppose that spending is decided by local governments and revenues are collected by the national government and allocated to localities on the basis of their spending decisions. Obviously, in this case localities do not internalize the full cost of taxation of their spending decisions since taxes are levied nationally. Most countries have arrangements which attempt to put a limit on these incentives, such as having some local taxes required to finance some type of spending, or having budget rules on local governments (as we will discuss later). In many cases, however, these arrangements are imperfect and a common pool problem remains. The relationship between local governments and the Central Government may also imply a case of soft budget constraint (see Kornai et al., 2003). Localities expect Central Government to bail them out and overspend. Pettersson-Lidbom (2010) provides a test using Swedish data.

[x] Tornell and Lane (1999) develop a model of a sort of common pool problem applicable more directly to developing countries with poorly developed institutions and large informal sectors. They develop a dynamic model of the economic growth process that contains two common characteristics of those developing countries that have grown slowly in the last decades, namely (i) the absence of strong legal and political institutions and (ii) the presence of multiple powerful groups in society. The focus is on the fiscal process as it is the mechanism through which powerful groups interact with the society (which is characterized by weak legal and political institutions) and where they can enforce discretionary fiscal redistribution—a kind of pork barrel transfer—as a way to appropriate national resources for themselves.

This discussion is of course related to the fundamental issues of fiscal federalism.[y] The trade off is well known. On the one hand, one wants to allow to federal countries some freedom of choice on their localities. On the other hand, such freedom should not imply a deficit bias at the national level.

7. INTERGENERATIONAL REDISTRIBUTION

Current generations, by means of government debt, redistribute from future generations to themselves. The argument is very appealing. However, it needs to take into account the fact that private bequest are positive, thus one needs to account for negative "public" bequest (government debt) and private positive bequests. In this respect Cukierman and Meltzer (1986) consider the standard framework with overlapping generation model, lump sum taxes and intergenerational transfers from parent to child, and no uncertainty. Individuals differ in their abilities, (and therefore in wage earnings) and in their nonhuman wealth. Some of them desire to leave positive bequests, and others would prefer to borrow resources from future generations. Individuals who would choose to leave negative bequests are "bequest-constrained" individuals. These individuals favor any fiscal policy that increases their lifetime income at the expense of future generations. Individuals who are not bequest constrained are indifferent to an intergenerational reallocation of taxes. In fact they can adjust up or down their private bequest when public bequests (government debt or assets) move up or down. By majority rule, if the decisive voter is bequest constrained, he will choose lower current taxes financed by additional debt, which cannot be defaulted. If instead the decisive voter is not bequest constrained, he is indifferent to a reallocation of taxes and social security over time that maintains present value. Thus, in this model by majority rule we will easily have an accumulation of debt. The likelihood to have deficits increases with an extension of the franchise to low wealth individuals who are likely to be bequest constrained. This is a simple but very powerful idea which strikes us as just right.

Tabellini (1991) explores a different argument, that is the redistribution consequences of debt repudiation in an overlapping generation framework implying both intra and intergenerational redistributions. The main idea is that issuing debt creates a constituency in support of repaying it. Thus, issuing debt makes a coalition of voters favorable to repaying it in order to avoid intragenerational redistributive consequences of the debt repudiation. In particular, parents have a first-mover advantage since they can vote on how much debt they want to be issued (ie, how much resources they want to extract from future, yet-unborn generation), without the future generation to have a word. Issuing government debt results in intergenerational redistribution to be tight to intragenerational consequences of choosing how much debt to repay. In particular, debt reputation harms the old, but it harms the wealthy more than the poor.

[y] See Oates (2011) for the classic work.

Young voters (specifically the children of the wealthiest debt holder parents) want to avoid intragenerational redistribution (ie, repudiation would result in redistributing wealth from rich to poor families) and for this reason they are willing to accept to repay some debt (ie, transferring resources to the parents), an action that would have been opposed by them ex-ante.[z] Therefore, there is a coalition that includes both old and young voters (the wealthiest) who vote in favor of debt repayment. The most interesting and valuable aspect of this chapter is the joint consideration of intra and intergenerational redistribution, a topic which is surprisingly understudied both theoretically and empirically. In many countries pension systems redistribute both across and within generations, to the extent that poor citizen get proportionally more than rich ones from pensions. This is an excellent topic for further theoretical and empirical research.

Song et al. (2012) develop a dynamic general equilibrium model of small open economies where voters in each period choose domestic public goods and the financing via taxes and debt. Within each country, old agents support high spending on public goods, high labor taxes and large debt. Instead, the young dislike debt, since it crowds out public good provision when they will be old. Specifically, the model consists of a set of small open economies populated by overlapping generations of two-period-lived agents who work in the first period and live off savings in the second period. In each country j there two types of goods: a private good c and a domestic public good g provided by each economy's government. There are two types of agents, the young and the old, each with a different preference towards the public good, which are represented, respectively, by the parameters θ_j and $\lambda\theta_j$. λ represents a preference weight that old put on the public good. Intuitively, this parameter can take value 0—individuals do not value the public good—or positive values—not necessarily bounded to 1. There are cross-country differences in θ which may reflect cultural diversity or differences in the efficiency and quality of public good provision, related to the technology and organization of the public sector. Capital is perfectly mobile across countries and it fully depreciates after one period. The private good is produced by using both capital and labor as inputs in the production function. The domestic fiscal policy is determined through repeated elections and government debt is traded on worldwide markets. Given an inherited debt b_j, the elected government chooses the labor tax rate τ_j, public expenditure g_j and debt accumulation $b_j^{\prime}$, subject to a standard dynamic government budget constraint. A probabilistic voting model delivers an equilibrium in which fiscal policy maximizes a weighted sum of young and old voters' utility. The weights assigned to each group represent the relative political influence of

[z] This is because, ex-ante issuing debt has only intergenerational, but not intragenerational effect. Given that agents would prefer not to redistribute resources, they would vote against this policy ex-ante. However, ex-post the policy has also intragenerational effect and the young generation would prefer to transfer resources to their parents rather than to the fraction of poor people in the same cohort.

each group. The model yields a trade-off between the marginal costs of taxation, due to the reduction in private consumption *c* suffered by the young, and the marginal benefit of public good provision. Such a trade-off reveals a conflict of interest between young and old voters. The old want higher taxes and current spending on public goods. Thus, the more power held by the old, the greater the reduction in private consumption. The preference for public good provision affects this trade-off: a higher θ or a higher λ reduces private consumption *c*. Moreover, there exists a sort of "disciplining effect" exercised by the young voters. In particular, they anticipate that increasing debt will prompt a fiscal adjustment reducing their future public good consumption. A key result is that the model provides a politico-economic theory of the determination of the debt level. In particular, in spite of the complete lack of intergenerational altruism (assumed through finite lives) debt converges to a finite level, strictly below the natural borrowing constraint. This results from the combination of forward-looking repeated voting and distortionary taxation. Higher debt can be financed by increasing taxes or cutting public good provision. As debt grows larger, the convexity of tax distorsions (a Laffer curve effect) implies that most of the adjustment will be in the form of less future public goods. The concern for avoiding a future situation of private affluence and public poverty makes young voters oppose debt increases. Given the prediction of a determined debt level, the model yields mean-reverting debt dynamics. Suppose that the economy is hit by a one-time fiscal shock (eg, a surprise war) requiring an exogenous spending. The government reacts by increasing taxes and decreasing nonwar expenditure in wartime. After the war, debt, taxes, and expenditure revert slowly to the original steady state. These predictions accord well with the empirical evidence of Bohn (1998), who finds the US debt-to-output ratio to be highly persistent, but mean reverting and Müller et al. (2016) which provide similar evidence for the period 1950–2010 for a panel of OECD countries.

Müller et al. (2016) extend their model by assuming that there are two types of voters, left wing (*l*-type) and right wing (*r*-type), who differ in their trade-off between private consumption and public good consumption: *l*-type voters like government expenditure and public good provision more than do *r*-type voters. Voters choose sequentially a fiscal policy which includes labor taxation, government expenditure on public goods, and debt policy, subject to the government's dynamic budget constraint. The novelty of this model compared to Song et al. (2012) is that, here there are political shocks which can be interpreted as shocks over time to the preference for public goods. In particular, during a left-wing wave the government increases taxation and public expenditure while reducing debt. Instead, during a right-wing wave the opposite occurs. In fact the driver of fiscal discipline of the young is based on their preferences for public good when old—that is how much the young expect that they will appreciate public good provision as they become old. During left-wing governments, the demand for fiscal discipline is stronger because the young left-wing voters—who are more concerned for future public good provision than right-wing voters of the same age—detain more political influence.

This is because *r*-type voters have less appeal to public good and more for private consumption. Thus, when the right-wing party is in power is less concerned to the provision of public good in the future and instead it would push up current debt today in order to use the resources as subsidies for private consumption. Left-wing voters are instead concerned with future public good provision, and would oppose such fiscal policy. The key predictions of the model are that, on the one hand, right-leaning governments are more prone to issue debt in normal times, while on the other hand left-leaning government engage in more proactive countercyclical fiscal policy—including issuing more debt during recessions. In other words, during normal times left-leaning governments do more public savings but use the debt to smooth income shortfalls associated with recessions.[aa] This result is reminiscent of the model by Persson and Svensson (1989) reviewed earlier, in a nonoverlapping generation framework.[ab]

It should be mentioned that all the models discussed earlier imply voting. Mulligan and Sala-i Martin (1999) argue that indeed spending on pensions is high in nondemocracies as well as democracies, namely variables like the aging of population and the relative size of young and old matter in both regimes. In fact the relative "strength" (ie, political influence) of the constituencies of young and old may be relevant in both democracies and nondemocracies even though the nature of the way in which this relative strength manifests itself is of course different. These differences in the intergenerational games in perfect and imperfect democracies and in dictatorships is an excellent topic for additional research.[ac]

8. RENT SEEKING

Acemoglu et al. (2008, 2010, 2011) study the dynamic taxation in a standard neoclassical model under the assumption that taxes and public good provision are decided by a self-interested politician who cannot commit to policies. Citizens can discipline politicians by means of election as in Barro (1973) and Ferejohn (1986) in a dynamic game. The self-interested politician creates distortions, namely he wants to extract rents from being

[aa] They show that these theoretical predictions are consistent with US postwar data on debt, and also with a panel of OECD countries.

[ab] However, the key difference between the two papers is that in Persson and Svensson (1989) a conservative government expecting to be replaced in the future strategically issues more debt. In contrast, the results in Müller et al. (2016) are unrelated to persistence or reelection probabilities. The robust prediction of their theory is that a left-leaning government issues less debt, irrespective of the probability of being replaced.

[ac] Azzimonti et al. (2014) make the case that the secular increase in debt to output ratios can be due to the liberalization of financial markets that took place in the mid eighties. While the political-economy comes from probabilistic voting, the paper provides an alternative theory of debt (to that of tax smoothing) and an explanation of why we could observe inefficiently higher debt to GDP ratios in the recent years. Specifically, they propose a multicountry political economy model with incomplete markets and endogenous government borrowing and show that governments choose higher levels of public debt when financial markets become internationally integrated and inequality increases.

in office. This adds an additional constraint in the economy, the political economy constraint. This constraint implies that politicians in power compare the lifetime utility from extracting rents in each period vs the one-time shot deviation of extracting all the resources available in the economy in one period and being voted out of office. Distortions are generated by the fact that citizens have to provide incentives to politicians to stay in office. These distortions may or may not disappear in the long run. In particular, if politicians are as patient or more patient than citizens, they value more staying in office and thus they set a tax rate equal to zero. If politicians are less patient than citizens, it may be optimal to set positive taxation. The idea is that, starting from a situation with no distortions as before, an increase in taxation has a second-order effect on the welfare of the citizens holding politician rents constant, but reduces the resources available in the economy and, thus, the rents that should be provided to politicians by a first-order amount.[ad] Thus, it is less costly to reduce the potential output in the economy, than to provide a higher rents to politicians to stay in office. These types of models therefore focus on the role of taxation as a tool to govern the interaction between citizens and self-interested politicians. There is no role for government deficit.

Yared (2010) develops a rent seeking model with implications on the accumulation of public debt using a Lucas and Stokey (1983) model. Yared considers a closed economy with no capital, with shocks to the productivity of public spending, and with complete markets. The self-interested politician has a utility function which is increasing in rents (namely tax revenues not used for productive public goods, ie, spending with no social value). A politician cannot commit to policies once in office and citizens cannot commit to keeping the incumbent in power in the future. Thus, in an infinitely repeated game, reputation sustains equilibrium policies. The focus is on "Efficient Sustainable Equilibria" in which a politician who pursues rent seeking extractive policies is voted out of office, and a politician who purses the policies expected by citizens is rewarded with future office.[ae] Therefore, the incumbent politician follows equilibrium policies as long as rents are sufficiently high, since this raises the value of cooperation, and as long as government debt is sufficiently high, since this limits what he can acquire through maximally extractive policies prior to removal from office. There is no default. Citizens reward a

[ad] Specifically, the marginal cost of additional savings for the citizens is higher in equilibrium than in the undistorted allocation, because a greater level of the resources in the economy increases the politician's temptation to deviate and thus necessitates greater rents to the politician to satisfy the political sustainability constraint.

[ae] The equilibrium refinement used is the sustainable equilibrium as in Chari and Kehoe (1993). In particular, individual households are anonymous and nonstrategic in their private market behavior (ie, buying government debt), while the representative citizen is strategic in the replacement decision. The politician in office is strategic in his decision regarding the policies, which have to satisfy the government dynamic budget constraint. The set of sustainable equilibrium are those in which citizens solve their optimal decision with respect to consumption, labor supply and bonds' decision given their individual budget constraints. Within the set of sustainable equilibrium, the focus is on the efficient ones, ie, the ones that maximize citizens' utility.

well-behaved incumbent by not replacing him as long as equilibrium taxes are sufficiently low and productive public spending is sufficiently high. Note that given the fact that citizens are all identical, there is no conflict in the political decision. Efficient sustainable policies thus solve the standard program of the benevolent government subject to incentive compatibility constraints for the politician and the representative citizen.

Consider now the rent seeking politicians. Given the lack of commitment, there are two set of incentives that have to be satisfied, the politician's and the citizens' incentives. The incumbent politician knows that citizens will remove him from office at the beginning of the following period if he misbehaves. In particular, a politician who is removed after period t receives period t rents and a punishment which is a function of χ^p, ie, an exogenous parameter representing the strength of political institutions, namely the institutional constraints on politicians. The optimal policy for the citizens has to satisfy the constraint that the politician does not want to extract maximal rents and be removed from office. Maximal rents implies getting as much revenues as possible today, take out as much debt as possible today, delivering zero public goods, and repaying current debt. Therefore, the incumbent politician is less likely to deviate from the equilibrium policies if: (i) he is receiving a high level of equilibrium rents today and in the future because in this case the value of cooperation is high; (ii) if government debt is high because there is little space for him to expropriate resources through increasing his rents. Satisfaction of this incentive compatibility constraint implies a lower bound on taxes and an upper bound on public spending which both bind whenever the incentive compatibility constraint binds. This is because there has to be a limit on the size of resources owed to the government in each period. Indeed, if the size of these resources is too large, there is a high incentive for the politician to deviate and appropriate them as rents. This implies that resources going into a given period cannot be too large, and government activity must be financed mostly with current and future taxes, instead of past taxes.

The second set of incentives to take into account are those for the citizens. In this model, citizens may have an incentive to replace an incumbent politician even if he is well behaving. In this sense, citizens cannot commit to a plan where they keep an incumbent in power no matter what. Therefore, the incumbent politician has to set fiscal policies such that they define a sufficiently low level of taxation and/or a sufficiently high level of public expenditure in order to have some chances to stay in office the subsequent period. In this framework, replacing an incumbent politician provides a benefit for the citizens which is a function of the exogenous parameter χ^c. Here, χ^c represents the lack of popularity of the incumbent.[af] These conditions provide upper bounds on revenues and lower bounds on public spending.

[af] Another interpretation may be the gains for the citizens from having a new incumbent, reflected in the policies that are promoted during the electoral campaign. The author interprets it as a general "social benefit of political turnover."

Summing up, satisfying the incentives of politicians requires sufficiently high revenues and sufficiently low levels of public spending. In contrast, satisfying the incentives of citizens requires sufficiently low level of taxes and sufficiently high level of public spending. The best policy is therefore found to be the one that maximizes citizens' lifetime utility subject to the two set of incentive compatible constraints. This political distortion leads to several departures from the social planner policies. In particular, taxes are not constant but *volatile*. This is because the constant revenue policy characterizing the benevolent government is associated with too much rent seeking by politicians. Second, the increase in debt reduces the potential rents that the politician can appropriate and thus make it easier for citizens to provide the incentives to politicians. This approach is elegant, although contingent debt as in Lucas and Stokey (1983) is not issued by real world governments.

9. BUDGET RULES

Given that for so many reasons there are incentives for the government to run excessive deficits, is it feasible to devise rules and institutions that limit or eliminate those problems? By rules we mean numerical targets like a balanced budget rules, or a limit on the level of deficit, perhaps adjusted by the cycles, or excluding certain items such as public investment.[ag]

9.1 Balanced Budget Rule for National Governments

The pros and cons of national balanced budget rules, namely rules which imply zero or negative deficits (surpluses) are clear. A balanced budget rule does not allow to smooth out spending shocks (ie, to run deficits when the need for spending are especially large) or fluctuations of tax revenues over the cycle for given tax rates. However, to the extent that political distortions are so large that governments may be far from the optimal policy, then a balanced budget rule might be a second best solution to massive political distortions.

The political debate on balanced budget rules is extensive, since the pros and cons are, in principle, straightforward but there are strong prior views about which costs or benefits are bigger and those views are not likely to be changed by the available, relatively scant, evidence.[ah] An additional set of issues relates to the enforceability of balanced budget rules, namely whether governments restricted by these rules would engage in "creative accounting" to circumvent them or simply *de facto* ignore them.

[ag] For a review see Fatás and Mihov (2003b).

[ah] See Sabato (2008) for a presentation of the policy debate. Fatás and Mihov (2003a) present evidence on a cross section of countries consistent with the view that the presence of budget rules limits the volatility of fiscal policy.

Azzimonti et al. (2015) present a quantitative evaluation of the net benefits of a balanced budget rule (BBR) for the US economy using the political economy model developed by Battaglini and Coate (2008).[ai] As reviewed earlier, political economy frictions lead to inefficiently high levels of government debt in the long run. A constitutional requirement that imposes that tax revenues must be sufficient to cover spending and the interest on debt (eg, permitting surpluses but not deficits) may improve welfare by restraining policymakers from excessive debt creation. The authors show that the BBR leads to a gradual reduction of debt in equilibrium. Intuitively, the reduction in flexibility to smooth taxes imposed by the rule increases the expected costs of taxation. Therefore, savings become more valuable as a buffer against adverse shocks. By lowering the stock of debt in good times, legislators reduce interest payments, which decreases pressure on the budget in bad times. In the long run, this results in lower taxes and higher spending in equilibrium than in the unconstrained case, "pushing" the model on the direction of optimal fiscal policy. The impact of a BBR on welfare is theoretically ambiguous: in the short run, citizens experience a loss in utility since the government has to cut spending and raise taxes to reduce debt above what might be optimal. In the long run, citizens benefit from lower debt levels but, due to the inability to borrow in bad times, suffer from higher volatility. Because the net effect depends on parameters, the authors calibrate the model to the US economy using data between 1940 and 2013, and show that it can fit the path of US fiscal policy reasonably well. One immediately wonders whether including the Second World War years in this exercise is appropriate given that during a major war probably the balanced budget rule could be easily abandoned. By including a major war period they, in a sense, may set the stage for a framework with high costs for balanced budget rules. The authors find that the short run costs are too large to compensate for the steady state benefits of a lower stock of debt. However, quite apart from the parametrization (which, as always, could be debatable) the model makes an interesting point: the balanced budget rule could be costly in the short run and beneficial in the long run. This result leads to interesting and immediate consequences on the political economy implications on voting upon a balanced budget rule in say, an overlapping generations model.

Halac and Yared (2015) discuss the optimal design of centralized supranational fiscal rules like those for Euro area countries, and how they compare to decentralized (national) fiscal rules in an environment in which there is a trade-off between allowing flexibility while also reducing a government's deficit bias. They consider a two-period model in which a continuum of identical governments choose deficit-financed public spending. At the beginning of the first period, each government suffers an idiosyncratic shock to the social value of spending in that period. Governments are benevolent ex-ante, prior to the realization of the shock, but present-biased ex-post, when it is time to choose

[ai] See also Stockman (2001) for calibrations of balanced budget rules in RBC models.

spending—which can be interpreted as the results of the potential political turnover (ie, the political business cycle). The results of the chapter compare optimal rules—which maximize the social welfare of all countries—when it is set by a central authority or an individual government. The results can be summarized as follows: when governments are not too impatient when choosing public spending, then the optimal centralized fiscal rule is tighter than the decentralized one, and hence interest rates are lower under centralization. The idea is that, in choosing decentralized rules, an individual country does not internalize the fact that by allowing itself more flexibility, a country pushes the global interest rate up, and thus redistributing resources away from governments that borrow more towards governments that borrow less. Instead, committing ex-ante to tighter rules is good as this pushes down the global interest rate and therefore allows countries with higher marginal value of spending to borrow more cheaply. If governments' present bias is large, the optimal centralized fiscal rule is slacker than the decentralized one, and hence interest rates are higher under centralization. The idea is that governments choosing rules independently do not internalize the fact that by reducing their own discretion—ie, by choosing very tight borrowing limits—they lower interest rates, thus increasing governments' desire to borrow more and worsening fiscal discipline for all. Instead, committing ex-ante to more flexibility is socially beneficial: the cost of increasing discretion for over borrowing countries is mitigated by the rising interest rate, which induces everyone to borrow less. The interest rate has a disciplining effect in the sense that it reduces the incentives for over borrowing countries to borrow more.

Aguiar et al. (2015) investigate the conditions under which the imposition of debt ceilings is welfare improving. Specifically, they study the interaction between fiscal and monetary policy in a monetary union with the potential for rollover crises in sovereign debt markets. Each member-country chooses how much to consume and borrow by issuing nominal bonds. A common monetary authority chooses inflation for the union, taking as given the fiscal policy of its member countries. Both types of policies are implemented without commitment. The lack of commitment on fiscal policy is especially critical because it may lead to the possibility of default. They show the existence of a "fiscal externality" in this type of environment. This externality leads countries to over borrow and thus, higher inflation and lower welfare. This gives credit to the imposition of debt ceiling in a monetary union which overcome the problem of lack of commitment on fiscal policy. Aguiar et al. (2015) go further and investigate the impact of the composition of debt in a monetary union, that is the fraction of high-debt vs low-debt members, on the occurrence of self-fulfilling debt crisis. Specifically, they show that a high-debt country may be less vulnerable to crises and have a higher welfare when it belongs to a union with an intermediate mix of high- and low-debt members, than one where all other members are low debt.

One could also think of balanced budget rule with escape clauses. An obvious one, mentioned earlier already would be a major world war. This (fortunately) rare event may be used as a relatively easy contingency to verify, but if the contingencies become too frequent then not only the stringency of the rule but even its enforceability is called into question. For instance, how does one define a "major" war? Clearly the Second World War was major, but would the Iraq war be a major one? Also one might think of cyclically adjusted balanced budget rules to overcome some of the rigidity of the latter, but then debates about how to measure the cyclical adjustment might lead to strategic manipulation of the rule itself. With specific reference to the United States, Primo (2007) discusses the pitfalls of balanced budget rules with complicated escape clauses.

An additional argument against formal budget rules is that financial markets might impose increasing borrowing costs on government which move far away from the optimal policy and accumulate large debts. Increasing borrowing costs would lead to more discipline even without rules. The recent experience of the Euro area and its fiscal crisis, casts doubts on this argument. Until 2008 the interest rate spread on, say German government bonds and even Greek ones was virtually nil. In fact, as a result of this low spreads several countries accumulated large debts in the first decade of the monetary union even when these countries were growing at respectable rates, including Greece whose economy was booming and debt skyrocketing. The reason of this is that probably investors did not believe the no bail out case of European treaties and assumed (largely correctly) that in case of a debt crisis they would be protected. In fact, probably because market discipline was not considered sufficient the funding fathers of the monetary union introduced contingent budget rules, like the stability and growth pact. These rules have been changed repeatedly and generally implied a maximum level of deficit (3% of GDP) with various escape clauses in case of major recessions. The discussion about the optimality of such rules in the Euro area is immense and we do not review it here (see the excellent discussion in Wyplosz, 2014).[aj] However, we want to make three points here. One is that the enforceability of these rules has been questionable. Even as early as 2002 Germany itself broke the rule and then many countries followed this example. The complexity and contingency of these rules did not help. The second is that probably now some European countries are feeling the bite of such rules, binding during a prolonged recession. The third is that especially at the time of the introduction of the Euro much creative accounting was widely used to satisfy "on paper" the 3% rule. These procedures introduced confusion and decreased trust amongst members of the Euro area.[ak]

How can balanced budget rules for a sovereign national government can be enforced? One possibility is to have the law in the constitution so that it would take a Constitutional revision to change it. An alternative would be to require a qualified majority. Such rules need to be stable, namely they should not imply that the rule itself can be changed, as in Barbera and Jackson (2004). For some discussion of this issue, see Primo (2007) which

[aj] For some empirical evidence on the stability and growth pact see von Hagen and Wolff (2006).

[ak] Von Hagen (2006) compare the effectiveness of budget rules in the EU vs Japan.

elaborates over the Baron and Ferejohn (1989) approach with specific reference to the US institutional setting. This is an excellent topic for future research not only within the specific American institutions.

9.2 Balanced Budget Rules for Local Governments

The pros and cons of balanced budget rules discussed earlier for national government apply also to subnational ones. However, there are reasons to believe that balanced budget rules for local governments may be more attractive than for national governments. First, as we discussed earlier, local governments add an additional political distortion: a common pool problem given by the fact that their local spending is at least in part financed by national transfers and therefore local governments do not fully internalize the taxation costs of their spending decisions. Second, some (or most) of the countercyclical fiscal stabilizers may be national not local. In fact, balanced budget rules for local governments should be accompanied by nationally based automatic stabilizers, to avoid procyclical fiscal policy, unless, as were discussed earlier, a balanced budget rule is chosen also for the national government. Third, enforcement of local balanced budget rule may be easier since it may be done by the national governments. Fourth, a balanced budged rule for local governments would avoid accumulation of unsustainable debts with the related uncertainty, disruption and costs associated with bail outs of excessively indebted localities. In summary, balanced budget rules for local government may be a tool of an optimal allocation of fiscal responsibilities between national and local governments.[al]

Indeed, work by Alt and Lowry (1994), Poterba (1995), Bayoumi and Eichengreen (1994), Bohn and Inman (1996), and Alesina and Bayoumi (1996) show that more strength budget rules in the United States, namely tight fiscal controls which impose restrictions on government deficit, have been more effective at creating incentives to states more quickly responding to spending or revenue shocks.[am]

9.3 Other Types of Budget Rules

The policy discussion over balanced budget rules has also dealt with other types of budget restrictions. One is the so-called "golden rule," namely a rule which allows budget deficits only to finance public investments but not current expenditures. Bassetto and Sargent (2006) discuss the optimality of such rules. In principle, this may be a "good" rule especially for developing countries in need of investment in infrastructures. The problem, however, is that this rule may lead to creative accounting, namely simply reporting as spending in infrastructures what is really current spending. For developed countries one may wonder whether the political incentives to spend in physical

[al] See Inman (1997) and Poterba (1996) for a review of this literature.

[am] Canova and Pappa (2006), however, present results suggesting that in some cases US states managed to circumvent the rules.

infrastructures which would be induced by this rule is really necessary. In Western Europe, in particular, the emphasis on physical infrastructures seem overplayed already, relative to other fiscal problems in this continent, and a budget rule of this type may add to this misperception and lead to overinvestment in physical infrastructures.

Another possible budget rule would impose limits on spending. The issue here is that while we have a theory of optimal deficit management, reasonable people can disagree on the optimal size of government spending because of different views about the role of the state and the size of welfare policies, for instance. Thus, while pork barrel inefficient programs (like bridges to nowhere) might be constrained by spending limits, the latter may interfere with programs desired by the majority.

10. BUDGET INSTITUTIONS

10.1 Theory

The definition and approval of a budget in an advanced democracy is often a complex process, possibly kept strategically complex to achieve behind the scene deals or to be able to introduce them in some corner of the budget provisions in a sufficiently obscure manner to escape detection of the voters. One can identify three phases in the budget process: (1) the formulation of a budget proposal within the executive; (2) the presentation and approval of the budget in the legislature; and (3) the implementation of the budget by the bureaucracy. Two issues are crucial: the voting procedures leading to the formulation and approval of the budget, and the degree of transparency of the budget. We begin with the former.

We focus upon a key trade-off between two types of institutions. One type, which we label "hierarchical," limits the democratic accountability of the budget process with a high degree of delegation. The second type, we label "collegial," has the opposite features. Hierarchical institutions are those that, for instance, attribute strong prerogatives to the prime minister (or the Finance or Treasury minister) to overrule spending ministers within intergovernmental negotiations on the formulation of the budget. Hierarchical institutions also limit in a variety of ways the capacity of the legislature to amend the budget proposed by the government. Collegial institutions emphasize the democratic rule in every stage, like the prerogatives of spending ministers within the government, the prerogatives of the legislature vis-a-vis the government, and the rights of the minority opposition in the legislature. There is a trade-off between these two types of institutions: hierarchical institutions are more likely to enforce fiscal restraints, avoid large and persistent deficits, and implement fiscal adjustments more promptly. On the other hand, they are less respectful of the rights of the minority, and more likely to generate budgets heavily tilted in favor of the interests of the majority. Collegial institutions have the opposite features.

Let's begin with the definition of the budget within the government where we have a division of responsibilities between spending ministers and the Treasury minister. The latter has the role of aggregating the spending proposals of other ministers and produce a budget document. Spending ministers prefer a larger fraction of the budget devoted to their department: more money means more favors to constituencies. Thus, more hierarchical institutions are those which attribute stronger prerogatives to the Treasury. In the legislature, as we discussed earlier, different amendment rules may aggravate or reduce the common pool problem. Much of this research is based, directly or indirectly, upon a view of the budget as the result of conflicting interests of representatives with geographically based constituencies. The literature on procedures has addressed three related questions: what procedural rules mitigate or aggravate the problem of oversupply of pork barrel projects? What procedural rules make the choice of projects, given a certain total budget, more or less efficient? How do different procedural rules influence the final allocation of net benefits among districts? Two issues are particularly interesting for our purposes: (a) the sequence of voting on the budget, and (b) the type of admissible amendments on the proposed budget. Intuitively, one may argue that by voting first on the maximum size of the budget (and eventually of the deficit) one would limit the excessive multiplication of budget proposal. Ferejohn and Krehbiel (1987) study theoretically the determination of the size of the budget under the two alternative voting procedures. They assume that the budget can be allocated to two projects and different legislators have different preferences for the relative benefits of these two projects. It is not always the case that the size of the budget is smaller when the legislatures vote first on the size and then on the composition, relative to the case in which the overall budget size is determined as a residual. While the size of the budget is in general not independent on the order of votes, the relative size of the budget with different orders of votes depends on the distribution of legislatures' preferences for budget composition.[an]

In parliamentary democracies, the agenda setter in the budget process is the government. Thus, closed rules attribute more power to the government and less to the floor of the legislature. The result is that closed rules are more hierarchical as we discussed earlier. They give more influence to the government and lead to an immediate approval of the budget than the government poses. Open rules require more time for voting and with those rules the government gets a lower surplus relative to the nongovernmental minority. With a closed rule you achieve quick approval of a proposal, at the cost of implementing "unfair" budgets. Budgets are unfair in the sense that they are tilted in favor of those who make the first proposal, and always distribute benefits to the smallest possible majority. Hierarchical procedures are obviously preferable when the key problem is the control of the size of the budget and the implied deficit.

[an] The same issue has been revisited by Hallerberg and Von Hagen (1999).

Finally, the issue of transparency. The budgets of modern economies are very complex, sometimes unnecessarily so. This complexity, partly unavoidable, partly artificially created, helps in various practices to "hide" the real balance (current and future) of costs and benefits for the taxpayers. Politicians have incentives to hide taxes, overemphasize the benefits of spending, and hide government liabilities (the equivalent of future taxes). At least two theoretical arguments support this claim. The first is the theory of "fiscal illusion" reviewed earlier. By taking advantage of voters' irrational confusion, politicians can engage in strategic fiscal policy choices for reelection. The second argument does not rely on voters' irrationality and confusion. Several papers, although in different contexts (eg, Cukierman and Meltzer, 1986 and Alesina and Cukierman, 1990), highlight the benefit for policymakers of a certain amount of ambiguity even when they face a rational electorate. The idea is that, by creating confusion and, in particular, by making it less clear how policies translate into outcomes, policymakers can retain a strategic advantage vs rational, but not fully informed, voters. This advantage would disappear with "transparent" procedures; therefore, policymakers would often choose to adopt ambiguous procedures. Milesi-Ferretti (2004) shows that politicians who want to run excessive deficits would choose nontransparent procedures, and the latter would help them to achieve their (distorted) goals. As we discussed earlier, Rogoff and Sibert (1988) and Rogoff (1990) make a similar point in the context of political business cycle models. They show that if voters cannot easily observe the composition of the budget (on the spending or on the financing side), then policymakers can follow loose fiscal policies before elections and increase their chances of reappointment. Gavazza and Lizzeri (2009) develop a model in which the lack of voters' information about the complexity of the budget lead to transfers to voters even when taxation is distortionary and voters are homogeneous. Transfers are financed with debt and the latter is higher the less transparent the system is, that is the less likely it is that voters can fully observe fiscal variables.[ao]

How, in reality, do policymakers obfuscate the budget? and what to do about it? In practice, a variety of tricks can serve the purpose of strategically influencing the beliefs and information of taxpayers/voters. For instance: (1) Overestimate the expected growth of the economy, so as to overestimate tax revenues, and underestimate the level of interest rates, so as to underestimate outlays. At the end of the fiscal year, the "unexpected" deficit can be attributed to unforeseen macroeconomic developments, for which the government can claim no responsibility; (2) Project overly optimistic forecasts of the effect on the budget of various policies, so that, for instance, a small new tax is forecast to have major revenue effects, thus postponing to the following budget the problem of a real adjustment; (3) Keep various items off budget; (4) Use budget projections strategically. For example, in all the discussions about future budgets, a key element is the

[ao] The same authors (Gavazza and Lizzeri, 2011) investigate how lack of transparency may lead to the choice of inefficient fiscal tolls for redistribution.

"baseline." By inflating the baseline, politicians can claim to be fiscally conservative without having to create real costs for the constituencies. In this way, they create an illusion: they appear conservative in the eyes of the taxpayers, worried about the size of the budget, but they do not really hurt key constituencies with spending cuts. Clearly, this illusion cannot last forever, since adjustment, rigorous only relative to inflated baseline, in the end will not stop the growth of the debt. However, this procedure creates confusion and, at the very least, delays the electorate's realistic perception of the actual state of public finance; (5) Strategic use of multiyear budgeting. By announcing a, say, 3-year adjustment plan in which all the hard policies occur in years 2 and 3, politicians can look responsible and can buy time; then, they can revise the next 3-year budget policies to further postpone the hard choices.[ap]

We can think of three possibilities for increasing transparency. The first and most commonly followed is a "legalistic" approach. That is, more and more rules and regulations are imposed on how the budget should be prepared, organized, and executed. This approach is unlikely to be successful: complicated rules and regulations provide fertile ground for nontransparent budget procedures. A second alternative is to create legislative bodies in charge of evaluating the transparency, accuracy, and projections of the government budget. This approach is superior to the legalistic one, but it relies heavily on the political independence of this public body. This independence may be problematic, particularly in a parliamentary system where the government parties control a majority in the legislature. A third alternative, the most radical but the most effective, is to delegate to a respected private institution the task of verifying the accuracy and transparency of the budget process. In addition, the government budget should be based on an average of the economic forecasts of and projections derived by international organizations or private institutions.

10.2 Empirical Evidence

The empirical evidence on the relationship between rules and deficit is, generally speaking, supportive of the idea that hierarchical institutions are associated with lower deficits. Hallerberg et al. (2009), in a book which also summarizes and consolidate previous works by the same authors, classify budget institutions for the EU countries in terms of delegation of prerogatives to the Treasury minister versus a contracting approach within ministers, the presence of targets, voting rules in parliament, relationship between central and local governments. They argue that institutions matter and delegations and targets (ie, hierarchical institutions) are effective at containing deficits and debts. Alesina et al. (1998) and Stein et al. (1999) consider Latin America countries and construct an index of their budget institutions based upon surveys of local officials. In doing so they can distinguish up to a point between *de iure* and *de facto* procedures. These authors correlated

[ap] See Alesina et al. (2015) for a detailed study of multiyear fiscal adjustment plans.

positively an index of hierarchical of budget institutions and of transparency to lower levels of debt. Fabrizio and Mody (2006) obtain similar results for Center and Eastern European countries. Dabla-Norris et al. (2010) on a vast sample of developing countries. These results should be taken very cautiously since they are based upon a handful of countries and often the classification of procedures is open to question. For instance, *de iure* and *de facto* procedures may differ substantially. Also comparing along those lines very different countries might be challenging, for instance think of a comparison of United States vs parliamentary democracies budget institutions. Debrun et al. (2008) compile a detailed data set for European Union countries for the period 1990–2005. They consider numerical fiscal rules on any fiscal aggregate, their legal status (normal law, constitutional law, supranational rules, accepted norms) and consider both national and subnational governments. Based upon this vast data set they build an index of stringency of the rules and they find that it strongly correlates with fiscal performance. More stringent rules reduce a deficit bias and improves upon the countercyclical stance of fiscal policy in EU countries. Miano (2015) has shown that national rules have the effect of reducing deficits. A recent work at IMF (Budina et al., 2012) provide extensive data on budget institutions for many countries and examine how the recent financial and fiscal crisis in many countries have led to reforms in budget institutions. These data have not been used yet for extensive empirical analysis.

11. QUESTIONS FOR FUTURE RESEARCH

In this final section, we elaborate on some issues which in our view are left open in this literature.

11.1 Endogenous Institutions

The literature which we have reviewed thus far uses certain political institutions (eg, type of government, electoral rules, presidential vs parliamentary systems) as exogenous or at least predetermined in explaining economic variables. In the present chapter we focus on debt and deficits but a vast literature also considers other related variables like the size of government and the level of redistribution for instance.

The assumption of exogeneity of predetermined institutions as “cause” of deficits can, however, be called into question. The same historical, sociological, cultural variables which may have led to the choice of certain institutions may also be correlated with fiscal policies.[aq] For instance, suppose that a parliamentary proportional system (generating a multiparty system with many veto players) was adopted because it was the only way to guarantee representation to very polarized and divided societies (across income, ideological, religious or ethnic lines). Those same characteristics of society might lead to

[aq] See Alesina and Giuliano (2015) for a discussion of the relationship between culture and institutions.

certain choices of fiscal policies (spending, deficits, debt). Thus, proportional representation and deficits would correlate but causality is called into question. Along those lines, Alesina and Glaeser (2005) review the literature showing that in many European countries proportional representation was introduced after the First or Second World War under pressure from Socialist and Communist parties. The presence of the latter clearly is not exogenous to fiscal policy decisions. Aghion et al. (2004) discuss how certain types of voting rules would be chosen optimally or not (ie, with or without a veil of ignorance) in divided societies.[ar] Empirically, they show how ethnic fractionalization is correlated with various institutional variables. Galor and Klemp (2015) present results along similar lines using different measures of diversity. On the other hand a vast literature on ethnic fractionalization (see the survey by Alesina and La Ferrara, 2005) show how the latter variable is correlated with several economic variables which may be directly or indirectly correlated with deficits and debt. Thus, diversity of populations may "cause" both institutions and fiscal outcomes. The correlation between the latter two does not imply causality, strictly speaking. Persson and Tabellini (2000) in their work on institutional determinants of fiscal policies are aware of this limitation and make some progress in addressing causality, but this remains an open question. The literature on fiscal policy which appeals to institutional variables as causal explanation for deviations from optimality (especially when thinking of long run horizons) needs to make the extra step. At this point, the correlations seem clear, identification of causality is not.

These arguments apply even more strongly when focusing specifically to budget institutions. The latter may work very differently in different countries depending upon their interaction with other features of the country itself. Hallerberg et al. (2009) argue that delegations to the Treasury minister does not work well in countries with sharp differences in the preferences of different parties for fiscal policy, a result which is consistent also with the model of political delegation by Trebbi et al. (2008). With a deep political conflict delegation to one decision maker is hard, undesirable by the minority and possibly counterproductive. Budget institutions are clearly endogenous. Why do countries choose different budget institutions and therefore to what extent the latter can be used as right hand side variables in a regression with debt and deficits on the left hand side? Countries with lower polarization and more homogeneous governments may be more likely to choose more hierarchical fiscal institutions, since delegation is easier, as argued earlier. But then it may be that the lower political conflict leads to more restrained fiscal policies; in this case, institutions are just an "intermediate" variable. In other words, paradoxically countries which needs stringent budget rules the least, since they have a lower tendency to run deficits, may be those which adopt more stringent budget rules. As noted by Hallerberg et al. (2009), some institutional reforms in the direction of making them more hierarchical have followed deep crisis, like the case of Sweden in the nineties. But again,

[ar] See also Trebbi et al. (2008) for an application to US cities.

causality is an issue: perhaps changes in attitudes due to the crisis might have led to a political equilibrium with more fiscal restraints regardless of the institutions. It is virtually impossible to establish causality from budget institutions to fiscal outcomes, although the correlations are interesting. Debrun et al. (2008) are fully aware of this problem and attempt to instrument their index of stringency of rules with some institutional variables but the exclusionary restriction is highly questionable. Miano (2015) shows how the adopting of various budget institutions are endogenous to a host of sociopolitical variables and are affected by the timing of elections. Overall, the argument that budget institutions "cause" fiscal discipline is virtually impossible to make empirically given the endogeneity of these institutions. Countries with a culture of fiscal profligacy will not adopt them (or will not enforce them) while countries with a culture of rigor will adopt and enforce them. The evidence presented earlier is consistent with a weaker argument namely that countries which, for whatever reason, cultural or otherwise, prefer budget discipline will be helped in their goal by choosing certain institutions rather than others. We think that we need more research on this point: to what extent institutions "cause" fiscal policies? Perhaps more natural experiment-based research may help address this question.

A second line of argument relates to the time consistency of institutional rules. To what extent institutional choices would be time consistent and not reversed as a result of various shocks? Halac and Yared (2014) address precisely this issue in a model where a government has an incentive to overspend. The government chooses a fiscal rule to trade off its desire to commit to not overspend against its desire to have the flexibility to react to shocks. These authors show that in the case of persistent shocks the ex-ante optimal rule is not sequentially optimal. The optimal rule in fact is time dependent with large fiscal shocks leading to an erosion of future fiscal discipline. It would be very useful to investigate the choice of budget rules under a Rawlsian veil of ignorance at the constitutional table or in a situation in which the veil of ignorance has holes, as in related work by Trebbi et al. (2008) on voting rules.

11.2 Culture

A rapidly growing literature has recently explored how various cultural traits affect economic decisions in a variety of dimensions including, savings, investment, trade, labor markets and the private or public provisions of safety networks and, more generally, growth and development.[as] Cultural traits like trust, relationship between family members (including intergenerational generosity), individualism, respect of the rules of laws, propensity to save and in which form, have been widely studied and their relevance for economic behavior is well established. Many of these attitudes are relevant for a society's acceptance of government deficits, including their intergenerational redistributive effects. Also the acceptability of policies geared towards reducing excessive deficits

[as] Guiso et al. (2006) and Alesina and Giuliano (2015) provide surveys of this literature.

may be different in different cultural settings. For instance, Guiso et al. (2015) investigate how cultural differences among Euro area countries may have led to the aggravation of conflict over debt policies and delayed resolutions of the latter. Cultural values certainly affect decisions about tax evasion,[at] another variables which clearly determines the accumulation of debt. While a relatively vast literature studies tax evasion, we are not aware of much work linking it to the accumulation of debt.[au]

The connection between institutions and culture is important (Alesina and Giuliano, 2015; Bisin and Verdier, 2015). The adoption of certain budget institutions may be endogenous to certain cultural traits. Countries more prone to thriftiness (say Germany) may be more likely to adopt certain budget rules and institutions, others may do the opposite. In addition, the rigorous application of certain budget rules (say a balanced budget amendments) may be endogenous to certain cultural traits having to do, for instance with the social acceptability towards "bending the rules," which may vary greatly across countries.[av] Both cross-country and within-country evidence would be useful. The latter could hold constant national institutions and examine the effect of difference cultural attitudes within the same national institutions.

The control of politicians is also a "public good" which may be under supplied in certain cultures, as shown by Nannicini et al. (2012) who develop an intuition by Banfield (1958). When "social capital" is low, people do not feel compelled to participate in political activities, control politicians and punish the latter when they misbehave. In fact, with low social capital individuals may expect private favors rather than public goods. Politicians then feel more free to exert less effort, be self-motivated or corrupt. Less control by voters may also allow powerful lobbies to have easier access to politicians. For instance, Campante and Do (2014) show that more isolated capital cities show more levels of corruption and are associated with a greater role for money in state-level elections. In particular, firms and individuals contribute disproportionately more compared to nonisolated capital cities. Thus, lower social capital may be associated with more political distortions and rent seeking of policymakers which may aggravate the deficit bias problem.

11.3 Delegation

In the case of monetary policy, the benefit of delegation to an independent (up to a point) agency is widely accepted. For fiscal policy this kind of delegation is virtually nonexistent. The question is why and whether some delegation in fiscal policy (and how and to whom) might be useful.

[at] See Richardson (2008).

[au] An exception on Italy is Alesina and Maré (1996) on Italy.

[av] On this point see for instance Guiso et al. (2011), Tabellini (2010), and Guiso et al. (2015).

The fundamental reason why delegation of an independent agency in monetary policy is more acceptable than fiscal policy goes back to where we started in this chapter. Fiscal policy is perceived as much more closely linked to redistributions of various type than monetary policy. In the case of the latter, instead a policy based upon some form of Taylor rule is (at least in normal times) considered as beneficial for society as a whole and redistribution issues may eventually be corrected by fiscal policy (say unemployment benefits during a recession). Alesina and Tabellini (2007) and Alesina et al. (2008) discuss issues of delegation and show results consistent with this argument: delegation is much less agreed upon when it involves redistribution while it is easier to achieve for more technical questions (say the conduct of monetary policy) with less direct distributional consequences.[aw] Blinder (1997) argues that even aspects of fiscal policies may benefit from some delegation. He notes that the benefits of Central Bank independence derived from the technical nature of the task, the long term effects of certain decisions, the desire to delegate to bureaucrats through choices when needed (say creating unemployment to fight inflation and diffuse the blame away from politicians) and the tendency of policymakers to inflate too much, possibly close to elections. This author correctly notes how many of these features apply also to certain fiscal policy decisions, especially in the case of tax policy. During the financial crisis, the close connections between monetary and fiscal policy (immortalized by the dramatic joint appearance of Henry Paulson and Ben Bernanke in front of Congress at the outset of the crisis) also made the sharp distinction between independent central banks and totally "political" governments even more striking and possibly artificial.

An intermediate step which does not imply delegation can be to create an independent fiscal council which examines the fiscal policy of the government and expresses an evaluation in terms of its short and long run effects and its technical problems. In the United States, the Congressional Budget Office with a reputation of skills and independence has this role. In Sweden a highly respected fiscal council issues an influential document every year to review the policy of the Swedish government. In the matter of delegation, even to a Council, probably cultural variables examined above play a role. In countries with high level of trust, delegation is easier and the independence of, say, a fiscal council would be (correctly) believed. This might be precisely the case of Sweden. In countries with low levels of trust (say Italy, Spain, or France), the independence of the council would not be believed, and this skepticism might not be unreasonable. Thus, the status of the council would be compromised and it would be viewed as politically influenced and would lose its legitimacy and its potentially useful role. This is another example of the interaction between institutions and culture discussed earlier. What and how to delegate in the area of fiscal policy remains an excellent topic of research.

[aw] Pettersson-Lidbom (2012) discuss evidence on legislature and bureaucratic relationship as a determinant of the size of government using two natural experiments.

11.4 Lobbyist and Bureaucrats

The role of the bureaucracy in the implementation of the budget is hardly studied by economists.[ax] Highly ranked bureaucrats may have an influence which goes well beyond the implementation of executive decisions. Thus, even without any formal delegation (discussed earlier) highly ranked bureaucrats when applying the fiscal provisions of the budget may have sufficient discretion to favor this or that pressure groups. Up to a point this may be a sort of "unwanted" delegation, that is a delegation which *de facto* but not *de iure* has the bureaucracy gains. This may increase the difficulty in implementing reforms because of a status quo preferences of existing bureaucratic bodies.

Finally, virtually all of the models we have considered model the polity by means of voting. A different view about the political process sees voting in legislatures simply as a result of lobbying pressure and therefore modeling lobbies' behavior is the fundamental step. While a rich literature on lobbies exists (see Grossman and Helpman, 2008), especially with regard to trade issues, we are not aware of lobbying models related to optimal debt management. Lobbyist and bureaucrats may be connected because the former may have access to the latter and may obtain favors in the implementation of various fiscal measures. This is especially the case when budget procedures and prescriptions are sufficiently opaque so as to guarantee a *de facto* discretion of bureaucrats. In turn, this lack of transparency may be strategically preserved precisely to allow for such pressures from lobbyist, with the related gains for policymakers. Linking the lobbying literature to government debt is an excellent topic of research.

11.5 Empirical Work

Much of the politico-economic literature reviewed earlier is theoretical. We think that there are high payoff in empirical research. Probably cross-country regressions have exhausted what they can teach us in most (but necessarily all) cases. Other tools are available. One is of course dynamic general equilibrium models where one could introduce political constraints or distortions and quantify their effects. A good example of this type of empirical work is the paper by Azzimonti et al. (2015) on the balanced budget rule reviewed earlier. At the opposite extreme of methodology one can think of historical case studies which would be especially helped by "natural experiment." For instance, imagine natural experiments which imply institutional changes (or other kind of changes) which can be considered relatively exogenous to fiscal policy. These studies may help address the question of endogeneity emphasized earlier. The use of historical evidence with time period spanning over institutional changes can be especially useful.

Within-country studies can also be helpful. Imagine a situation in which different localities within a country display very different policy stance regarding deficits. These

[ax] See Bertrand et al. (2015) and Gratton et al. (2015) for some recent work on the bureaucracy in India and Italy, respectively.

studies may shed some light on determinants of deficits, holding institutions constant. Evidence on localities is useful for two reasons. One because local public finance is important and interesting per se. Second, because, holding constant national institutions, we can investigate variations in other determinants of deficits. Much of this type of research is on US localities. Thus, there is room for work on other countries.

Another dimension in which progresses could be made is in the disaggregation of fiscal variables. Most of the literature refers to government spending, taxes and debt, without distinguishing within these broad categories. This is true (with few exceptions) both for the macro literature on fiscal policy and for the political economy literature. There is much unexplored territory here.

ACKNOWLEDGMENTS

We thank Marina Azzimonti, Marco Battaglini, Stephen Coate, Casey Mulligan, Per Pettersson-Lindbom, Guido Tabellini, Pierre Yared, Fabrizio Zilibotti. and the editors for very useful conversations and comments on earlier drafts.

REFERENCES

Acemoglu, D., Golosov, M., Tsyvinski, A., 2008. Markets versus governments. J. Monet. Econ. 55 (1), 159–189.

Acemoglu, D., Golosov, M., Tsyvinski, A., 2010. Dynamic mirrlees taxation under political economy constraints. Rev. Econ. Stud. 77 (3), 841–881.

Acemoglu, D., Golosov, M., Tsyvinski, A., 2011. Political economy of Ramsey taxation. J. Public Econ. 95 (7-8), 467–475.

Aghion, P., Bolton, P., 1990. Government domestic debt and the risk of default: a political-economic model of the strategic role of debt. In: Dornbusch, R., Draghi, M. (Eds.), Public Debt Management: Theory and History. Cambridge University Press, Cambridge, MA.

Aghion, P., Alesina, A., Trebbi, F., 2004. Endogenous political institutions. Q. J. Econ. 119 (2), 565–611.

Aguiar, M., Amador, M., Farhi, E., Gopinath, G., 2015. Coordination and crisis in monetary unions. Q. J. Econ. 130 (4), 1–50.

Aiyagari, S.R., Marcet, A., Sargent, T.J., Seppälä, J., 2002. Optimal taxation without state-contingent debt. J. Polit. Econ. 110 (6), 1220–1254.

Akhmedov, A., Zhuravskaya, E., 2003. Opportunistic political cycles: test in a young democracy setting. Q. J. Econ. 119 (4), 1301–1338.

Alesina, A., Bayoumi, T., 1996. The costs and benefits of fiscal rules: evidence from U.S. states. NBER Working Paper Series, http://www.nber.org/papers/w5614.pdf.

Alesina, A., Cukierman, A., 1990. The politics of ambiguity. Q. J. Econ. 105 (4), 829–850.

Alesina, A., Drazen, A., 1991. Why are stabilizations delayed? Am. Econ. Rev. 81 (5), 1170–1177.

Alesina, A., Giuliano, P., 2012. Preferences for redistribution. In: Bisin, A., Jackson, M.O. (Eds.), Handbook of Social Economics, vol. 1. North Holland, The Netherlands, pp. 93–131.

Alesina, A., Giuliano, P., 2015. Culture and Institutions. J. Econ. Lit., Am. Econ. Assoc. 53 (4), 898–944.

Alesina, A., Glaeser, E.L., 2005. Fighting Poverty in the US and Europe: A World of Difference. Oxford University Press, Oxford, UK.

Alesina, A., La Ferrara, E., 2005. Ethnic diversity and economic performance. J. Econ. Lit. 43, 762–800.

Alesina, A., Maré, M., 1996. Evasione e Debito. In: Monorchio, A. (Ed.), La Finanza Italiana Dopo la Svolta del 1992.

Alesina, A., Paradisi, M., 2014. Political budget cycles: evidence from Italian cities. NBER Working Paper 20570.

Alesina, A., Perotti, R., 1995. The political economy of budget deficits. NBER Working Paper Series 4637.

Alesina, A., Stella, A., 2010. The politics of monetary policy. In: Friedman, B.M., Woodford, M. (Eds.), Handbook of Monetary Economics, vol. 3. Elsevier Inc., pp. 1001–1054

Alesina, A., Tabellini, G., 1989. External debt, capital flight and political risk. J. Int. Econ. 27, 199–220.

Alesina, A., Tabellini, G., 1990. A positive theory of fiscal deficits and government debt. Rev. Econ. Stud. 57 (3), 403–414.

Alesina, A., Tabellini, G., 2007. Bureaucrats or politicians? Part I: A single policy task. Am. Econ. Rev. 97 (1), 169–179.

Alesina, A., Cohen, G.D., Roubini, N., 1993. Electoral business cycle in industrial democracies. Eur. J. Polit. Econ. 9 (1), 1–23.

Alesina, A., Perotti, R., Tavares, J., 1998. The political economy of fiscal adjustments. Brook. Pap. Econ. Act. 1 (1), 197–266.

Alesina, A., Tabellini, G., Campante, F.R., 2008. Why is fiscal policy often procyclical? J. Eur. Econ. Assoc. 6 (5), 1006–1036.

Alesina, A., Ardagna, S., Galasso, V., 2010. The Euro and structural reforms. In: Review of Economics and Institutions, vol. 2. National Bureau of Economic Research, Inc., University of Chicago Press and NBER, Chicago, pp. 1–37.

Alesina, A.F., Carloni, D., Lecce, G., 2012. The electoral consequences of large fiscal adjustments. In: Alesina, A., Giavazzi, F. (Eds.), Fiscal Policy after the Financial Crisis. National Bureau of Economic Research, Inc., The University of Chicago Press, Chicago and London, pp. 531–570.

Alesina, A., Favero, C., Giavazzi, F., 2015. The output effect of fiscal consolidation plans. J. Int. Econ. 96, 19–42.

Alt, J.E., Lassen, D.D., 2006. Fiscal transparency, political parties, and debt in OECD countries. Eur. Econ. Rev. 50 (6), 1403–1439.

Alt, J.E., Lowry, R.C., 1994. Divided government, fiscal institutions, and budget deficits: evidence from the states. Am. Polit. Sci. Rev. 88 (4), 811–828.

Arvate, P.R., Avelino, G., Tavares, J., 2009. Fiscal conservatism in a new democracy: "sophisticated" versus "naïve" voters. Econ. Lett. 102 (2), 125–127.

Auerbach, A.J., Gorkhale, J., Kotlikoff, L.J., 1991. Generational accounts: a meaningful alternative to deficit accounting. Tax Policy Econ. 5, 55–110.

Azzimonti, M., Francisco, E.D., Quadrini, V., 2014. Financial globalization, inequality, and the raising of public debt. Am. Econ. Rev. 104 (8), 2267–2302.

Azzimonti, M., Battaglini, M., Coate, S., 2015. Costs and benefits of balanced budget rules: Lessons from a political economy model of fiscal policy. MPRA Paper 25935.

Banfield, E., 1958. The Moral Basis of a Backward Society. The Free Press, Glencoe, Illinois.

Baqir, R., 2002. Districting and government overspending. J. Polit. Econ. 110, 1318–1354.

Barbera, S., Jackson, M.O., 2004. Choosing how to choose: Self-stable majority rules and constitutions. Q. J. Econ. 119 (3), 1011–1048.

Baron, D.P., Ferejohn, J.A., 1989. Bargaining in legislatures. Am. Polit. Sci. Rev. 83 (4), 1181–1206.

Barro, R.J., 1973. The control of politicians: an economic model. Public Choice 14-14 (1), 19–42.

Barro, R.J., 1979. On the determination of the public debt. J. Polit. Econ. 87 (5), 940.

Barseghyan, L., Battaglini, M., Coate, S., 2013. Fiscal policy over the real business cycle: a positive theory. J. Econ. Theory 148 (6), 2223–2265.

Bassetto, M., Sargent, T.J., 2006. Politics and efficiency of separating capital and ordinary government budgets. Q. J. Econ. 121 (4), 1167–1210.

Battaglini, M., 2014. A dynamic theory of electoral competition. Theor. Econ. 9 (2), 515–554.

Battaglini, M., Coate, S., 2008. A dynamic theory of public spending, taxation, and debt. Am. Econ. Rev. 98 (1), 201–236.

Battaglini, M., Coate, S., 2015. A political economy theory of fiscal policy and unemployment. J. Eur. Econ. Assoc. (forthcoming).

Bayoumi, T., Eichengreen, B., 1994. Restraining yourself: fiscal rules and stabilization. CEPR Discussion Papers 1029.

Bertrand, M., Burgess, R., Chawla, A., Xu, G., 2015. Determinants and consequences of bureaucrat effectiveness: evidence from the Indian administrative service. Unpublished.

Besley, T., Prat, A., 2006. Handcuffs for the grabbing hand? Media capture and government accountability. Am. Econ. Rev. 96 (3), 720–736.

Bisin, A., Verdier, T., 2015. On the joint evolution of culture and institutions. Unpublished.

Blinder, Alan S., 1997. Is Government Too Political? Foreign Affairs November/December 1997, 115–126.

Bohn, H., 1998. The behavior of U.S. public debt and deficits. Q. J. Econ. 113 (3), 949–963.

Bohn, H., Inman, R.P., 1996. Balanced budget rules and public deficits: Evidence from the U.S. states. Carn.-Roch. Conf. Ser. Public Policy 45 (1), 13–76.

Brender, A., 2003. The effect of fiscal performance on local government election results in israel: 1989-1998. J. Public Econ. 87 (9-10), 2187–2205.

Brender, A., Drazen, A., 2005. Political budget cycles in new versus established democracies. J. Monet. Econ. 52 (7), 1271–1295.

Brender, A., Drazen, A., 2008. How do budget deficits and economic growth affect reelection prospects? Evidence from a large panel of countries. Am. Econ. Rev. 98 (5), 2203–2220.

Buchanan, M.J., Wagner, E.R., 1977. Democracy in Deficit: The Political Legacy of Lord Keynes. Academic Press, Ney York, NY.

Budina, N., Schaechter, A., Weber, A., Kinda, T., 2012. Fiscal rules in response to the crisis: Toward the "next-generation" rules: A new dataset. International Monetary Fund, Working Papers 12/187.

Buti, M., Van Den Noord, P., 2004. Fiscal discretion and elections in the early years of EMU. J. Common Mark. Stud. 42 (4), 737–756.

Buti, M., Turrini, A., Van den Noord, P., Biroli, P., 2010. Reforms and re-elections in OECD countries. Econ. Policy 25, 61–116.

Campante, F.R., Do, Q.A., 2014. Isolated capital cities, accountability, and corruption: evidence from US states. Am. Econ. Rev. 104 (8), 2456–2481.

Canova, F., Pappa, E., 2006. The elusive costs and the immaterial gains of fiscal constraints. J. Public Econ. 90 (8-9), 1391–1414.

Casella, A., Eichengreen, B., 1996. Can foreign aid accelerate stabilisation. Econ. J. 106, 605–619.

Chari, V.V., Kehoe, P.J., 1993. Sustainable plans and debt. J. Econ. Theory 61 (2), 230–261.

Cukierman, A., Meltzer, A.H., 1986. A positive theory of discretionary policy, the cost of democratic government and the benefits of a constitution. Econ. Inq. 24 (3), 367–388.

Dabla-Norris, E., Allen, R., Zanna, L.F., Prakash, T., Kvintradze, E., Lledo, V.D., Yackovlev, I., Gollwitzer, S., 2010. Budget Institutions and Fiscal Performance in Low-Income Countries. p. 57.

Debrun, X., Moulin, L., Turrini, A., Ayuso-i Casals, J., Kumar, M.S., 2008. Tied to the mast? National fiscal rules in the European Union. Econ. Policy 23, 297–362.

Drazen, A., 2000. Political Economy in Macroeconomics. Princeton University Press, Princeton, NJ.

Drazen, A., Easterly, W., 2001. Do crises induce reform? Simple empirical tests of conventional wisdom. Econ. Polit. 13 (2), 129–157.

Drazen, A., Eslava, M., 2010a. Electoral manipulation via voter-friendly spending: theory and evidence. J. Dev. Econ. 92 (1), 39–52.

Drazen, A., Eslava, M., 2010b. Pork barrel cycles. NBER Working Paper Series 12190.

Drazen, A., Grilli, V., 1993. The benefits of crises for economic reforms. Am. Econ. Rev. 83, 598–607.

Easterly, W., 1993. How much do distortions affect growth? J. Monet. Econ. 32 (2), 187–212.

Elgie, R., McMenamin, I., 2008. Political fragmentation, fiscal deficits and political institutionalisation. Public Choice 136 (3-4), 255–267.

Fabrizio, S., Mody, A., 2006. Can budget institutions counteract political indiscipline? Econ. Policy 21 (48), 689–739.

Fatás, A., Mihov, I., 2003a. The case for restricting fiscal policy discretion. Q. J. Econ. 118 (4), 1419–1447.

Fatás, A., Mihov, I., 2003b. On constraining fiscal policy discretion in EMU. Oxford Rev. Econ. Policy 19 (1), 112–131.

Ferejohn, J., 1986. Incumbent performance and electoral control. Public Choice 50, 5–25.

Ferejohn, J.A., Krehbiel, K., 1987. The budget process and the size of the budget. Am. J. Polit. Sci. 31 (2), 296–320.

Foremny, D., Freier, M.D.M., Yeter, M., 2015. Overlapping political budget cycles. IEB Working Paper 02.

Galor, O., Klemp, M., 2015. Roots of autocracy. Unpublished.

Gavazza, A., Lizzeri, A., 2009. Transparency and economic policy. Rev. Econ. Stud. 76 (3), 1023–1048.

Gavazza, A., Lizzeri, A., 2011. Transparency and manipulation of public accounts. J. Public Econ. Theory 13 (3), 327–349.

Gavin, M., Perotti, R., 1997. Fiscal policy in Latin America. In: Bernanke, B., Rotemberg, J. (Eds.), NBER Macroeconomics Annual, vol. 12. MIT Press, Cambridge, USA, pp. 11–72.

Gonzalez, M.D.L.A., 2002. Do changes in democracy affect the political budget cycle? Evidence from Mexico. Rev. Dev. Econ. 6 (2), 204–224.

Gratton, G., Guiso, L., Michelacci, C., Morelli, M., 2015. From Weber to Kafka: political activism and the emergence of an inefficient bureaucracy. Unpublished.

Grilli, V., Masciandaro, D., Tabellini, G., Malinvaud, E., Pagano, M., 1991. Political and monetary institutions and public financial policies in the industrial countries. Econ. Policy 6 (13), 342–392.

Grossman, G.M., Helpman, E., 2008. Separation of powers and the budget process. J. Public Econ. 92 (3-4), 407–425.

Guiso, L., Sapienza, P., Zingales, L., 2006. Does culture affect economic outcomes? J. Econ. Perspect. 20 (2), 23–48.

Guiso, L., Sapienza, P., Zingales, L., 2011. Civic capital as the missing link. In: Benhabib, J., Bisin, A., Jackson, M.O. (Eds.), Handbook of Social Economics. In: vol. 1. North Holland, The Netherlands, pp. 417–480.

Guiso, L., Herrera, H., Morelli, M., 2015. A cultural clash view of the EU crisis. CEPR Discussion Papers 9679 (unpublished).

Halac, M., Yared, P., 2014. Fiscal rules and discretion under persistent shocks. Econometrica 82 (5), 1557–1614.

Halac, M., Yared, P., 2015. Fiscal rules and discretion in a world economy. NBER Working Paper 21492.

Hallerberg, M., Von Hagen, J., 1999. Electoral institutions, cabinet negotiations, and budget deficits in the European Union. In: Fiscal Institutions and Fiscal Performance. National Bureau of Economic Research, Inc., pp. 209–232.

Hallerberg, M., Strauch, R., Von Hagen, J., 2009. Fiscal Governance: Evidence from Europe. Cambridge University Press, Cambridge, UK.

Hassler, J., Krusell, P., Storesletten, K., Zilibotti, F., 2005. The dynamics of government. J. Monet. Econ. 52 (7), 1331–1358.

Haveman, R., 1994. Should generational accounts replace public budgets and deficits? J. Econ. Perspect. 8 (1), 95–111.

Inman, R.P., 1997. Rethinking federalism. J. Econ. Perspect. 11 (4), 43–64.

Kashin, K., King, G., Soneji, S., 2015. Systematic bias and nontransparency in US Social Security Administration forecasts. J. Econ. Perspect. 29, 239–258.

Klein, P., Krusell, P., Ríos-Rull, J.V., 2008. Time-consistent public policy. Rev. Econ. Stud. 75 (3), 789–808.

Kontopoulos, Y., Perotti, R., 1999. Government fragmentation and fiscal policy outcomes: evidence from OECD countries. In: Fiscal Institutions and Fiscal Performance. National Bureau of Economic Research, Inc., pp. 81–102.

Kornai, J., Maskin, E., Roland, G., 2003. Understanding the soft budget constraint. J. Econ. Lit. 41 (4), 1095–1136.

Krogstrup, S., Wyplosz, C., 2010. A common pool theory of supranational deficit ceilings. Eur. Econ. Rev. 54 (2), 269–278.

Lizzeri, A., 1999. Budget deficits and redistributive politics. Rev. Econ. Stud. 66 (4), 909–928.

Lucas, R.E.J., Stokey, N.L., 1983. Optimal fiscal and monetary policy in an economy without capital. J. Monet. Econ. 12 (1), 55–93.

Miano, A., 2015. Determinants and Consequences of Fiscal Rules: Evidence form the World Economy. Bocconi University.

Milesi-Ferretti, G.M., 2004. Good, bad or ugly? On the effects of fiscal rules with creative accounting. J. Public Econ. 88 (1-2), 377–394.

Milesi-Ferretti, G.M., Perotti, R., Rostagno, M., 2002. Electoral systems and public spending. Q. J. Econ. 117 (2), 609–657.

Müller, A., Storesletten, K., Zilibotti, F., 2016. The political color of fiscal responsibility. J. Eur. Econ. Assoc. 14 (1), 252–302.

Mulligan, C., Sala-i Martin, X., 1999. Gerontocracy, retirement, and social security. NBER Working Paper 7117.

Nannicini, T., Stella, A., Tabellini, G., Troiano, U., 2012. Social capital and political accountability. Am. Econ. J.: Econ. Policy 5 (230088), 222–250.

Oates, W.E., 2011. Fiscal Federalism. Edward Elgar Pub, Northampton, MA, USA.

OECD, 2014. Social spending–StatExtracs. http://stats.oecd.org/Index.aspx?DataSetCode=SOCX_AGG.

OECD, 2015. Social expenditure. www.oecd.org/els/social/expenditure.

Passarelli, F., Tabellini, G., 2013. Emotions and political unrest. CESifo Working Paper Series.

Peltzman, S., 1992. Voters as fiscal conservatives. Q. J. Econ. 107 (2), 327–361.

Persson, T., Svensson, L.E.O., 1989. Why a stubborn conservative would run a deficit: policy with time-inconsistent preferences. Q. J. Econ. 104 (2), 325–345.

Persson, T., Tabellini, G., 2000. Political Economics: Explaining Economic Policy. MIT Press, Cambridge.

Pettersson-Lidbom, P., 2001. An empirical investigation of the strategic use of debt. J. Polit. Econ. 109 (3), 570–583.

Pettersson-Lidbom, P., 2010. Dynamic commitment and the soft budget constraint: an empirical test. Am. Econ. J.: Econ. Policy 2, 154–179.

Pettersson-Lidbom, P., 2012. Does the size of the legislature affect the size of government? Evidence from two natural experiments. J. Public Econ. 96 (3-4), 269–278.

Ponticelli, J., Voth, H.J., 2011. Austerity and anarchy: budget cuts and social unrest in Europe, 1919-2008. CEPR Discussion Papers 8513, C.E.P.R. Discussion Papers.

Poterba, J.M., 1995. Capital budgets, borrowing rules, and state capital spending. J. Public Econ. 56, 165–187.

Poterba, J.M., 1996. Budget institutions and fiscal policy in the U.S. states. Am. Econ. Rev. 86 (2), 395–400.

Prat, A., Stromberg, D., 2013. The Political Economy of Mass Media. Cambridge University Press, Cambridge, UK.

Primo, D., 2007. Rules and Restraint: Government Spending and the Design of Institutions. University of Chicago Press, Chicago, IL.

Ramsey, F.P., 1927. A contribution to the theory of taxation. Econ. J. 37 (145), 47–61.

Reinhart, C.M., Rogoff, K.S., 2010. Growth in a time of debt. Am. Econ. Rev. 100 (2), 573–578.

Richardson, G., 2008. The relationship between culture and tax evasion across countries: additional evidence and extensions. J. Int. Account. Audit. Tax. 17 (2), 67–78.

Rogoff, K., 1990. Equlibrium political budget cycles. Am. Econ. Rev. 80 (1), 21–36.

Rogoff, K., Sibert, A., 1988. Elections and macroeconomic policy cycles. Rev. Econ. Stud. 55 (1), 1–16.

Sabato, L.J., 2008. A More Perfect Constitution: Why the Constitution Must Be Revised: Ideas to Inspire a New Generation. Walker Publishing Company, New York, NY.

Schuknecht, L., 2000. Fiscal policy cycles and public expenditure in developing countries. Public Choice 102 (1/2), 115–130.

Shi, M., Svensson, J., 2006. Political budget cycles: do they differ across countries and why? J. Public Econ. 90 (8-9), 1367–1389.

Song, Z., Storesletten, K., Zilibotti, F., 2012. Rotten parents and disciplined children: a politico-economic theory of public expenditure and debt. Econometrica 80 (6), 2785–2803.

Spolaore, E., 2004. Adjustments in different government systems. Econ. Polit. 16 (2), 117–146.

Stein, E., Talvi, E., Grisanti, A., 1999. Institutional arrangements and fiscal performance: the Latin American experience. In: Poterba, J.M., Von Hagen, J. (Eds.), Fiscal Institutions and Fiscal Performance. University of Chicago Press, Chicago, IL, pp. 103–134.

Stockman, D.R., 2001. Balanced-budget rules: welfare loss and optimal policies. Rev. Econ. Dyn. 4 (2), 438–459.

Tabellini, G., 1991. The politics of intergenerational redistribution. J. Polit. Econ. 99, 335.

Tabellini, G., 2010. Culture and institutions: economic development in the regions of Europe. J. Eur. Econ. Assoc. 8 (4), 677–716.

Tabellini, G., Alesina, A., 1990. Voting on the budget deficit. Am. Econ. Rev. 80 (1), 37–43.

Tornell, A., Lane, P.R., 1999. The voracity effect. Am. Econ. Rev. 89 (1), 22–46.

Trebbi, F., Aghion, P., Alesina, A., 2008. Electoral rules and minority representation in U.S. cities. Q. J. Econ. 123 (1), 325–357.

Velasco, A., 1999. A model of endogenous fiscal deficits and delayed fiscal reforms. In: Poterba, J.M., Von Hagen, J. (Eds.), Fiscal Institutions and Fiscal Performance. University of Chicago Press, Chicago, pp. 37–58.

Velasco, A., 2000. Debts and deficits with fragmented fiscal policymaking. J. Public Econ. 76 (1), 105–125.

Volkerink, B., De Haan, J., 2001. Fragmented government effects on fiscal policy: new evidence. Public choice 109 (3-4), 221–242.

Von Hagen, J., 2006. Fiscal rules and fiscal performance in the EU and Japan. Discussion Paper Series of SFB/TR 15 Governance and the Efficiency of Economic Systems 147.

von Hagen, J., Wolff, G.B., 2006. What do deficits tell us about debt? Empirical evidence on creative accounting with fiscal rules in the EU. J. Bank. Finance 30 (12), 3259–3279.

Weingast, B., Shepsle, K., Johnsen, C., 1981. The political economy of benefits and costs: a neoclassical approach to distributive politics. J. Polit. Econ. 84 (4), 642–664.

Woo, J., 2003. Economic, political, and institutional determinants of public deficits. J. Public Econ. 87 (3-4), 387–426.

Wyplosz, C., 2014. Fiscal rules: theoretical issues and historical experiences. In: Alesina, A., Giavazzi, F. (Eds.), Fiscal Policy After the Financial Crisis. University of Chicago Press and National Bureau of Economic Research, Chicago and London, pp. 495–529.

Yared, P., 2010. Politicians, taxes and debt. Rev. Econ. Stud. 77 (2), 806–840.